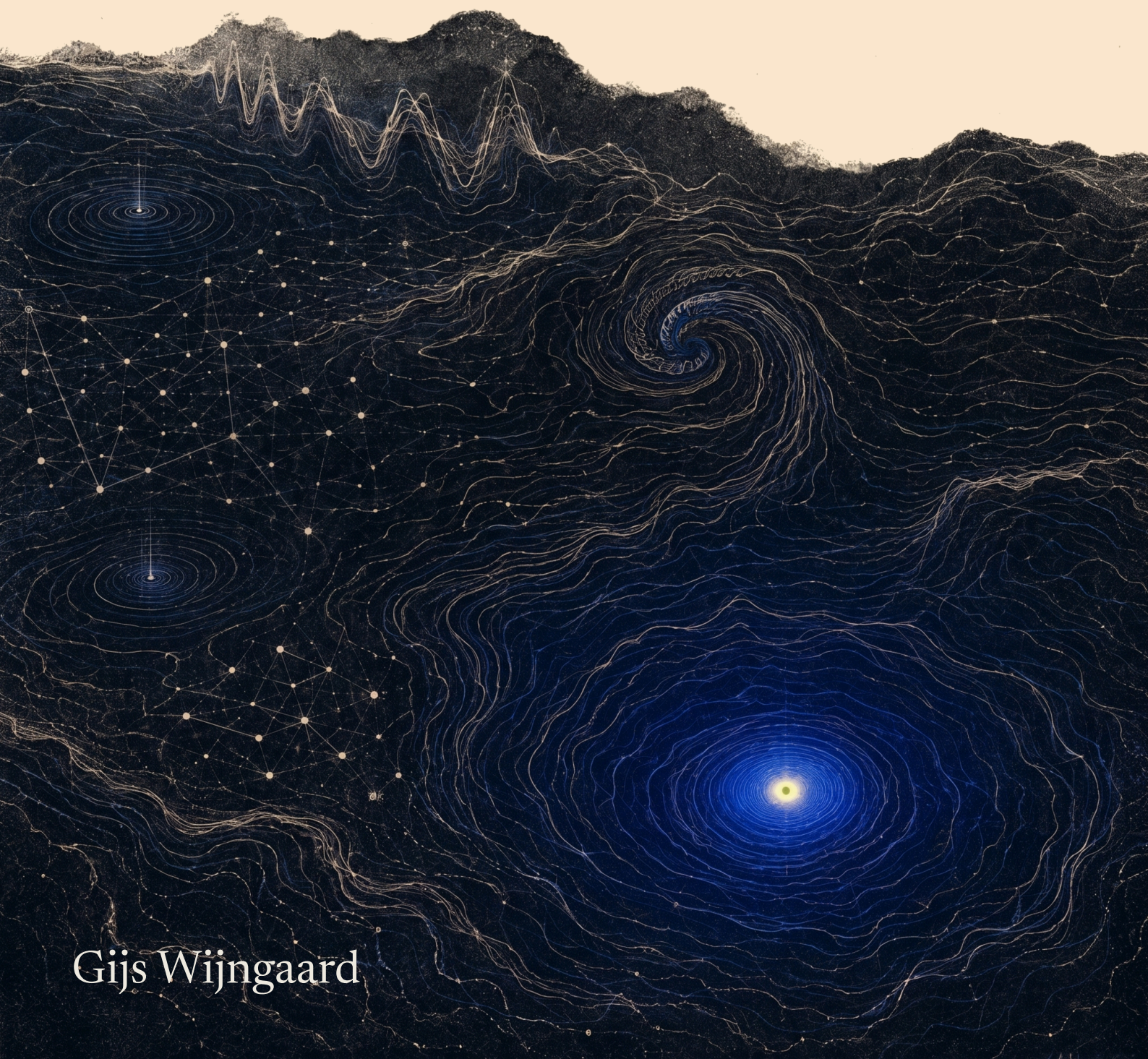


Data and Semantic Reasoning in Audio-Language Learning



Gijs Wijngaard



DOCTORAL THESIS

DATA AND SEMANTIC REASONING IN AUDIO-LANGUAGE LEARNING

GIJS WIJNGAARD

MAASTRICHT 2026



DATA AND SEMANTIC REASONING IN AUDIO-LANGUAGE LEARNING



Dissertation

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. Jan Smits,
in accordance with the decision of the Board of Deans,
to be defended in public
on Monday 7 September 2026, at 16:00 hours

by

GIJSBERT WILLEM ADRIAAN WIJNGAARD

SUPERVISORS

Prof. Dr. Elia Formisano, Maastricht University
Prof. Dr. Michel Dumontier, Maastricht University

ASSESSMENT COMMITTEE

CHAIR

Prof. Dr. Ir. Jan Scholtes, Maastricht University

MEMBERS

Prof. Dr. Paul Groth, University of Amsterdam
Dr. Aki Harma, Maastricht University
Dr. Chang Sun, Maastricht University
Prof. Dr. Toon van Waterschoot, KU Leuven

To Mar, Joos, and my parents

Contents

1	Introduction	1
1.1	Datasets of Sound	3
1.2	Human-Inspired Reasoning	4
1.3	Dissertation and Research Questions	5
2	Audio-Language Datasets of Scenes and Events	10
2.1	Introduction	12
2.2	Background	17
2.2.1	Audio-to-Language	17
2.2.2	Language-to-Audio	20
2.2.3	Datasets	22
2.2.4	YouTube-based datasets	26
2.2.5	Freesound-based datasets	32
2.2.6	Sound effect datasets	37
2.2.7	Field recording datasets	40
2.2.8	Movies & TV Sound Datasets	43
2.2.9	Datasets for Large Audio-Language models	44
2.3	Augmenting Audio-language datasets	45
2.3.1	Audio augmentation	46
2.3.2	Single Modal Text Augmentation	48
2.3.3	Audio-Text Multimodal Augmentation	51
2.3.4	Augmentation with other modalities	52
2.4	Audio Dataset Evaluation	53
2.4.1	Audio Dataset Analysis	53
2.4.2	Duplicate analysis	57
2.5	Discussion	60
2.5.1	Data quality	60
2.5.2	Bias in Audio Datasets	61
2.5.3	Dataset Accessibility	63
2.6	Conclusion	63
2.6.1	Future Research Directions and Recommendations	65
2.7	Appendix	67
2.7.1	List of audio-visual datasets	67
2.7.2	Example caption per dataset	67

3	The Data Advantage: Balanced Curriculum Learning for Enhancing Audio Question Answering	72
3.1	Introduction	74
3.2	Related work	75
3.3	Method	75
3.3.1	System Overview	75
3.3.2	Datasets	76
3.3.3	Experimental Setup	77
3.4	Results	80
3.4.1	Ablation Analysis	80
3.4.2	Alternative Approaches Tried	81
3.5	Conclusion	82
3.6	Acknowledgment	83
4	AudSemThinker: Enhancing Audio-Language Models through Reasoning over Semantics of Sound	84
4.1	Introduction	86
4.2	Related Work	89
4.2.1	Audio Datasets with Synthetically Generated Captions	89
4.2.2	Audio Models with Reasoning	89
4.3	AUDSEM Dataset	90
4.3.1	Dataset Creation	91
4.3.2	Dataset Processing	92
4.3.3	Dataset Generation	95
4.3.4	Semantic Descriptors	97
4.4	AUDSEMT THINKER Model	98
4.4.1	Supervised Fine-Tuning (SFT)	99
4.4.2	Reinforcement Learning with Group Relative Policy Optimization (GRPO)	99
4.5	Experiments	100
4.5.1	Experimental Setup	102
4.5.2	Main Results	102
4.5.3	Ablation Study	104
4.6	Conclusion	105
4.7	Acknowledgement	105
4.8	Appendix	106
4.8.1	Ethics Statement	106
4.8.2	Thinking Budget	107
4.8.3	Example Model Outputs	110
4.8.4	Prompts	112
5	ACES: Evaluating Automated Audio Captioning Models on the Semantics of Sounds	115
5.1	Introduction	117

5.2	Related work	118
5.2.1	Drawbacks of current metrics	118
5.2.2	Research in other fields	120
5.3	Methods	120
5.3.1	Model training	121
5.3.2	Metric definition	122
5.3.3	Hyperparameter tuning	124
5.3.4	Optimization	125
5.4	Experiments	125
5.4.1	Model evaluation	125
5.4.2	Human evaluation	127
5.5	Discussion	127
5.6	Conclusion	128
5.7	Acknowledgment	128
5.8	Appendix	128
5.8.1	Overview of hyperparameters	128
5.8.2	In depth comparison of ACES with previous metrics	130
6	AudioToolAgent: An Agentic Framework for Audio-Language Models	133
6.1	Introduction	135
6.2	Related Work	136
6.3	Methodology and Experimental Setup	137
6.3.1	Framework Overview	137
6.3.2	Evaluation Setup	139
6.3.3	Ablation Study Design	139
6.4	Results	140
6.4.1	Ablation Study	141
6.4.2	Discussion and Future Work	143
6.5	Conclusion	146
6.6	Acknowledgment	146
7	Discussion, Conclusion, and Future Work	147
7.1	Review of Contributions	148
7.2	Answering the Research Questions	150
7.3	Discussion, Limitations and Challenges	152
7.3.1	Data-Related Challenges	152
7.3.2	Evaluation and Benchmarking Challenges	154
7.3.3	Computational and Architectural Challenges	156
7.3.4	Legal, Ethical, and Societal Challenges	158
7.3.5	Environmental and Sustainability Concerns	160
7.4	Future Research Directions	160
7.4.1	Temporally Grounded Sound Events with Captions	161
7.4.2	Reasoning in Latent Space	165
7.4.3	End-to-End Training for Tool Use	166

7.4.4	Neuro-symbolic AI and Knowledge Graph Integration	167
7.4.5	Prompt Sensitivity and Automated Prompt Optimization	168
7.4.6	World Models and Multi-Sensory Perception	169
	Bibliography	171
	Summary	210
	Samenvatting	212
	Valorization	214
	Acknowledgments	217
	Published work	219
	About the author	220

Introduction



Hearing shapes how people live. Sound enables musical enjoyment, communication, and information about the environment. A laugh signals joy. A siren warns of danger. Movies and TV shows use sound effects to enhance the experience. Physical devices, such as blinking car lights, use sound to carry information. Animals rely on sound to survive: bats navigate by echolocation, and whales communicate across oceans. Sounds are everywhere and carry rich information.

Humans can hear sounds from an early age: babies can hear voices, music, and other sounds in the womb, even before birth [312]. Hearing is one of the first senses to develop. The ability to hear sounds supports children's development and learning. As children grow, their hearing skills improve through experience and language. Children learn to connect sounds with objects, events, and actions, forming categories of meaning [468] and build knowledge that links what they hear to what they know [506]. Later in life, the auditory system helps to recognize sounds and events [123, 37].

Machines learn from sound similarly. Computational models can analyze audio and generate text that describes it or answers questions about it. This field of machine learning is known as **audio understanding** or **machine listening** [487]. This dissertation focuses on this domain, in particular, on **audio-language models (ALMs)**, systems that connect an audio model with a language model to describe sound using natural language [101]. ALMs benefit from training on large, diverse datasets that pair audio with rich textual descriptions [171]. Two core tasks define ALMs' capabilities: **automated audio captioning (AAC)** and **audio question answering (AQA)**. AAC models generate text descriptions of audio, whereas AQA models produce textual answers to questions about audio.

For a model to process audio, the audio must first be converted to a digital format. Sound travels through air as pressure waves, which are captured by microphones

that convert these mechanical vibrations into electrical signals. These signals are then sampled at discrete time intervals, typically between 16,000 and 48,000 times per second, so they can be stored numerically on a computer. Before audio models can analyze these signals, researchers usually transform the raw waveforms into time-frequency representations. A common choice is the spectrogram, computed using the Short-Time Fourier Transform (STFT), which shows how frequency content evolves over time. Further transformation to the mel scale produces mel-spectrograms, which reflect human auditory perception, particularly humans' greater sensitivity to low frequencies [426]. The audio encoder, the first component of an ALM, takes the log mel spectrograms as input and converts them into audio embeddings. These embeddings are then combined with a language model to generate text output.

ALMs operate across three domains: music, sound events, and speech. Historically, audio understanding relied on domain-specific architectures, with each area (speech, music, and sound events) requiring its own dedicated models. Early speech systems focused on transcribing audio using end-to-end deep neural networks [21, 158]. In parallel, music domain research advanced, and tasks such as pitch analysis and instrument classification were introduced [206]. Similarly, sound-event research developed models to detect and classify individual acoustic events [427]. Because of this specialization, each domain required separate training data, models and evaluation metrics.

The shift toward unified end-to-end architectures began around 2020, driven by advances in sequence-to-sequence learning [437] and transformer-based models [470]. When the work for this dissertation began in 2021, ALMs were primarily domain-specific (environmental sound, speech, and music). Since then, AAC and AQA have emerged as core tasks for ALMs (see 2.2 for an overview of capabilities), although specialized models still outperform general-purpose ALMs in tasks such as speech transcription [308, 331].

ALM capabilities now support applications across accessibility, media, healthcare, environmental sensing, industry, and robotics. Assistive technologies translate alarms, doorbells, and approaching vehicles into text, vibration, or visual cues on phones and wearables for people who are deaf or hard of hearing [93], while automated audio subtitling labels non-speech sounds in user-generated online videos [8] and natural-language audio retrieval searches large archives by content rather than filenames or tags [207]. In the home, smart devices track leaks, glass breaking, and door openings to detect intrusions or accidents [354], and acoustic event detection systems flag falls, screams, and other distress events in elderly home settings to alert caregivers in real time [9].

In environmental sensing, ALMs that recognize bird songs and other wildlife vocalizations support biodiversity monitoring at scale [411, 383, 302], including bird-song recognition on low-power microcontrollers for field deployment [184]. Edge environmental sound classifiers bring similar capabilities to resource-constrained

devices in homes and cities [310, 469], and unsupervised models flag urban noise anomalies in smart-city deployments [156]. In industry, anomalous sound detection identifies early signs of wear or malfunction in machines and vehicles to support predictive maintenance [336]. In robotics, audio integrates with vision and touch in vision-language-action pipelines, where contact sounds inform manipulation [502] and multisensory fusion improves fine-grained interaction [242]; audio world models also predict the acoustic consequences of robot actions [575], so household robots react to events outside the visual field, such as a kettle boiling or a pot overflowing.

This dissertation follows two recurring themes: data and datasets, and ideas inspired by human cognition and neuroscience. The first theme examines how data quality, dataset design, and curation strategies affect model performance. The second theme builds on concepts from cognitive neuroscience and reasoning, such as semantic sound descriptors and explicit reasoning steps before generating an answer. The intention of this dissertation is to provide a collection of treatises: each main chapter corresponds to a separate article.

1.1 Datasets of Sound

As with children learning from examples, researchers train models to understand and reason about audio by providing pairs of sounds and text. High-quality data are essential: systematically improving data quality can yield stronger models than replacing architectures [335, 121]. Audio datasets follow the same principle [126]. The growth of the Internet has increased the availability of audio data, with platforms such as YouTube and Freesound hosting large collections of audio content widely used for model training.

Kaplan et al. [194] and Hoffmann et al. [173] show that model performance scales with data size, model parameters and compute. This phenomenon is known as *scaling laws*. Richard Sutton (a pioneer in reinforcement learning) argued in his paper “The Bitter Lesson” that general methods that scale with computation ultimately prevail [438]. The Chinchilla study [173] refined these principles by showing that models should be trained with substantially more data as parameter count increases, with approximately 20 tokens per parameter for compute-optimal training. As the size of the datasets used to train language models continues doubling approximately every six months [378], progress in ALM performance is likely to follow. State-of-the-art models such as Qwen 3 and Llama 4 require around 36 trillion and 40 trillion tokens, respectively [559, 301].

However, more data does not guarantee better models. The principle of *garbage in, garbage out* applies: models trained on low-quality, noisy, or biased data will inherit those flaws regardless of scale [335]. In audio-language learning, two common issues arise: (1) YouTube datasets often contain off-screen sounds or captions describing visual rather than auditory content [435, 222], and (2) User-created metadata from

platforms such as Freesound frequently contains irrelevant or incomplete information [116, 115, 603]. Such issues make data cleaning and curation crucial, alongside efforts to scale datasets.

With the growth of audio collections, the field also faces structural challenges related to dataset design and documentation. Analyses and surveys highlight that audio-language datasets differ widely in annotation pipelines, label quality and granularity, and licensing [435, 23, 543, 162]. Recurring problems include overlapping training and evaluation splits, repeated reuse of identical audio across datasets, limited language and domain diversity, and legal constraints on downloading material from platforms such as YouTube complicate reproducibility and long-term access.

Embedding-based similarity measures offer a way to quantify dataset overlap and coverage. Encoders convert audio into embedding vectors, and contrastively trained encoders place similar audios near one another in the latent space. Cosine similarity between embeddings measures their overlap. CLAP audio-text embeddings [525] are particularly useful because audio and text describing the same sound yield similar embeddings. These embeddings can be used to filter candidate duplicate or near-duplicate examples during dataset construction: candidates too similar to an existing instance are removed.

1.2 Human-Inspired Reasoning

This dissertation draws ideas from how humans listen and reason, with two key elements. First, the *semantics of sound*: listeners describe everyday sounds using attributes such as who or what makes the sound, how it is produced, and where or when it occurs (see Table 4.2 for a detailed description). Previous work suggests that the brain engages distinct neural mechanisms for perceiving and recognizing objects (living or non-living) and actions [44, 237, 382]. Following Giordano et al. [133], these semantic descriptors provide a structured vocabulary to analyze and compare audio descriptions.

The second component involves *thinking before answering*. In cognitively demanding tasks such as playing chess, longer thinking time improves performance [467, 49], reflecting the well-known speed-accuracy tradeoff in decision-making [166]. Similarly, Kahneman [191] distinguishes two types of thinking: System 1 thinking is fast and intuitive, while System 2 thinking is slow, logical, and analytical. System 2 thinking is associated with longer thinking time and better performance in demanding tasks [191].

In ALMs and, more broadly, in language modeling the terminology of thinking before answering is referred to as **reasoning**. A model can place its reasoning in tokens, enclosed by a `<think>` and `</think>` tag. Because attention in transformer models incorporates previous tokens, the model can reason about audio using its own intermediate thoughts. This reasoning process is known as **Chain-of-Thought (CoT)** [499],

and analogies have been made to system 2 thinking of humans [248, 484]. Research shows that large-scale reasoning models outperform similar non-reasoning models on demanding tasks [499, 148].

1.3 Dissertation and Research Questions

This dissertation examines audio-language modeling through two themes: (i) data quality and (ii) human-inspired structure and reasoning. The central thesis is that balanced datasets with low-overlap reduce contamination and improve training stability; semantic descriptors, semantics-aware metrics, and tool-based systems support interpretable and verifiable outputs and improve performance on audio-language tasks.

This dissertation addresses five research questions across five chapters:

RESEARCH QUESTION 1

Dataset Landscape and Its Impact on ALMs

What is the current landscape of audio-language datasets, and how do their origins, overlaps, and augmentation strategies shape the development and evaluation of audio-language models?

Chapter 2 surveys 69 audio-language datasets up to September 2024, mapping their origins across five main sources: YouTube, Freesound, sound effect libraries, field recordings, and movies. Datasets are organized by their original task (captioning, retrieval, question answering, generation, and source separation), and statistics such as caption counts, total hours, vocabulary size, and accessibility are summarized. The mapping reveals that YouTube-based and Freesound-based datasets dominate, while other resources provide complementary coverage.

Using CLAP embeddings [525] and large-scale principal component analyses, Chapter 2 visualizes semantic coverage of each dataset in both audio and text spaces. The analysis reveals that a small set of categories, such as indoor environments, music, and male speech, account for nearly half of all samples. It identifies datasets with distinct centroids (e.g., AnimalSpeak, AudioCaption, MACS, and AudioSet) and contrasts them with other resources such as Freesound. Pairwise cosine similarity analyses reveal significant acoustic and textual overlap among popular datasets, such as AudioCaps, WavCaps, and Clotho and their derivatives.

A section on data augmentation categorizes techniques into audio-only, text-only, multimodal and cross-modal (e.g. visual) methods. While augmentation generates more training data and improves quality by filtering and rewriting captions, it may also reinforce existing biases. The survey also shows that, by 2024, most new datasets rely on LLM-generated captions, often produced for the same underlying audio. This potentially increases redundancy and reinforce biases.

Overall, the chapter argues that dataset quality, accessibility, and diversity are more important than scale alone. Recommendations include combining low-overlap field-recording and sound-effect datasets for balanced training corpora, distributing data via platforms like Zenodo or HuggingFace for reproducibility, and adopting systematic duplicate-detection pipelines. Later chapters build on these curated datasets and analyses.

RESEARCH QUESTION 2

Curriculum Learning and Data Balancing for AQA

How can curriculum learning and data balancing improve audio question answering performance?

Chapter 3 examines these issues in Audio Question Answering research: (1) contamination between training and evaluation data, (2) strong class imbalance in audio categories (with music and speech dominating), and (3) instabilities when applying reinforcement learning.

The chapter highlights that AQA benchmarks, such as MMAU [393], inherit large portions of their audio from widely used datasets such as AudioSet, Clotho, and WavCaps, often used as training data. This complicates claims of generalization.

To address these issues, Chapter 3 introduces a data-balanced curriculum pipeline. It constructs a clean version of the DCASE 2025 Task 5 development dataset¹ with no audio overlap across splits. Questions and answers are categorized with Phi-4-mini [2] into semantic classes, and statistical thresholds limit overrepresented categories, producing a cleaner, more balanced dataset.

On this curated data, the chapter applies curriculum-stabilized Group Relative Policy Optimization (GRPO). The curriculum first presents easy examples, based on question type and model confidence, to stabilize early reward signals before harder examples are introduced. The full pipeline combines supervised fine-tuning with low-rank

¹<https://dcase.community/challenge2025/task-audio-question-answering>

adaptation (LoRA) and reinforcement learning on the filtered, curriculum-ordered data.

Experiments on the DCASE 2025 Task 5 development dataset show that the data-balanced curriculum pipeline improves accuracy by 3.3% over a baseline that uses standard supervised fine-tuning and GRPO. Ablations confirm that moderate filtering and an easy-to-hard curriculum yield the most stable and effective training.

RESEARCH QUESTION 3

Semantic-Descriptor-Based Human Inspired Reasoning

How can incorporating human-inspired reasoning through semantic descriptors enhance audio-language model performance?

Chapter 4 addresses this question by introducing AUDSEMT THINKER, an audio-language model whose reasoning is structured around a set of semantic descriptors inspired by auditory cognitive neuroscience. The model explicitly analyzes who or what produces a sound, how it is generated, and where and when it occurs before producing an answer or caption. This design mirrors how human listeners describe everyday sounds. These descriptors are also used in Chapter 5.

To support this approach, Chapter 4 introduces AUDSEM, a dataset created from YouTube closed captions using a multi-stage pipeline involving audio models, language models, and visual cues. Captions are rewritten to match audible content, harmful material is removed, and CLAP embeddings filter audio overlapping with existing benchmarks. The dataset contains question answering, captioning, and creative prompts tailored for semantic reasoning.

AUDSEMT THINKER generates outputs in three-phases: a free-form thinking phase, a structured semantic analysis, and a concise answer phase. The chapter explores both supervised fine-tuning and GRPO [409] with length-constraint rewards that control the model’s thinking budget.

Evaluations on MMAU [393], OmniBench [247], and AudioBench [475] show that AUDSEMT THINKER matches or outperforms strong baselines, especially in music and environmental sound reasoning. Ablations indicate that explicit semantic descriptors improve performance in supervised settings and that GRPO can further refine reasoning patterns when reward design and length constraints are chosen carefully.

RESEARCH QUESTION 4

Semantics-Aware Evaluation Metric for Audio Captioning

How can an audio captioning metric better align with human perception by using semantic descriptors?

Chapter 5 introduces ACES (Audio Captioning Evaluation on Semantics of Sounds), a metric that evaluates captions based on semantic descriptors rather than n-gram overlap. ACES identifies entities corresponding to who/what, how, where/when, and related perceptual attributes using a fine-tuned RoBERTa model [266] that performs token-level labeling and aggregates descriptors into entity representations, and cosine similarities between contextualized embeddings measure alignment between candidate and reference captions. ACES incorporates a METEOR-inspired F-score [27] that emphasizes recall over precision, adds a length penalty for overly short captions, and includes a fluency detector.

On the FENSE benchmark [600], ACES performs comparably to or better than existing embedding-based metrics, especially on the human-incorrect and human-machine categories of the Clotho-Eval dataset. These categories test the ability of a metric to separate incorrect captions from correct ones and to distinguish machine-generated captions from human references. Qualitative analyses reveal cases where ACES aligns more closely with human judgments because it focuses on key sound descriptors rather than global sentence similarity.

The chapter also discusses limitations, including sensitivity to missing descriptor types and cases with no overlapping entities. Nonetheless, ACES provides a semantically grounded alternative to token- or sentence-embedding metrics. This semantics-aware perspective is also used in Chapter 6.

RESEARCH QUESTION 5

Agentic Framework for Coordinating Audio Tools

How can agentic frameworks that coordinate specialized audio tools achieve state-of-the-art performance without additional training data?

Chapter 6 presents AudioToolAgent, a framework that orchestrates audio-language models and speech recognizers as tools controlled by a text-only central agent. The agent receives an audio file path, a question, and answer options and then determines

which tools to call, in what order, and with what prompts. Tool adapters support tasks such as audio question answering and speech-to-text conversion, and the agent issues follow-up queries when responses are incomplete or inconsistent.

The chapter describes open-source and closed-source configurations of AudioToolAgent. The open-source configuration combines the Qwen 3 235B agent [559] with Qwen3 Omni [559], Qwen2.5 Omni [544], Audio Flamingo 3 [134], and Whisper [374] as tools. The closed-source configuration uses Gemini 3 Pro [124] as the agent and Gemini 3 Pro, Qwen2.5 Omni, Audio Flamingo 3, and Whisper as tools. The framework does not require additional fine-tuning of the underlying models.

Experiments on MMAU [393], MMAR [280] and MMAU-Pro [224] show that AudioToolAgent reaches or surpasses state-of-the-art accuracy across speech-heavy and mixed-sound tasks. The framework matches or surpasses strong multimodal baselines that operate as single models. Gains stem from combining specialized tools, for example, pairing robust automatic speech recognition models with general audio-language models, within one reasoning pipeline.

Ablations quantify the contribution of individual tools and agents, showing that general-purpose audio models contribute more to the agentic system than narrow speech recognizers. Furthermore, the performance of models as tools does not correlate with the benchmark scores of those models. The chapter also analyzes practical constraints, such as latency, compute cost, and outlines future directions for the selection of learned tools and more efficient orchestration. Together, these results demonstrate that an agentic framework can transform existing audio models into more capable systems without new training data.

Chapter 7 synthesizes findings across all chapters by revisiting the central themes of data quality and human-inspired reasoning and discussing their interaction. It summarizes the dissertation’s contributions and outlines open challenges, including multilingual and domain-specific datasets, latent-space reasoning, prompt sensitivity, and multisensory world models for robotics.

For each chapter, the PhD candidate conceived the research idea, curated the data, designed the methodologies, implemented and ran all experiments, conducted the analyses, and drafted the papers in full, which were subsequently refined with input from his supervisors.

Audio-Language Datasets of Scenes and Events



ADAPTED FROM

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. “Audio-Language Datasets of Scenes and Events: A Survey”. In: *IEEE Access: Practical Innovations, Open Solutions* 13 (Jan. 2025), pp. 20328–20360. doi: 10.1109/ACCESS.2025.3534621

Abstract

Audio-language models (ALMs) generate linguistic descriptions of sound-producing events and scenes. Advances in dataset creation and computational power have led to progress in this domain. This chapter surveys 69 datasets used to train ALMs, covering research up to September 2024. The survey provides an analysis of the origins of the dataset, the audio and linguistic characteristics, and the use cases. Key sources include YouTube-based datasets, such as AudioSet, with more than two million samples, and community platforms such as Freesound, with more than one million samples. The survey evaluates the acoustic and linguistic variability between datasets by analyzing the principal components of audio and text embeddings. The survey also analyzes data leakage through CLAP embeddings and examines sound category distributions to identify imbalances. The survey examines augmentation techniques for audio-language datasets, including traditional methods such as MixUp and SpecAugment, LLM-based text augmentation, and multimodal approaches that take advantage of both audio and textual transformations. Finally, the survey identifies development challenges: dataset overlap, biases, accessibility barriers, and English-language dominance. The survey also shows specific areas requiring attention: multilingual dataset development, specialized domain coverage, and improved dataset accessibility. In the GitHub repository¹ all scripts are available to reproduce the analysis.

Preface

This chapter answers Research Question 1 (Dataset Landscape and Its Impact on ALMs) by characterizing audio-language datasets, their sources, overlap, and augmentation practices, and by analyzing how these properties shape training and evaluation. It contributes to the data-quality theme by showing how dataset quality, accessibility, diversity, and leakage constrain what acoustic events models can learn and how reliably performance can be measured. The resulting dataset map and overlap analyses motivate the dataset curation choices and evaluation cautions used in subsequent chapters.

The initial version of this chapter was accepted at the IEEE Access journal [509]. This dissertation version includes an additional section on data augmentation techniques (Section 2.3) that was omitted from the submitted version.

¹<https://github.com/GLJS/audio-datasets>

2.1 Introduction

Audio-Language learning, also known as Audio-Text learning, is a rapidly growing research field that focuses on processing, understanding, and describing sounds using natural language. Audio-language models fall under the broader category of Multimodal Large Language Models, which can process and generate responses based on diverse input modalities [570].

Advancements in machine learning and the increasing availability of datasets that pair sounds with corresponding textual descriptions are driving the growth of audio-language learning. The field has witnessed a surge in activity due to the development of large language models (LLMs), which have inspired new research directions [586, 327]. For example, natural language processing models based on the Transformer architecture [470] have been adapted for audio-language Learning in various settings [293, 552, 321, 485, 231]. In addition, techniques such as contrastive learning, designed to differentiate similar and dissimilar data pairs, and transfer learning have led to improvement in performance in various audio-language tasks [88, 105, 525, 17, 352].

The availability of high-quality datasets has been another key driver of progress in this field. Two milestones were the release of the Freesound platform in 2013 [117] and AudioSet in 2017 [126]. These resources have enabled training models for various sounds and tasks. Since then, numerous studies have followed [199, 102, 525, 296], leading to development and refinement of datasets tailored to the needs of specific audio-language models [139, 88, 73].

Recent studies indicate scaling laws for the development of foundation models [194, 173]: a correlation between model performance and the amount of data and training time these models use. In Audio-Language Learning, for example, the ESC-50 audio classification dataset requires around 2 million audio-text data points for a zero-shot model to match human parity [588]. A major challenge in acquiring audio-language data is that the amount of training audio data available is inevitably smaller than that of natural language processing tasks [440]. Current audio datasets are mainly based on sound effects databases, YouTube, or Freesound for their audio, whereas LLMs are trained on a large portion of the open Internet [376]. A limited training dataset has presented the issue of *distributional shift*: a mismatch between training data and production data, which makes a model poorly equipped for real-world data [53]. Models trained with contrastive learning (CLAP) [525, 105], benefit from large and diverse datasets for better performance in audio-language domains [203, 261, 520, 334, 105].

A recent development in audio-language datasets is the use of LLMs to annotate, generate, or augment training data for audio-language models (e.g., [296, 435, 129, 214, 383]; see also [598]). LLM-generated sound captions offer a way to address the scarcity of human-annotated high-quality sound datasets. Another complementary approach, as demonstrated in this work, involves utilizing a wider range of existing, diverse data sources.

This chapter reviews the available datasets in the audio-language domain, analyzing 69 datasets, expanding on previous surveys that reviewed 2 datasets [485], 3 datasets [293, 552, 321], 5 datasets [231], and 20 datasets [461]. This chapter contributes the following:

1. **Dataset Overview:** A detailed overview of datasets used in the field of audio-language models, including relevant statistics (Table 2.1) such as the average length of audio and text (captions), total hours, sources, and application domains.
2. **Data Leak Analysis:** A data overlap analysis (Section 2.4.2) that assesses the extent of data leakage between datasets. This analysis is critical for combining datasets, developing new ones, and properly evaluating zero-shot learning tasks.

An additional section on data augmentation techniques categorizes augmentation into four approaches: single-modal audio techniques, single-modal text augmentation, multimodal augmentation that combines both modalities, and augmentation with other modalities, including visual features and retrieval augmentation. The analysis reveals that text augmentation can have bigger impact than audio augmentation in traditional setups. In addition, it demonstrates how contemporary LLMs serve multiple roles in augmentation, such as caption generation, paraphrasing, and quality assessment of training data.

Through this analysis of existing datasets, this chapter identifies key challenges in the development of ALM datasets. These include data quality and bias issues (particularly in visually-grounded datasets), imbalanced representation of sound categories and languages, dataset accessibility and reproducibility concerns, and risks of model collapse in LLM-augmented datasets. Specifically, the survey highlights the need for multilingual datasets beyond English, specialized domain coverage (e.g., environmental audio, industrial applications, natural soundscapes), and improved dataset accessibility and reproducibility (particularly for YouTube-based datasets).

“Zero-shot” learning, the ability of models to recognize sounds they were not explicitly trained on, is becoming a relevant aspect for evaluating audio-language models [444, 131, 181, 88]. However, achieving true zero-shot learning is challenging in practice. Pretrained models may have already been exposed to similar data during training due to data leakage or overlap between datasets [447]. For example, a model pretrained on AudioSet [126] may already have seen “*dog barks*” sound samples, which questions whether recognizing “*dog barks*” in a new dataset represents genuine zero-shot learning. This indicates the need for a thorough dataset analysis to understand the capabilities and limitations of audio-language models and to evaluate zero-shot performance robustly. The data leak analysis presented in this survey directly addresses these concerns by examining overlap between datasets.

Given the extensive scope of the audio-language field, this chapter focuses on two specific aspects. First, it examines the domain of sound and event machine listening,

deliberately excluding speech and music modeling from its scope (for surveys on speech and music modeling, see [289, 365, 231, 3]). Second, it distinguishes between different types of machine listening tasks. Traditional tasks, such as Sound Event Detection (SED) and Audio Tagging (AT) are limited to predicting within the scope of the labeled data on which they were trained. For example, a model trained on the AudioSet dataset can only predict labels from its 527 classes. In contrast, recent audio-language models use an encoder-decoder setup, integrating an audio understanding and representation model with a text understanding and representation model. This survey focuses on audio-language models and the tasks they support, including Automated Audio Captioning (AAC), Audio-Text Retrieval (ATR), Automated Audio Generation (AAG), and Text-to-Audio Retrieval (ATR). These tasks, along with others, are briefly introduced in Section 2.2.

The survey is organized as follows: The Background section (Section 2.2) introduces the field and its models. Then it reviews related work on audio-language research and describes the research methodology used. An overview of the identified datasets in the field is provided in Section 2.2.3. The survey also includes an analysis of data leakage (Section 2.4.2), followed by a discussion of bias and data quality in audio datasets (Section 2.5). Finally, the article concludes with a summary of findings and future research directions (Section 2.6).

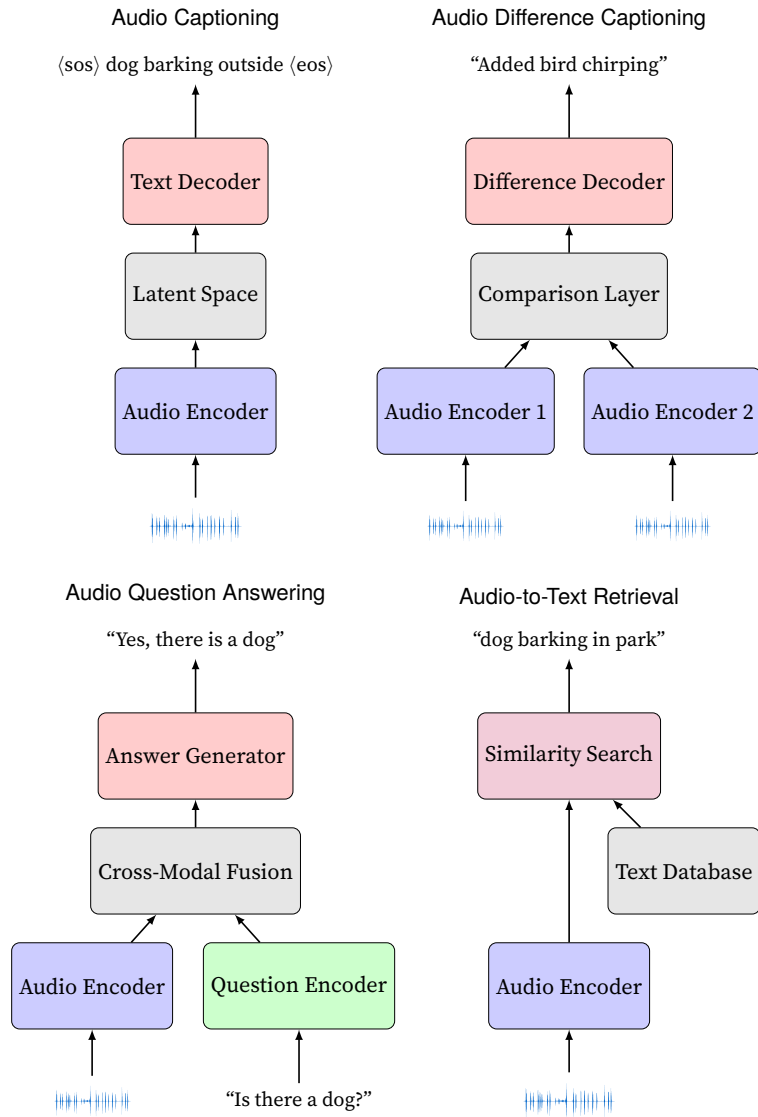


Figure 2.1: Audio-to-language models process audio inputs to produce textual outputs across captioning, comparison, question answering, and retrieval tasks.

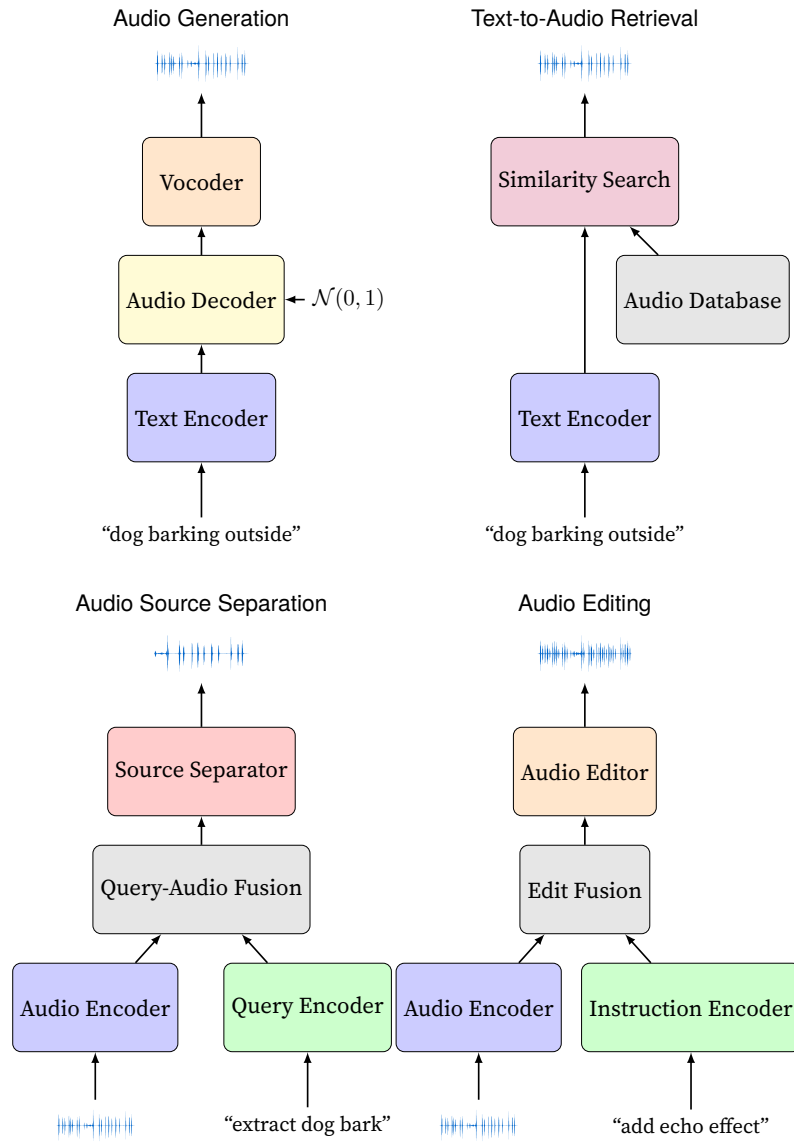


Figure 2.2: Language-to-audio models consume textual prompts to retrieve, separate, edit, or generate audio signals.

2.2 Background

This section provides an introductory overview of Audio-Language Learning, categorizing the field into two primary task types: **Audio-to-Language** and **Language-to-Audio**. Figure 2.1 outlines and illustrates the key tasks within these categories. For detailed descriptions of specific tasks, see the surveys by Xu et al. [552] and Wang et al. [485].

2.2.1 Audio-to-Language

The Audio-to-Language category comprises models that transform audio inputs into textual outputs, either by generating text or retrieving it from a dataset. This category encompasses several subfields:

1. **Audio Captioning:** Generating natural language descriptions for audio events.
2. **Audio Difference Captioning:** Showing content differences between pairs of similar audio clips (e.g. the addition or removal of specific sounds).
3. **Audio Representation Learning:** Learning audio representations that are effective for downstream tasks.
4. **Audio-Text Retrieval:** Using a database of audio-language pairs to find textual descriptions corresponding to given audio inputs.
5. **Audio-Question Answering:** Answering natural language questions about specific audio content.

(Automated) Audio Captioning (AAC)

As introduced by Drossos et al. [101], AAC refers to converting audio signals into natural language descriptions. Prominent datasets within this domain include Clotho [102] and AudioCaps [199] (see Section 2.2.3). AAC models predominantly use an encoder-decoder framework. The **encoder** processes audio signals into latent representations. The **decoder** converts these representations into textual descriptions. Encoders are frequently pretrained in datasets such as AudioSet [126] for audio classification tasks [208, 291], facilitating transfer learning to AAC tasks, as demonstrated in various studies [517, 290, 190]. Common encoder architectures include Convolutional Neural Networks (CNNs) [126, 212] and Transformer encoders [137]. Decoders employ Transformers [375, 238]. Transformers and RNNs allow for variable-length audio input, while CNNs are often limited to fixed-length input due to their pre-defined kernel size and spatial dimensions. Various other architectures have been proposed, such as GANs [294], diffusion models [546], and variational audioencoders [581].

AAC models are trained on paired audio-text data (τ, α) , using various training methods. Early models used encoder-decoder architectures, where an audio encoder

produced embeddings for a text decoder. Contemporary approaches include cross-attention mechanisms that enable multimodal audio-language fusion [140, 209, 58, 290], integration using audio embeddings as prefixes for text decoders [405, 204], and zero-shot approaches that adapt pretrained multimodal models and LLMs without direct training [408, 396, 302]. Innovative variations include retrieval-augmented captioning [129] and unified encoder architectures [73, 444], which allow for integration of audio and text modalities.

A powerful training technique that has emerged for audio captioning models is Instruction Tuning, which improves model performance with reformatted training data of instruction-response pairs. This approach has demonstrated superior results compared to traditional training techniques [80]. Two primary strategies have been explored: using templates crafted manually [498, 74, 399, 490] or using LLMs to automatically generate instruction datasets [175, 489]. Notable examples include Vicuna [590], Stanford Alpaca [446], and LLaMA-GPT4 [362] all developed by fine-tuning LLaMA using LLM-generated instruction datasets. This technique has also proven effective in multimodal contexts, such as guiding visual encoders in extracting and explaining image features using instructions [602].

The Audio Logging (AL) task extends the concept of AAC by generating longer, paragraph-length descriptions of audio content. For example, Bai et al. [24] introduced a method that uses LLMs to describe audio recordings up to three minutes long.

The task of Soundscape Captioning, as proposed by [176], aims to generate context-aware descriptions of soundscapes by capturing the acoustic scene, event information, and the corresponding human affective qualities. This task seeks to automate traditional soundscape analysis, which is based on subjective ratings and labor-intensive surveys.

Audio Entailment, introduced by [90], involves evaluating whether a text description can be inferred from an audio recording. Each text description is categorized as either an entailment, neutral, or contradiction with respect to the audio. This task assesses an audio-language model's deductive reasoning capabilities by determining whether a text hypothesis can be logically inferred from an audio.

Text-queried Sound Event Localization (SEL), proposed by [585], involves identifying the location or direction of a specific sound event described in natural language. Unlike traditional Sound Event Localization and Detection (SELD) systems, which are limited to predefined classes, SEL models allow users to describe target sounds in natural language, making the task flexible and user-centric.

Audio Difference Captioning (ADC)

Audio Difference Captioning (ADC) addresses a limitation in traditional audio captioning, where models often generate similar descriptions for similar sounds without adequately capturing the differences in their content. ADC models use annotations that explicitly highlight differences between captions of similar audio clips, enabling better differentiation. For example, Tsubaki et al. [463] demonstrated the application of ADC in detecting small anomalies in machine sounds to demonstrate its usefulness in real-world scenarios. Prominent datasets for ADC include: (1) the MIMII-Change dataset [463], which focuses on linguistic descriptions of anomalies in machine-sounds. (2) The AudioDiffCaps dataset [441], which features audio clips with artificially mixed background and foreground sounds derived from existing environmental sound datasets.

A related technique, termed *Audio Difference Learning*, is described by Komatsu et al. [211], specifically as a pretraining strategy for Audio Captioning models. In this approach, two audio inputs, an input and a reference audio, are mixed and encoded. The encoded reference is subtracted from the encoded input and the result is passed to a decoder to generate the caption. This method bypasses the need to create specialized ADC datasets, simplifying the training process.

Additionally, a closely related task called *Natural Language Audio Reasoning* has been proposed by [249]. NLAR focuses on distinguishing, comparing, and summarizing differences between two audio clips. To support this task, the researchers introduced the NLAR dataset, which has proven effective in evaluating the reasoning capabilities of audio models.

Audio Representation Learning (ARL)

Audio Representation Learning (ARL) focuses on learning audio representations useful for downstream tasks. This involves training a mapping function f that transforms an audio clip x into a meaningful representation y . The mapping function f is often learned to use techniques such as contrastive learning [521] or self-supervised learning [58, 290]. These methods aim to create robust embeddings that capture essential features of audio signals, enabling improved performance across various applications.

Audio-to-Text Retrieval (ATR)

Audio-to-Text Retrieval (ATR) involves finding the textual descriptions corresponding to a given audio query. This task is challenging due to the inherent modality gap between audio and text modalities, which requires models to align these distinct data

types. Recent research has emphasized the development of cross-modal alignment techniques to bridge this gap effectively [295, 274].

Contrastive learning has emerged as a key approach in ATR. Contrastive learning trains models to maximize similarity between paired audio-text samples while minimizing similarity with unrelated pairs. For example, Wu et al. [527] employed a CLIP-inspired contrastive learning framework to align audio and text embeddings in a shared space. Similarly, Deshmukh et al. [89] introduced a framework that combines two audio encoders with a text encoder and contrastive learning to improve the retrieval performance of variable-length audio content.

Audio-Question Answering (AQA)

Audio-Question Answering (AQA) is a multimodal task in which a system receives audio input and a related textual question to generate accurate answers about the audio content. This requires integrating audio signal processing with natural language understanding to produce contextually relevant responses. Advances in this field have been driven by attention-based neural network models that use self-attention to create representations for both audio and text. Cross-attention mechanisms map relevant audio features to textual queries, improving precision [241, 432].

AQA datasets differ from typical audio-language datasets using a triplet structure: an audio file, a text-based question, and a corresponding text-based answer. This setup supports interactive models beyond understanding or transcribing audio to interpreting and answering specific queries. The question and answer components are textual, with answers often structured for classification tasks, such as simple yes/no responses. For analytical purposes, questions and answers are sometimes concatenated into a single text string (see Table 2.1).

A sub-task of AQA is Spatial Sound Question Answering, introduced by Zheng et al. [591]. This variation aims to spatially detect and localize sounds by determining their direction of arrival and distance, adding a spatial reasoning dimension to the task.

2.2.2 Language-to-Audio

The Language-to-Audio category encompasses tasks in which language input generates, retrieves, or modifies audio. This category is divided into four fields:

1. **Audio Generation:** The creation of audio based on natural language prompts.
2. **Text-To-Audio Retrieval:** Retrieving audio clips from a database using natural language queries.
3. **Audio Editing:** The modification of existing audio content based on language instructions.

-
4. **Language-Queried Audio Source Separation:** The isolation of specific sounds from a mixture using natural language queries.

Audio Generation (AG)

Audio generation, also called Audio Synthesis or Text-To-Audio (TTA), focuses on producing synthetic audio that can mimic sounds, speech, or music. These models accept text input (prompts) and generate various audio outputs. The approaches to audio generation include:

1. Language-model-based methods: Autoregressive models trained on text-audio pairs, such as AudioGen [219] and AudioLM [41], generate audio directly from textual prompts.
2. Diffusion-based models: Latent diffusion models, such as Diffsound [562] and AudioLDM [256], generate audio from scratch or conditionally based on input text or audio. These have become the dominant approach in audio generation [25].
3. Consistency models: Models such as AudioLCM [258] synthesize audio with high efficiency, requiring only two iterations.

Text-to-Audio Retrieval (TAR)

Text-to-audio retrieval (TAR), or language-based audio retrieval, involves searching for and retrieving audio clips that match a given text query [207]. This task parallels Audio-to-Text Retrieval, focusing on mapping audio and text into a shared embedding space for cross-modal retrieval. Approaches include contrastive learning [527, 229], and training objectives such as triplet loss [534, 361]. Subtasks include Audio-Text Relevance Learning, where the goal is to learn the shared semantic properties of audio samples and text descriptions, using similarity functions [533], and Audio Moment Retrieval [319], where the models identify specific timestamps in audio that correspond to a text query.

Audio Editing

Audio editing focuses on altering existing audio content according to language instructions. Recent advancements include training-free methods, like AudioEditor [186], which repurpose pretrained text-to-audio models using techniques such as null-text inversion and EOT suppression, and specifically trained models, like AUDIT [491], which employ triplet training data and targeted modification learning. These

methods enable tasks such as adding background sounds, replacing instruments, and repairing audio while maintaining the integrity of unedited sections.

Language-queried audio source separation (LASS)

Language-queried audio source separation (LASS) uses natural language descriptions to isolate specific sound sources from a mixture of sounds. This field is also known as Text-to-Audio Grounding [550, 551], Language-assisted Sound Event Detection (SED-LM) [495], Language-Queried Target Sound Extraction [547], Query-based Sound Separation (QSS) [263], and Query-conditioned target sound extraction [278].

The prominent models in this field include LASSNet [264], SoundWords [198], CLIPSep [98], CLAPSep [278], AudioSep [263] and FlowSep [574]. In 2024, LASS was introduced as a new challenge in the DCASE competition; the winner was AudioSep-DP [569].

2.2.3 Datasets

Audio captioning datasets can be categorized into several groups based on their origins:

- **YouTube-Based:** Includes datasets derived from YouTube, such as AudioSet and AudioCaps, which originate from AudioSet.
- **Freesound-Based:** Clotho being a primary example, these are datasets from the Freesound database.
- **Sound-Effect Datasets:** Examples include WavText5K and the BBC Sound Effects Library.
- **Field-recording Datasets:** These datasets are sourced from field recordings, such as the MACS dataset. Field recordings are audio samples captured in natural environments, such as urban, rural, or natural settings.
- **Movies & TV Shows Datasets:** These datasets are sourced from television shows, such as the AudioCaption dataset.

Table 2.1 presents a summary of datasets relevant to the field of audio-language learning. This subsection contains a brief description of each dataset. The table in Appendix 2.3 contains example captions for each dataset discussed here. Moreover, Table 2.2 illustrates the combinations of datasets used to train large-scale audio-language models.

Table 2.1: Comparative overview of 69 datasets in the audio-language domain. The statistics are self-calculated where possible. Indentation indicates reuse of the audio or that the dataset is a subset of another dataset. Easily accessible (Acc.) datasets are publicly available and do not require special access, scraping, payment or complicated access procedures. LLM Generated Captions (LLM Gen.) indicate whether the dataset contains captions generated by LLMs such as ChatGPT.

Number of Captions	Avg Chars ($\pm$ SD)	Unique Words	Number of Sounds	Total Hours	Size (GB)	Acc.	LLM Gen.
YouTube-based Datasets							
AudioSet [126] — <i>Sound vocabulary based on an ontology</i>							
2,084,317	19.97 $\pm$ 18.92	1,067	1,955,135	5,383.11	1,155.23	✗	✗
AudioSet Strong [170] — <i>Improved temporal precision of labels</i>							
1,074,359	14.7 $\pm$ 8.88	645	120,080	329.23	44.78	✗	✗
AF-AudioSet [214] — <i>Large scale with synthetic captions</i>							
696,079	143.65 $\pm$ 57.58	18,498	331,421	912.51	195.83	✗	✓
ACalt4 [331] — <i>Provides four captions per audio</i>							
167,140	63.7 $\pm$ 29.89	12,338	30,116	82.51	17.71	✗	✓
EzAudioCaps [154] — <i>Diverse synthetic captions</i>							
1,088,808	81.43 $\pm$ 38.8	28,976	849,816	2,344.60	472.99	✗	✓
AudioSet Dialogues [135] — <i>Dialogues for audio understanding</i>							
172,566			78,084			✗	✓
AudioTime [540] — <i>Temporally aligned annotations</i>							
22,000	77.73 $\pm$ 27.62	13,584	22,000	59.18	12.70	✓	✓
AudioSetMix [543] — <i>Audio-language learning with modifiers</i>							
49,971	20.00		49,971		138.81	✗	✓
Sound-VECaps_A [573] — <i>Improved performance on complex prompts</i>							
1,657,029	180.72 $\pm$ 46.87	141,558	1,645,920	4,531.74	972.52	✗	✓
DAQA [110] — <i>Synthetic dataset for temporal reasoning</i>							
598,878	77.24 $\pm$ 19.72	271	100,000	2,089.36	224.17	✗	✗
CompA-R [131] — <i>Synthetic dataset for complex reasoning</i>							
200,205	284.32 $\pm$ 58.05	49,039	62,617	171.98	36.91	✓	✓
MULTIS [589] — <i>Dialogue dataset for instruction tuning</i>							
8,835	116.18 $\pm$ 48.01	5,156	7,645	21.07	4.52	✓	✓
SpatialSoundQA [591] — <i>Spatial reasoning for 3D sound</i>							
1,749,114	100.17 $\pm$ 30.25	1,639	1,749,114	4,858.65	2,085.18	✗	✓
AudioCaps [199] — <i>First large-scale captioning</i>							
57,188	49.22 $\pm$ 24.09	6,623	36,706	100.55	21.58	✗	✗
Audio-ALPACA [283] — <i>Winner and loser audios</i>							
15,025	56.94 $\pm$ 24.96	2,795	15,025	42.75	4.59	✓	✓
AudioHallucination [221] — <i>Object hallucination in ALMs</i>							
92,643	51.4 $\pm$ 5.71	1,113	686	1.88	0.40	✗	✗
BATON [250] — <i>Human preference feedback for TTA</i>							
4,509	67.59 $\pm$ 15.36	1,083	5,990	17.04	1.69	✓	✓

Continued on next page

Number of Captions	Avg Chars ($\pm$ SD)	Unique Words	Number of Sounds	Total Hours	Size (GB)	Acc.	LLM Gen.
TextToAudioGrounding [550] — <i>Fine-grained temporal labels</i>							
4,750	52.27 $\pm$ 24.33	1,715	4,588	12.71	2.73	✓	✗
AQUALLM [33] — <i>Large, high-quality, diverse AQA datasets</i>							
1,435,785			61,167			✗	✓
VGGSound [57] — <i>Large, audiovisual correspondence focus</i>							
199,467	16.52 $\pm$ 5.49	475	197,956	548.63	75.38	✓	✗
VGGSound-Caption [564] — <i>Language prompts for Video to Audio</i>							
178,730			178,730	496.47		✗	✓
Auto-ACD [435] — <i>Diverse with environmental context</i>							
1,922,878	103.97 $\pm$ 17.33	43,802	1,922,878	5,245.78	1,239.19	✗	✓
FAVDBench Audio [412] — <i>Diverse life-related scenes</i>							
9,643	78.56 $\pm$ 46.75	6,433	10,000	21.55	13.87	✓	✗
Freesound-based Datasets							
Freesound [117] — <i>Large, diverse, and expandable</i>							
1,018,339	43.81 $\pm$ 43.17	372,319	515,580	3,003.35	645.89	✓	✗
FSD50K Title+Desc [114] — <i>Open, large vocabulary, expandable</i>							
51,197	224.22 $\pm$ 277.64	86,349	51,197	108.59	32.12	✓	✗
AudioDiffCaps [441] — <i>Differences between audios</i>							
37,488	41.11 $\pm$ 18.16	973	7,756	21.43	12.68	✗	✗
Clotho [102] — <i>Focus on diversity and caption ambiguity</i>							
29,645	62.84 $\pm$ 15.72	8,168	6,972	43.64	12.91	✓	✗
Revised-Clotho [530] — <i>Revision of inaccurate captions</i>							
29,645		8,168	6,972	43.64	12.91	✗	✓
Clotho-Detail [587] — <i>Expressive and descriptive captions</i>							
3,939	324.83 $\pm$ 48.08	4,407	3,939	24.55	7.26	✓	✓
ClothoV2_GPT [367] — <i>Augments ClothoV2 with keywords</i>							
95,981	52.61 $\pm$ 14.86	23,676	3,839	23.99	7.09	✓	✓
Clotho-ChatGPT-Mixup [523] — <i>Increased complexity and diversity</i>							
50,000	100.39 $\pm$ 20.08	10,308	50,000	328.82	35.28	✓	✓
Clotho-Moment [319] — <i>Retrieves specific moments from long audio</i>							
44,261	13.33 $\pm$ 2.79	7,808	49,842	178.25	830.7	✓	✗
Clotho Entailment [90] — <i>Evaluate deductive reasoning</i>							
17,787	82.48 $\pm$ 19.58	7,745	5,929	37.06	10.96	✓	✓
ClothoAQA [254] — <i>Crowdsourced question answering</i>							
35,838	35.40 $\pm$ 11.97	4,407	1,991	12.46	3.68	✓	✗
mClothoAQA [34] — <i>Multilingual audio question answering</i>							
203,082	33.63 $\pm$ 11.49	27,522	1,991	12.46	3.68	✓	✗
NLAR [249] — <i>Natural language audio reasoning</i>							
200			200	1.2		✗	✓
WavCaps [296] — <i>Weakly-labeled GPT-generated captioning</i>							

Continued on next page

Number of Captions	Avg Chars ($\pm$ SD)	Unique Words	Number of Sounds	Total Hours	Size (GB)	Acc.	LLM Gen.
403,050	44.87 $\pm$ 21.18	43,802	403,050	1,119.58	264.41	✓	✓
AudioCondition [150] — <i>With text and control conditions</i>							
92,065	351.02 $\pm$ 275.91	43,802	92,065	255.74	60.37	✗	✓
ESC50 [364] — <i>With human accuracy estimates</i>							
2,000	9.04 $\pm$ 3.88	50	2,000	2.78	0.82	✓	✗
SynCaps [343] — <i>Synthetic, focused on temporal cues</i>							
5,370	39.55 $\pm$ 5.16	101	5,370	14.92	4.41	✓	✓
PicoAudio [537] — <i>Precise temporal control of audio events</i>							
12,400	40.42 $\pm$ 21.46	12,414	5,600	13.93	2.99	✓	✓
RDATS [553] — <i>Controls and describes detailed audio aspects</i>							
1,966	13.66 $\pm$ 4.38	834	2,000	6.86	1.47	✗	✓
SoundScaper [176] — <i>Acoustic, event, and affective information</i>							
25,440			25,440	212	125.38	✗	✓
DCASE LASS [264, 263] — <i>Dataset for DCASE Task on LASS</i>							
60,397	68.84 $\pm$ 22.71	8,204	58,397	128.59	34.26	✓	✗
Sound Effect Datasets							
LAION-Audio-630k [525] — <i>Large dataset for CLAP training</i>							
1,213,235	45.56 $\pm$ 43.71	391,541	391,541	4,017	867.07	✗	✓
Audiostock [18] — <i>Royalty free sound effects</i>							
10,001	35.36 $\pm$ 14.59	6,710	10,001	20.89	5.79	✓	✗
Epidemic Sound [45] — <i>Large catalog of sound effects</i>							
151,290	56.33 $\pm$ 47.83	16,750	75,645	220.41	60.10	✓	✓
Free To Use Sounds [119] — <i>Sound effects from field recordings</i>							
6,370	93.18 $\pm$ 45.33	4,979	6,370	175.73	50.69	✓	✗
ODEON (Paramount) [338] — <i>Cinematic sound effects</i>							
5,212	10.98 $\pm$ 4.38	3,709	5,212	20.01	2.03	✓	✗
Sonniss Game Effects [119] — <i>Sound effects for Game Audio</i>							
5,561	42.37 $\pm$ 22.92	7,854	5,561	84.61	22.50	✓	✗
We Sound Effects [174] — <i>Defunct sound effect library</i>							
5,561	41.40 $\pm$ 21.58	7,854	5,561	84.61	22.50	✓	✗
WavText5K [89] — <i>Improves audio-retrieval performance</i>							
4,348	73.00 $\pm$ 49.01	8,096	4,525	25.48	14.43	✓	✗
BigSoundBank [404] — <i>Including ambiences and soundscapes</i>							
2,290	51.59 $\pm$ 34.38	3,459	2,467	19.34	11.03	✗	✗
SoundBible [119] — <i>Free sound effects for multimedia projects</i>							
1,232	31.95 $\pm$ 12.28	1,827	1,320	4.74	0.55	✓	✓
BBC Sound Effects [31] — <i>Sound effects from TV and Radio</i>							
15,968	56.44 $\pm$ 27.72	13,278	15,973	479.99	77.25	✓	✗
SoundDescs [207] — <i>Large vocabulary, varied durations</i>							
32,979	104.75 $\pm$ 52.23	26,362	32,836	1,056.63	621.77	✗	✗

Continued on next page

Number of Captions	Avg Chars ($\pm$ SD)	Unique Words	Number of Sounds	Total Hours	Size (GB)	Acc.	LLM Gen.
Adobe Audition SFX [5] — <i>Uncompressed royalty-free sound effects</i>							
9,524	45.07 $\pm$ 13.19	2,871	9,528	22.42	14.62	✓	✗
Sound Jay [120] — <i>Free and royalty-free sound effects</i>							
2,239	32.67 $\pm$ 13.28	1,028	1,139	3.57	2.72	✓	✗
Zapsplat [100] — <i>Professionally recorded sound effects</i>							
64,421	8.51 $\pm$ 2.97	11,040	64,421	117.67	27.78	✗	✗
Pro Sound Effects [366] — <i>Large commercial sound effect library</i>							
430,434	74.67 $\pm$ 43.77	51,182	430,434	4,470.56	5,964.95	✗	✗
Field Recording Datasets							
MACS [288, 313] — <i>Metric-based reliable ground truth</i>							
17,275	54.26 $\pm$ 21.65	2,775	14,400	10.92	39.56	✓	✗
AudioLog [24] — <i>Temporal information of long audio</i>							
7,909			7,909	33.66	29.87	✗	✓
AnimalSpeak [383] — <i>Large bioacoustic dataset</i>							
895,190	39.38 $\pm$ 34.81	42,968	878,359	11,841.12	1,131.02	✗	✓
SoundingEarth [164] — <i>Urban sounds for remote sensing</i>							
51,058	220.96 $\pm$ 195.34	156,983	51,078	3,453.36	400.69	✗	✗
AudioEgoVLP [344] — <i>Audio from egocentric viewpoints</i>							
8,954	66.76 $\pm$ 23.63	4.23	3,346	12.35	3.65	✓	✓
MIMII-Change [463] — <i>Changes between two audio samples</i>							
17,999			17,999	50	5.36	✗	✗
CAPTDURE [340] — <i>Single-source sound captions</i>							
8,034	72.96 $\pm$ 39.70	6,195	2,088	4.90	2.67	✓	✗
Movies & TV Show Datasets							
AudioCaption [522, 549] — <i>Specifically for Mandarin Chinese</i>							
29,117	92.10 $\pm$ 41.89	8,542	7,336	21.43	12.96	✓	✗
SAM-S [163] — <i>Human-centric and movie-specific sounds</i>							
116,000			110,000			✗	✓

2.2.4 YouTube-based datasets

YouTube-based datasets, such as AudioSet and VGGSound, have been instrumental in the development of audio captioning models. However, using these datasets comes with specific challenges due to YouTube’s policies. Content creators can license their videos under the Creative Commons CC BY license, permitting reuse. Despite this, YouTube’s Terms of Service restrict content download unless explicitly authorized or permitted by YouTube.

As a result, researchers cannot directly share audio files. Instead, YouTube-derived datasets include only video links with the corresponding start and end timestamps. Researchers also provide scripts for scraping these videos. Consequently, anyone looking to replicate the models must scrape the data themselves.

A further complication arises from *link rot*, where video owners may delete their videos, or YouTube may remove them for policy violations. This leads to the unavailability of some portions of the dataset, which complicates efforts to reproduce research results or train new models.

AudioSet

Launched by Google in 2017 [126], AudioSet is one of the first large-scale datasets for audio classification, containing 2,085,544 audio clips, each approximately 10 seconds in duration. AudioSet features 527 categories of audio events, each manually annotated using Amazon Mechanical Turk, providing multi-label annotations. For creating audio-language models, various researchers concatenate the audio event labels together to form captions.

The categories are organized into an ontology with top-level labels such as Human sounds, Source-ambiguous sounds, Animals, Sounds of things, Music, Natural sounds, and Environmental and background sounds. AudioSet is divided into balanced and unbalanced training sets, with the balanced set comprising about 1% of the total and designed to ensure at least 59 examples per category. The unbalanced set contains the majority of the examples, particularly dominated by categories such as music and speech.

Due to the transient nature of YouTube content, the exact composition of AudioSet varies over time, as videos are sometimes removed from the platform. As of March 2024, about 18% of AudioSet samples are no longer available [464]. This variation can affect the size of the dataset used in different studies by up to $\pm 5\%$, potentially influencing outcomes and reported metrics such as mean average precision (mAP) [243].

Several other derivative datasets have been created from AudioSet. These include the following:

AudioSet Strong [170], which is a direct derivative, provides enhanced annotations with *strong* labels that indicate the exact start and end times of each sound event within the clips. This allows for more precise training on sound event detection tasks. However, it covers only 356 of the original 527 AudioSet classes. Also, these datasets are not necessarily effective for audio-language learning since the dataset only provides specific sound events per start-end time, instead of full sentences.

AF-AudioSet [214] is a synthetic audio captioning dataset based on AudioSet produced with a data labeling pipeline. The pipeline includes the audio-language model Audio Flamingo [213] that is prompted with the audio clips and the instruction “Can you briefly describe what you hear in this audio?” to generate 20 captions per audio with sampling using $\text{top-}k = 50$ and $\text{top-}p = 95\%$. Afterwards, CLAP [525] is used between the audio and the caption to rank and filter the caption, and the top three

most correlated captions are selected, removing captions having similarities below 35%. Although this approach provides multiple high-quality synthetic captions per audio clip, their synthetic nature implies that they could potentially contain biases from the audio-language model.

AudioCaps Alternative 4 Captions (ACalt4) [331] is a dataset that uses LLMs to generate captions similarly to Auto-ACD (see 2.2.4), where a frame is taken from the video of the AudioSet and VGGSound datasets, and an image captioning algorithm generates the image caption [435]. A ChatGPT LLM then generates a new caption given the image caption and AudioSet labels. Although this approach generates four different captions per audio using both visual and audio information, the dataset's reliance on video frames for caption generation may introduce irrelevant visual biases.

EzAudioCaps originates from four datasets: AudioSet, AudioSet Strong, VGGSound, and AudioCaps; part of the dataset is based on Auto-ACD. A large part of EzAudioCaps includes two synthetically generated datasets AS-Qwen-Caps and AS-SL-GPT4-Caps based on AudioSet [154]. AS-Qwen-Caps is a dataset with captions generated using Qwen-Audio [73], one of the leading audio-language models. AS-SL-GPT4-Caps is a dataset generated using GPT-4o-mini, based on the temporally-resolved annotations from AudioSet Strong. It is a large-scale, albeit noisy, dataset that combines multiple other larger datasets.

AudioSet Dialogues [135] is a dialog-based dataset of multi-turn sentences and comparison question-answer pairs for general sounds, although with low granularity. It is generated using a prompting-based approach with GPT-4 and its source data was from AudioSet. The dataset contains samples with one to four dialogs covering general sounds and focuses on creating strong correlations between rounds of dialog through the presence of pronouns, follow-up questions, and complex contexts. The paper also describes a dataset that covers the music domain.

AudioTime [540] is a dataset for the generation of temporal-controlled text audio. AudioTime includes timestamps, duration, frequency, and ordering information to help control audio generation. The dataset filters AudioSet Strong clips using CLAP similarity scores greater than 0.3 for segmented sound events. The TextToAudio-Grounding model [550] removes segments with scores below 0.6 when using event labels as queries.

AudioSetMix [543] is an audio capture dataset generated by applying audio transformations (speed, pitch, volume, duration, mixing, and concatenation) to several clips together from the AudioSet Strong dataset, combined with an LLM to form a weakly annotated caption. In particular, the dataset puts emphasis on the ordering, and the LLM is prompted to put the temporal information of the audio transformations into the caption.

Sound-VECaps [573], and its derivative **AudioCaps-Enhanced**, are AudioSet-based datasets that provide rich captions with audio event orders, locations, and environ-

mental information to improve audio generation models. The captions are generated using a visual captioning model named CogVLM [486] that processes each second of video, although this visual-based approach increases data complexity during inference. Llama3 [142] is prompted together with an audio caption using EnCLAP [202] and its AudioSet label. The Sound-VECaps provides two different categories of datasets. Sound-VECaps_F describes full detailed information, including visual features, for example, texts, names, and colors. Although Sound-VECaps_A removes visual-only information and contains only audible content or environmental-descriptive information.

DAQA: The Diagnostic Audio Question Answering (DAQA) dataset [110] by Meta (Facebook), provides a controlled environment to evaluate temporal reasoning capabilities of machine learning models engaged in Audio Question Answering (AQA). This dataset comprises sequences synthesized from 400 distinct natural sound events, with programmatically generated questions and answers that minimize biases. The audio content is sourced primarily from the AudioSet database, along with recordings made specifically for this dataset. However, due to its synthetic nature, DAQA may not fully capture the complexity of real-world audio scenarios. Moreover, it requires a large amount of computational resources to generate it.

CompA-R: CompA-R is a large and diverse instruction-tuning dataset synthesized by GPT4 to enable complex reasoning in audio-language models on audio-related metadata [131]. Each sample in CompA-R includes multi-aspect information on the audio, such as event tags and temporal markers. Its testing dataset, CompA-R-test, is a human-annotated evaluation set built to assess the reasoning capabilities of audio-language models on audio question-answering tasks. This dataset serves as a benchmark to test models fine-tuned on CompA-R for their ability to infer detailed scene descriptions based on complex audio input. The dataset is based on AudioSet Strong.

MULTIS: Zhao et al. [589] introduced the Audio Conversation 10k as part of the MULTIS datasets, which focuses on multimodal instructional tasks. This particular dataset is formatted for instruction tuning and includes audio-language pairs aimed at enhancing the capabilities of the large language model in audio-question-answering scenarios. An example interaction in this dataset features a human asking “What is making the sound in the audio?” to which the model responds, “The sound in the audio is made by a waterfall pouring into a stream.” All audio clips in this dataset are sourced from AudioSet.

SpatialSoundQA [591] developed a spatial sound question-answering dataset for 3D audio understanding using natural language. The dataset provides a platform for evaluating spatial audio perception and reasoning tasks such as sound event detection and direction of arrival estimation. The data is originally from AudioSet and adapted to spatial audio using the SoundSpaces 2.0 simulator [55]. It is mainly used for spatial

question answering, and models trained on this dataset may not generalize well to non-spatial audio understanding tasks.

AudioCaps

AudioCaps is a research dataset on audio-language learning sourced from AudioSet [199]. Each audio clip is 10 seconds long. The dataset is widely used as a training set and as a reference in audio captioning research, though its captions tend to be simple and may not capture all audio nuances. Due to the YouTube source, the size of the dataset has decreased over time. The validation and test sets provide five captions per audio clip. The training set includes only one caption per audio. The captions of AudioCaps are created by workers on the Amazon Mechanical Turk platform.

A notable issue within AudioCaps is the presence of duplicate samples [42]. Using clustering techniques, the researchers identified 257 audio clips with repetitive elements. These duplicates include identical introductory audio effects. This redundancy could affect the training and performance evaluations of the models developed using the dataset [42].

Several other derivative datasets have been created from AudioCaps. These are the following:

Audio-ALPACA is a preference dataset based on AudioCaps used in TANGO 2 [283] to enhance audio generation. Provides preference data derived from text prompts and their corresponding audio samples, although the samples are not always strongly aligned to their text prompts. The dataset creation involves generating audio samples from text prompts using a pretrained Tango model. The dataset uses three distinct strategies to ensure diversity and complexity. The first strategy generates multiple audio samples with varying denoising steps from the AudioCaps captions. The second strategy introduces semantic changes to the original captions using GPT-4. This strategy creates audio from the original and altered captions to capture nuances in audio-text alignment. The third strategy modifies the sequence of events in the captions to test the temporal relationships in audio generation.

AudioHallucination [221] adapted AudioCaps to a question-answering task, providing a structured approach to assess object hallucination in audio-language models. It extracted objects from the captions using NLP techniques and formatted them into five prompts to benchmark audio-language models. The datasets are separated into 3 categories, based on the type of sampling strategy used to create the questions: random sampling, popular sampling, and adversarial sampling. Random sampling receives k objects that are not present in the current audio. Popular sampling selects the top- k most frequent objects. Adversarial sampling ranks all objects by their frequency of co-occurrence.

BATON [250] is a small dataset created to train an audio diffusion model. Although its size could limit generalizability, the dataset incorporates feedback from human preferences to align text-to-audio models with human judgments. The dataset construction began by selecting the 200 most frequent categories from AudioCaps and combining them in pairs and triplets. GPT-4 then generated sentences connecting these category combinations using conjunctions like “and” and “followed by”. Expert human annotators manually rephrased a subset of these generated sentences. In the final dataset, each caption is used to generate five different audio samples using the TANGO audio generation model [127]. The dataset is evenly split between human-annotated and LLM-generated captions.

TextToAudioGrounding[550] is a dataset specifically designed for Text-To-Audio Grounding, a task that aims to predict the onsets and offsets of sound events described by natural language (see Section 2.2.2). The dataset features caption-oriented sound event tagging for each audio file and provides segmentation timestamps for each identified sound event. It was created by initially extracting sound event phrases from AudioSet and subsequently refining them through manual phrase merging.

AQUALLM [33]: The AQUALLM dataset comprises three subsets: AQUALLM-AudioCaps, AQUALLM-Clotho, and AQUALLM-MACS. These subsets were developed using an automated framework designed to ensure breadth, quality, and diversity in the content provided. The framework consists of four modules. The first module generates an initial set of potential responses from the captions following predefined rules. The second module uses an LLM to formulate a question for each potential answer. The third module validates each question-answer pair using an LLM. Successful pairs are then coupled with the corresponding audio clip. The final step involves the Question Paraphrasing Module (QPM), which paraphrases the generated questions.

VGGSound

VGGSound is a large-scale dataset [57]. The dataset was created using a combination of computer vision and machine listening techniques to automatically identify and verify audio-visual correspondences in videos. Audio events cover various categories including human actions, animal sounds, musical instruments, vehicles, and natural phenomena.

VGGSound-Caption: VGGSound has been adapted in Audio-Language Learning to train an audio generation model with video and loudness map as input [564].

Auto-ACD

The Auto-ACD dataset [435] is a large-scale collection based on VGGSound [57] and AudioSet [126] that provides detailed captions, although it contains many noisy entries requiring filtering. Also, since the dataset relies on visual features, it may contain biases from the visual domain. This dataset improves the original keywords from these sources using rich audiovisual information processed through six different models. Specifically, four algorithms analyze the middle frame of the video as an image: (1) BLIP-2 for image captioning, (2) Grounding Dino for object detection, (3) OpenAI’s CLIP model, which categorizes images into 1000 ImageNet categories using the prompt “a photo of a label”, and (4) PlaceCNN for place recognition.

FAVDBench Audio

The FAVDBench dataset is a specialized dataset to support the fine-grained audible video description task (FAVD) [412]. This dataset includes video clips, each accompanied by a summary, 4-6 visual descriptions, and 1-2 audio-related descriptions. Although visual descriptions are primarily tailored for visual understanding and are not ideal for Audio-Language Learning, audio-related descriptions are specifically crafted to capture sounds and auditory events that are discernible without visual context. For this reason, audio-related descriptions are used in this survey, while visual descriptions are excluded.

These videos were sourced from YouTube and are available exclusively under the Creative Commons license. The dataset was manually annotated in Chinese by a team of 60 people and subsequently translated into English. The annotation process for each video involved one primary annotator and two reviewers. In addition, each sentence in the descriptions was required to comprise more than five Chinese characters.

The same group also released another dataset, under the name TAVGBench [285]. This dataset focuses on audiovisual understanding and contains visual cues.

2.2.5 Freesound-based datasets

Freesound is a widely used resource in the audio domain and hosts various audio files. On Freesound, users can upload sound clips they have created or found, attaching relevant metadata such as filenames, keywords, and descriptions Font Corbera et al. [117]. Although user-generated metadata can vary in quality and structure, often including extraneous information or featuring long length audio clips, raw data requires preprocessing to make it suitable for machine learning applications.

One of the main advantages of using Freesound-based datasets is the licensing. Freesound allows for the broad utilization of its sounds under various licenses, making it a valuable source for training datasets in audio-related machine learning tasks.

This section will explore several Freesound-derived datasets, detailing how they have been processed and used for specific research purposes.

Freesound

Various researchers have included almost all Freesound in their training data for audio-language models [525, 88, 132]. Although Freesound continues to grow through user uploads, its user-provided descriptions are often noisy and require extensive processing. For this reason, Freesound is often used during pretraining in a self-supervised fashion.

FSD50k

FSD50k [114] is primarily a dataset for the classification of sound events that is openly distributed and contains a large vocabulary of everyday sounds. Although it does not include captions, researchers have adapted it for the training of audio-language models [104]. This adaptation involves concatenating the title and description provided in the metadata for each audio file. Although this method generates long descriptions, it introduces a degree of noisiness due to the unstructured nature of the metadata. A limitation of the dataset is that its classes are unbalanced and that some classes have many more examples than others.

AudioDiffCaps

The AudioDiffCaps dataset [441] is specifically designed to describe differences between pairs of similar but distinct audio clips through human-annotated descriptions. Although the dataset effectively captures subtle differences between audio pairs, its clips are artificially synthesized using the Scaper library [395], which may not fully reflect real-world conditions. The pairs are created by mixing foreground event sounds with background noises from the FSD50K and ESC-50 datasets. Each audio clip within the pairs is precisely 10 seconds in length. This dataset complements the Audio Pair with Difference dataset introduced earlier by the same researchers [442].

Clotho

Clotho [102] is one of the most widely used datasets in the field of audio captioning. It features a training, validation, and evaluation split, each ranging from 15 to 30 seconds in length. Every audio file in Clotho is paired with multiple captions created using Amazon Mechanical Turk. Additionally, Clotho includes a test set of audio samples without accompanying captions hosted on a separate Zenodo page, specifically

designed for inference purposes. This dataset is also prominently used in the DCASE challenges for Automated Audio Captioning. Its advantage lies in that it contains five captions per training, validation, and evaluation sample.

A study by Xie et al. [532] involved a crowdsourcing task on Mechanical Turk to evaluate the precision of 200 Clotho captions. Participants were asked to score how well the sound content matched the given caption and other relevant captions not attributed to the audio. From this study, which collected 18,204 responses from 340 workers, it was found that approximately 60% of the audio content in Clotho matches accurately its given caption. This dataset is proposed as GrRel in subsequent work, including relevance rankings from 0 to 100 of the audio samples given a caption [533].

Several researchers have developed revised or extended versions of the Clotho dataset to address specific challenges and enhance its utility for training audio captioning models. Key developments include:

Revised-Clotho [530] : This dataset is a modified version of Clotho that improves caption accuracy through GPT-3.5-based validation. Although it addresses issues with captions that poorly represent audio content by evaluating and revising them based on audio file name tags, the automated revision process may introduce biases that affect the natural diversity of the original captions.

Clotho-Detail [587]: Is a small dataset designed for instruction tuning that provides expressive and descriptive captions for deep understanding of audio content. It was generated automatically using GPT-4, which consolidated the five original captions per audio into a single caption in a few-shot setting.

ClothoV2_GPT [367] : Using GPT-3.5, this version involves generating new captions for the Clotho dataset. This process used both the keywords and the original captions to produce five distinct captions for each audio file. While the additional metadata-augmented captions help reduce overfitting during fine-tuning, the GPT-generated captions only marginally improve performance since Clotho already contains diverse human-generated captions.

Clotho-ChatGPT-Mixup [523]: The winner of the DCASE Challenge 2023 on Audio Captioning uses ChatGPT to propose caption mix-ups, increasing the complexity and diversity of the training data. They provide ChatGPT two randomly sampled captions and ask it to come up with a mix of the two while keeping the number of words under 25. The audio is mixed by scaling the two waveforms to ensure that their relative root mean square energy is within ± 5 dB before adding them together [59].

Clotho-Moment [319] is a proposed dataset for Audio Moment Retrieval that enables large-scale training by simulating long recorded audio in the city. The audio and text pairs are generated by overlaying Clotho samples with the Walking Tour dataset [472] at randomly sampled intervals. Although this approach allows for diverse moment annotations and robust model evaluation, the simulated nature may not perfectly

capture real-world audio complexities. The resulting dataset consists of 1-minute-long samples.

Clotho Entailment (and AudioCaps Entailment) [90] are two datasets that are used to check whether a model can assess whether a text description (hypothesis) of audio content can be deduced from an audio recording (premise). The dataset has three options for this. Either entailment, neutral, or contradiction. For each of these three, a model should also provide its evidence.

ClothoAQA

ClothoAQA [254] is derived from the Clotho dataset to be specifically used to answer audio questions. The dataset integrates crowd-sourced annotations through Amazon Mechanical Turk. Each audio sample is accompanied by six questions and three corresponding answers, totaling 18 pairs of questions with answers. Each pair of question-answers per audio was annotated by a different worker. The questions are structured to vary in difficulty: two per audio file are ‘yes-no’ questions designed for binary responses, while the remaining four require a single-word answer, which could be a limiting factor.

ClothoAQA subsequently has 2 datasets that are based on it, named NLAR and mClothoAQA [254]:

NLAR [249] is a small dataset created for the task of natural language audio reasoning. It is based on ClothoAQA, by cleaning the dataset. NLAR uses ChatGPT-Turbo to summarize acoustic characteristics and uses ChatGPT to generate five question-answer pairs given pairs of audio and its description.

mClothoAQA [34] expands the audio question-answering framework of ClothoAQA [254] into a multilingual context. By providing audio-caption pairs in eight languages (English, French, Hindi, German, Spanish, Italian, Dutch, and Portuguese), it addresses the lack of multilingual datasets in the field. The dataset was constructed by automatically transcribing the English question-answering pairs of ClothoAQA into seven other languages using Google’s machine translation API, though this automated translation process may introduce language bias issues.

WavCaps

WavCaps [296] is a weakly labeled dataset that brings together audio clips from four diverse sources: Freesound, BBC Sound Effects, SoundBible, and AudioSet (strong labels), of which the largest portion of the sounds come from Freesound. The dataset is generated using ChatGPT from raw descriptions. Additionally, WavCaps includes both prefiltering and post-processing steps aimed at refining the dataset by eliminating undesirable data, which resulted in a high discard rate.

AudioCondition: [150] is a dataset for conditional audio generation with text and control conditions based on AudioSet Strong, with its caption from WavCaps. For each category of AudioSet Strong, a new caption, albeit limited to the category, is created with timestamps based on the information from the WavCaps caption. The dataset enables fine-grained control over audio generation by incorporating timestamp, pitch contour, and energy contour for precise and diverse audio synthesis. For audio-to-language models, this dataset is less suitable due to its unstructured text captions.

ESC-50

ESC-50 [364] is a well-balanced environmental sound classification dataset that contains 2000 5-second audio clips in 50 classes of common environmental sounds. The classes are organized into five major categories: animal sounds, natural soundscapes and water sounds, human non-speech sounds, interior/domestic sounds, and exterior/urban noises. Although the dataset provides high-quality labeled recordings with sound events in the foreground and limited background noise, its small size of only 40 clips per class limits its utility for complex machine learning models. The dataset was created by extracting clips from Freesound recordings and was manually annotated by multiple participants through the CrowdFlower crowdsourcing platform, achieving 81.3% human classification accuracy. Additionally, without longer textual descriptions, the dataset's short class labels make it less suitable for audio-language models requiring more detailed captions.

SynCaps [343] is a synthetically created and used dataset to fine-tune text-to-audio and audio-to-text retrieval. An LLM generates textual descriptions in the style of AudioCaps from ESC50 sound labels. The dataset provides uniform temporal cues in the captions by concatenating two sounds with arbitrarily selected temporal order using only future and past cues. This helps models learn temporal relationships in audio, although the temporal cues are limited.

PicoAudio

Picoaudio [537] is a dataset created to evaluate the temporal control performance of audio generation models. The dataset provides high-quality, temporally aligned audio-text data that enable precise control over timestamps and frequency in audio generation tasks. Each audio has either 1, 2 or 3 sound events, with the sound events originating from Freesound. This limited number of audio events could restrict its scope of application. The sound events and the on-set time are randomly assigned, with two times as many clips containing 1 or 2 sound events as 3 sound events.

RDATS

RichDetailAudioTextSimulation [553] consists of simulated audio-text pairs with rich details in four aspects: temporal relationship, loudness, speaker identity, and occurrence number. The data were generated using curated single-event sounds from Freesound, mixed and concatenated according to randomly sampled attributes, and paired with descriptions generated by ChatGPT to provide more detailed captions.

SoundSCaper

SoundScaper is a model capable of generating soundscape captions, given an audio clip [176]. It is composed of an audio classification model, capable of predicting the acoustic scene and events, together with its 8-dimensional perceived affective quality values (such as Pleasantness-Eventfulness, or Calmness-Vibrancy). A GPT3.5 LLM generates a soundscape caption based on these values. The data originate from ARAUS [345], which contains the Urban SoundScapes of the World dataset, but also Freesound and Xeno-Canto. It is used to generate soundscape captions, capturing acoustic environment information from multiple perspectives.

DCASE LASS

Language-Queried Audio Source Separation [264, 263] is a 2024 DCASE Challenge to separate arbitrary sound sources using textual descriptions of the desired source. Its training set is based on FSD50k. The DCASE-Synth and DCASE-Real datasets are for validation and evaluation purposes. The DCASE-Real dataset consists of audio clips from FreeSound with either 2 or 3 manually annotated captions per audio. The DCASE-Synth dataset consists of synthesized mixtures, with a text description and its corresponding target source.

2.2.6 Sound effect datasets

Sound effects (SFX), originally designed for multimedia enhancements, have found utility in the audio domain. These datasets are valuable for their primary purpose—adding audio layers to visual media such as comedy series or action movies—and for their structured captions, which facilitate quick and accurate retrieval of specific sounds used when searching for specific content on SFX search websites. This structured metadata is a key advantage for training audio models as it allows for use as a caption describing the sound in these datasets.

A notable subset of SFX is the Foley sounds [84, 11]. Foley artists use everyday objects to mimic the sounds associated with actions or objects in a scene. For example, the sound of breaking bones might be simulated by pulling celery sticks. This creative

process produces similar sound effects and provides data for training audio models to recognize and replicate these sounds. However, some Foley sounds could be sonically different sounds with the same textual representation [513]. This section explores popular datasets that are based on these SFX sounds.

LAION-Audio-630k

LAION-Audio-630k [525] is the training set on which LAION-CLAP is trained. With 633,526 audio-text pairs, this extensive dataset with varied sources enables improved model generalization across diverse audio tasks. The following datasets are part of LAION-Audio-630k, next to Freesound and BBC SFX:

Audiostock [18] is a repository of royalty-free high-quality sound effects with a subscription fee of \$29.99 per month.

Epidemic Sound [45], a sound effects catalog that offers royalty-free high-quality music and sound effects for videos. Although it offers a subscription for \$10.99 per month for personal use, some users have reported receiving copyright claims on platforms like YouTube despite having paid subscriptions [425].

Free To Use Sounds [286], an indie sound effect catalog offering a \$25 all in one bundle, provides royalty-free sounds recorded worldwide. However, like other Epidemic Sound, users have reported receiving copyright strikes despite having purchased licenses [273].

ODEON Sound Effects Library of Paramount Motion Sound Effects [338] provides an extensive library in 53 categories, although it primarily contains movie and TV effects. The library costs \$49.95.

Sonniss Game Effects [421]: A yearly release of free game effects for the Game Developers Conference in various bundles. Although the sound effects are free, they primarily target game audio, and thus have limited diversity. The LAION-Audio-630k dataset contains the data up to 2020.

We Sound Effects [174]: a defunct website that offered high-quality sound libraries from top sound designers. Can still be accessed through archive.org.

WavText5k

WavText5k [89], created by Microsoft, improves audio retrieval performance on benchmarks like AudioCaps and Clotho through its contrastive learning approach. The dataset does not share audio files directly but provides links to original sounds with detailed metadata. The captions come from free-form descriptions created by the

sound effect creators that can lack consistency compared to human-curated annotations. This may limit the robustness of the dataset for complex retrieval scenarios. Its audios originate from two sound effect libraries:

BigSoundBank [404]: a sound effect website by Joseph Sardin offering free, high-quality, royalty-free sounds across 82 categories and 834 subcategories. Some subcategories are in French.

SoundBible [119]: offers public domain sounds to download in both wav and mp3 format, most of them created by Daniel Simion. While each sound has a title, description, and keywords, the dataset is small and restrictive since it primarily targets video editors, movie scorers, and game designers.

BBC Sound Effects

BBC Sound Effects database [31] offers a collection of 33,066 audio clips spread over 23 categories, such as nature, transport, and machines. This library features historical sounds from the BBC Radiophonic Workshop, the Blitz in London, various BBC TV and radio productions, and recordings from the BBC Natural History Unit archive. The accompanying captions are noisy and begin with nouns that identify the sound source or scene, followed by more detailed descriptions, which are often incomplete.

SoundDescs [207] serves as a reference audio dataset. It contains various audio content and vocabulary in the descriptions. The wide variation in audio duration of the dataset, with 109 files exceeding 10 min, helps train models to handle audio of different lengths. The dataset is organized into training (70%), validation and test splits (15% each), covering 23 categories such as nature, clocks, and fire. Recent evaluations have identified data leakage in SoundDescs, including issues with duplicates, reprocessed recordings, and partial overlaps, and proposed a new split of the dataset to eliminate these issues [497]. As no LLM was used to generate the captions, this dataset still contains unstructured noisy captions.

Adobe Audition SFX

Adobe offers a dataset of uncompressed sound effects as part of Adobe Audition and Adobe Creative Cloud [5]. The files are grouped by type and compressed into ZIP archives for easy download. In total, the dataset provides various different sound effects in 24 categories of audio sounds. Its filenames are used as captions, which results in noisy descriptions. The dataset has been used for training audio generation models, both latent diffusion models [181] and latent consistency models [258]. However, the sound effects are only accessible to active Creative Cloud members, and they can be used royalty-free in projects.

Sound Jay

Sound Jay [120] is an easy-to-use website that offers a small but diverse collection of free sound effects, although its use is limited to personal purposes. The website groups the sounds into 10 main categories, each with subcategories. It offers sounds in both WAV and MP3 format and provides captions as well as keywords for each sound.

Zapsplat

Zapsplat offers an extensive library of 150K+ free sound effects on its website under a \$5 subscription, where a free version offers a limited 4 downloads per hour and requires attribution when using the sounds. Its sounds are divided into various categories and subcategories.

Pro Sound Effects

The Pro Sound Effects Library, a commercial resource with more than a million professionally curated sounds, was the first dataset to be used in an Automated Audio Captioning setting [101]. Although its complete bundle costs up to \$15,000, the library offers high-quality sounds spread over 501 categories. After an initial automatic curation process, the dataset was refined to 152,619 samples in [101]. In our analysis in 2.2, all the current website offers are described in Table 2.1.

Subsequent studies have used this dataset for specialized applications. Kilgour et al. [198] used 110,000 samples for the text-driven separation of sounds, a technique that improves the clarity and isolation of specific audio elements. Furthermore, Miao et al. [302] used samples from this library to advance research in bioacoustics detection of wildlife.

2.2.7 Field recording datasets

This section explores a range of audio-language datasets that were not derived from common sources such as YouTube, Freesound, or standard sound effects websites. Instead, these datasets are derived from field or studio recordings. These sources include MACS and AudioLog, both based on TUT Sound Events datasets. These sources include bioacoustics websites such as AnimalSpeak, where users have recorded sounds from birds in the wild. Adaptations of existing datasets from related domains, such as MACS. Content from other video platforms, including SAM-S and AudioCaption. Self-recorded datasets, such as CAPTDURE, and collections from other online sources, such as SoundingEarth.

MACS

MACS [288, 313], a human-annotated audio dataset, features 3,930 files, each annotated with an average of 4.4 captions. Created from the TAU Urban Acoustic Scenes 2019 Dataset [165], MACS focuses specifically on urban environments, encapsulating three distinct sound categories: airport, public square, and park sounds. Each audio file is precisely 10 seconds long, with captions averaging 9.247 words in length, contributed by 133 annotators.

AudioLog

AudioLog is a dataset for training audio-language models to generate long paragraphs of descriptions of events that occur in audio [24]. It effectively integrates acoustic scene classification and sound event detection with contrastive learning, though its ability to generalize may be limited due to relying on specific acoustic scene annotations. The dataset is generated by an LLM capable of summarizing temporal information in long audio sequences. It uses the MAESTRO dataset [298], the TUT Acoustic Scenes 2016&2017 and the TUT Sound Events 2016&2017. The TUT Acoustic Scenes 2017 contains audio within two classes, “beach” and “office”, which also exist in the TAU Urban Acoustic Scenes 2019 Dataset that MACS is based on.

AnimalSpeak

The AnimalSpeak dataset [383], is a collection of bioacoustic sounds aimed at improving bioacoustic research. This dataset was initially compiled using unstructured labels with the assistance of ChatGPT, and further refined through screening with a language model. Captions that did not pass the screening were revised and captioned again to ensure accuracy and relevance. By pairing over a million audio recordings with human language captions, it enables scalable analysis across diverse species, though its species distribution shows some bias toward North American and European species. Due to its many Latin names, it contains novel vocabulary that may not be present in other datasets.

AnimalSpeak consists of several subsets, each sourced from distinct databases:

- iNaturalist²: This platform offers a crowd-sourced collection of animal observations, contributing 300,267 samples to the dataset.
- Xeno-Canto³: Known for sharing wildlife sounds, particularly birds, and recently expanding to include grasshoppers and bats, this website adds 679,676 samples.

²<https://www.inaturalist.org/>

³<https://xeno-canto.org/>

- Watkins Marine Mammal Database⁴: An open-access repository of marine mammal recordings collected throughout the career of William Watkins, providing 15,567 samples.
- Animal Sound Archive⁵: Maintained by the Museum für Naturkunde Berlin, this database initially contributed 21,473 samples. Unfortunately, most of this subset was lost due to a cyberattack.
- AudioCaps: see Section 2.2.4.

Each audio clip in the dataset, except those from AudioCaps, includes two captions: one referencing the animal species by its common name and the other by its scientific (Latin) name. In total, AnimalSpeak contains 1,062,983 samples, making it one of the largest bioacoustic datasets available.

SoundingEarth

The SoundingEarth dataset [164] is a collection of 50,454 geotagged ambient sound recordings from the Radio Aporee project, collectively accumulating over 3,500 hours of audio. While the dataset aims to archive field recordings globally, its reliance on Radio Aporee introduces an uneven spatial distribution that clusters heavily in Europe and the U.S., potentially limiting its global generalizability.

Each audio file in the dataset is accompanied by the corresponding Google Earth images, enabling self-supervised audiovisual learning without manual annotation. This feature provides a robust foundation for tasks such as remote sensing and environmental context interpretation. The recordings vary in length, with a median duration of approximately 3 minutes, although some of the longest samples extend beyond 30 minutes, with noisy but manually annotated captions.

AudioEpicMIR & AudioEgoMCQ (AudioEgoVLP)

Both AudioEpicMIR and AudioEgoMCQ are datasets from a proposed benchmark on audio retrieval [344]. AudioEpicMIR is based on the EpicMIR [81] task, which is based on EpicKitchens [81]. AudioEgoMCQ is based on EgoMCQ [252], which is based on Ego4D [143]. Both provide audio descriptions from an egocentric perspective. The LLM GPT3.5 is used to generate audio captions given visual descriptions within a prompt that also contains a few examples and an introduction to the task. Videos that did not contain any audio were removed, and another LLM GPT4 is used to generate descriptions into 3 categories: low, medium, and high. These categories scale the importance of audio for visual tasks. For the analysis in Table 2.1, the high category

⁴<https://whoicf2.whoi.edu/science/B/whalesounds/index.cfm>

⁵<https://www.museumfuernaturkunde.berlin/en/research/animal-sound-archive>

is used for both datasets, to include only the audio that is very informative, and to combine the two datasets.

MIMII-Change

The MIMII-Change dataset [463] is a specialized collection derived from the MIMII-DG dataset, which focuses on identifying malfunctions in industrial machines. Designed to study audio changes such as pitch and volume variations, this dataset comprises 1,500 manually annotated pairs of sounds, though it is limited to only five different types of machine sound. For each type of sound, there are 300 pairs evenly divided between normal operating sounds and their anomalous counterparts. The unique aspect of this dataset is the use of onomatopoeia in the annotations, where the annotators describe the specific changes in sound between each pair. Consequently, each pair of sounds is accompanied by three distinct captions.

CAPTDURE

CAPTDURE [340] is a manually created dataset comprised of 1,044 audio recordings captured by researchers in 14 different environments, each annotated with 3-5 captions to provide a description of the sound. It consists of single-source sounds and multiple-source sounds, each with captions and appropriateness scores for each caption. The captions were originally written in Japanese and translated into English. The annotation was carried out by crowd workers who provided captions and participated in a quality control process by evaluating the captions submitted by their peers.

2.2.8 Movies & TV Sound Datasets

AudioCaption

The AudioCaption project encompasses two distinct datasets. Each dataset focuses on sounds from specific environments: one in a hospital setting and the other inside a car. This scene-specific focus helps capture the typical sounds of these environments. The datasets were originally annotated in Mandarin and later translated into English using automated tools, which may introduce some translation inaccuracies. Due to their shared origins and annotation methodologies, these datasets are often used together.

- The Hospital dataset [522]: these clips were sourced from Chinese video platforms Youku, Iqiyi, and Tencent, predominantly featuring content from TV shows. The annotation process involved first asking annotators four questions about the sounds they heard, the sources of these sounds, their characteristics, and the sound location. The responses to these questions formed the basis for

the three captions assigned to each audio clip, culminating in a total of 11,121 captions. The dataset is divided into 3,337 training clips and 371 test clips.

- The Car dataset [549]: each clip in this dataset is annotated with five captions, making a total of 18,010 captions. These captions were generated by five Mandarin speakers per audio clip.

SAM-S

The Subtitle-Aligned Movie Sounds (SAM-S) dataset [163] provides a scalable, semi-automatic approach to extracting sound-specific data from closed-caption transcripts of 430 Hollywood movies released between 2014 and 2018. Although its closed-caption-based tagging process misses many potentially valuable audio events, the dataset still contains 116,000 captions that were originally enclosed in brackets, a method used to denote sound descriptions in film subtitles. These captions provide detailed information about the sounds, their sources and the quality of the audio events depicted in the movies. Of the total captions, 21,000 were identified as unique. The breakdown of these captions reveals a difference in length: 1,500 consist of a single word, 11,000 contain two words, and the remaining captions extend to three words or more.

2.2.9 Datasets for Large Audio-Language models

Current state-of-the-art audio-language models are trained on combinations of multiple datasets, which are described in this section. Table 2.2 provides an overview of datasets used in the main audio-language models. The early SoundWords model [198] from Google used an exceptionally large training set of 50 million captions, while most subsequent models work with 1-5 million captions. The most widely adopted datasets are AudioCaps and its parent dataset AudioSet. Of the 27 audio-language models analyzed, 21 used AudioCaps. The next most popular dataset was Clotho [102], followed by various Freesound-derived datasets. Sound-effects libraries such as BBC SFX, Sonnis, and Audiostock are also frequently used in audio-language models for training.

Dataset	Audio-Language models																												Count
	Netease AAC [157]	SoundWords [198]	MS-CLAP [104]	AudioGen [219]	LAION-CLAP [525]	AudioLDM [256]	Make-An-Audio [182]	TangoPromptBank [127]	ONE-PEACE [483]	LTU (OpenAQA) [139]	Pengi [88]	CoDi [445]	Make-An-Audio 2 [181]	FALL-E [192]	AudioLDM 2 [257]	Uniaudio [561]	CompA [132]	CLARA [334]	SALMONN [444]	Auffusion [556]	Audio Flamingo [213]	Cacophony [603]	MINT [584]	AudioLCM [258]	GAMA [131]	AutoReCap-XL [155]	AUDIO-TPE [492]		
AudioCaps [199]	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	21	
AudioSet [126]			✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	18	
Clotho [102]			✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	14	
MACS [288, 313]			✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	12	
Freesound [117]	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	13	
FSD50K [114]			✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	12	
WavCaps [296]								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10	
BBC SFX [31]				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10	
WavText5k [89]							✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	9	
Sonniss [421]				✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	8	
VGGSound [57]				✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	8	
Audiostock [18]					✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	7	
Epidemic Sound [45]					✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	7	
ESC-50 [364]							✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	7	
Paramount SFX [338]				✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	7	
UrbanSound8K [394]							✓	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	6	
SoundBible [119]	✓								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	6	
Free To Use Sounds [286]				✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	5	
We Sound Effects [174]				✓	✓		✓							✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	5	
SoundDescs [207]									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	4	
Adobe Audition SFX [5]											✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2	
Clotho-AQA [254]											✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2	
CochlScene [185]											✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2	
TUT [165]													✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2	
Chime-Home [118]																					✓	✓	✓	✓	✓	✓	✓	1	
FindSounds [112]											✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1	
NonSpeech7K [380]																					✓	✓	✓	✓	✓	✓	✓	1	
Sonyc-UST [52]																					✓	✓	✓	✓	✓	✓	✓	1	
Zapsplat [100]	✓																											1	
Sound Jay [120]	✓																											1	
Pro Sound Effects [366]		✓																										1	
YouTube		✓																										1	
Total captions (M)*	0.1	50	0.1	0.6	3.3		1.2	2.4	5.6	3.4	3.5	0.9	1.9	2.6		0.6			0.4	0.4	5.9	3.9	3	0.9	5.5	0.8	0.1		
Total hours (K)*	145.6			4	4.3	3.4	8					3.7	5.7		2		6.5	1	2		1.3	3.7		8.8	0.1				

Table 2.2: The sound event datasets (sorted by count) that were used in some major audio-language models (sorted by date). At the bottom is the total amount of sound event data these models were trained on. *Some papers only indicate the total amount of captions in their data, others the total hours of audio the data contains. In some cases both.

2.3 Augmenting Audio-language datasets

Over the years, researchers have proposed methods to enhance audio data for different benefits. First, augmentation addresses data scarcity: the audio-language field lacks datasets [203, 568] (see the rest of this Section 2.2.3), which hinders the development and performance of retrieval models. Second, augmentation improves generalization by reducing overfitting [501, 577]. Lastly, augmentation can improve model performance, which is important because smaller models trained on high-quality textbook-like data often outperform their larger counterparts [147]. Examples of models trained on augmented datasets include BLAT [554], WavJourney [265], Audiogen [219], Make-An-Audio [182], and LTU [139]. The published datasets (see Section 2.2.3) with LLM-generated captions include LAION-630k [525] (Section 2.2.6), CompA-

661K [132] (Section 2.2.4), WavCaps [296] (Section 2.2.5) and Auto-ACD [435] (Section 2.2.4).

This discussion excludes training techniques such as negative sampling [535, 261], transfer learning [290], Test-Time Augmentation (TTA) [200], teacher forcing [567], and curriculum learning (see Chapter 3). Although these methods aid in generalization and performance, they primarily modify training procedures. For a detailed overview of training strategies in audio captioning, Xu et al. [552] is recommended.

This section focuses on augmentation techniques used in audio-language research. Augmentations fall into four categories:

1. Single-modal augmentation for audio, focusing on techniques that enhance generalization and performance.
2. Single-modal augmentation for text, aimed at generating additional data.
3. Multimodal augmentation, where both text and audio are augmented.
4. Augmentation with other modalities.

This section explores both “traditional” methods, such as MixUp [577] and SpecAugment [356], which do not rely on machine learning models, and contemporary proposals that use audio generation models or language models, with an emphasis on advanced language models like ChatGPT.

2.3.1 Audio augmentation

Audio augmentation improves performance in audio-language models. It applies transformations to audio signals to mimic different acoustic environments or to distort the original audio. Common techniques include adding reverb, clipping, masking, pitch modulation, and introducing environmental sounds. The aim is to generate various versions of audio signals that improve the robustness of the model. Wei et al. [501] provide a comparative analysis of augmentation methods for audio classification and introduce Mixed Frequency Masking, which achieves gains on the Freesound-based dataset FSDKaggle2018 [116]. This subsection discusses both traditional methods (Mixup, SpecAugment, and audio effects) and approaches that use audio generation models or language models.

Mix-based methods

Researchers widely use mix-based techniques in audio augmentation. Methods such as (Audio) Mixup [577] increase training diversity and yield performance gains in multiple studies [441, 216]. Mixing both language and audio relates to Audio Mixup

and is discussed in Section 2.3.3. Audio Mixup [577] linearly combines pairs of audio samples using a scaling factor, as described by the equation:

$$\hat{a} = \lambda a_1 + (1 - \lambda) a_2 \quad (2.1)$$

where λ is a scaling factor drawn from uniform [132, 458, 330] or beta [577, 227, 256, 138, 225] distributions with task-specific parameters.

Mixing extends beyond simple linear combinations. Mixstyle [594] manipulates the frequency axis of spectrograms to transfer device-style characteristics between recordings. It alters spectrogram statistics and increases performance by 5.2% over Mixup [594, 369].

Some implementations also set a Mixup rate: mixing a fixed percentage of audio samples during training [256]. Other variants perform mixing in log-mel spectrogram space by scaling inputs to linear scale before mixing and reconverting to log scale [330].

SpecAugment

SpecAugment [356] originated in automatic speech recognition and has been applied to audio captioning [58, 478, 325, 567, 218] and text-to-audio retrieval [292]. It randomly masks frequency and time regions of log-mel spectrograms and deforms along the time axis, improving robustness [517], reducing overfitting, and enhancing performance [58, 478, 58]. Some works reduce the dimensions of the mask during inference [200]. SpecAugment++ [479] adds zero-value masking and mini-batch mixture masking and further improves audio captioning [433].

Audio Effects

Adding audio effects such as noise and reverb before training diversifies data, improves generalization, and increases performance [561, 201]. Unlike SpecAugment, these effects precede conversion to log-mel spectrograms. Common practices include adding random Gaussian noise to the waveform to reduce overfitting [413] and using speed, reverb [157], pitch, and overdrive, often combined with Mixup (2.3.3) for gains in audio captioning [225]. Temporal and frequency changes promote time and frequency invariance, respectively [413]. Gain augmentation randomizes the volume to promote gain invariance [369]. Huang et al. [183] also show that adding noise during training improves generalization and robustness.

When training models with fixed input sizes (e.g. CNNs), some researchers split longer audio into consistent 10-second segments. This ensures compatibility [383]. Similarly, audio clips are randomly cut when they are longer than 10 seconds or repeated when

they are shorter than 10 seconds [525]. Another technique is average pooling, to downsample audio to a fixed size [478].

Audio Augmentation Using Audio Generation Models

The generation of audio for augmentation yields mixed results. Ghosh et al. [132] synthesized compositional audio with state-of-the-art text-to-audio models [265, 256, 127] to improve audio-language training, but found limitations in generating diverse and cohesive audio for acoustic scenarios and reported a gap in capability remaining in current models.

Xiao et al. [529] used AudioLDM to synthesize audio from captions in COCO [253]. Integrating these text-audio pairs benefited audio captioning and well-annotated text from the vision-language domain. Later work scaled datasets and used large language models like ChatGPT⁶ to generate pseudo captions, increasing the dataset and improving audio captioning performance [139, 435, 296].

2.3.2 Single Modal Text Augmentation

Text augmentation, like its audio counterpart, mitigates data scarcity, improves performance, and strengthens generalization. Researchers use heuristic text cleaning and, more recently, large language models. This subsection outlines traditional and LLM-based approaches to enrich text data for training audio models.

Traditional Text Augmentation

Text augmentation maintains audio integrity while diversifying text. Primus et al. [369] report that in traditional settings, text augmentation can have a greater impact than audio augmentation. This subsection covers common techniques and applications, starting with data cleaning and then using external data.

Many works adopt data cleaning: a common technique for converting text to lowercase and removing punctuation [367, 107, 433, 414]. This reduces the size of the vocabulary [548] to simplify learning in small datasets: instead of 52 letters, each letter in the capital or lower case, the model only has to predict 26. However, this also returns uncapitalized and unpunctuated outputs at the validation or inference phase.

Cleaning also includes removing specific words (e.g., keywords or descriptions), checking spelling against US or UK dictionaries, deleting repeated keywords [397] and descriptions [190], and applying heuristic rules to remove non-English words [476]. Other heuristics filter texts with numbers, stop words, and lemmatization [219], and

⁶<https://chatgpt.com/>

remove patterns such as “recorded of/from,” pronouns and words that do not describe the audio content [157]. Researchers often apply these techniques to web-scraped datasets to improve training quality [296, 603].

In audio retrieval and audio generation, filtering preserves a representative dataset with K-means selection [53]. Algorithms such as minHash [46] remove duplicates and improve retrieval efficiency without distorting the distribution [476].

The BLAT audio captioning model uses AudioSet sound event tags to guide caption generation and excludes audio clips with tags absent in AudioCaps, to align a new caption dataset with existing dataset standards [554].

Previous work assesses temporal relations in audio-language models. Wu et al. [520] filtered AudioCaps and Clotho to sentences containing temporal conjunctions such as “before,” “after,” “then,” and “followed by” to test event ordering. Xie et al. [539] used a similar approach, filtering for temporal conjunctions and excluding conjunction words indicating simultaneous events. Both approaches captured the temporal structure but degraded the standard captioning metrics, attributed to the simple heuristic composition [211]. Abdelnour et al. [1] focused on question generation for audio-language models, and generated questions with “before” and “after” temporal relationships.

Another category is word replacement. Traditional methods such as TF-IDF [424] replace or select words to help prevent overfitting while preserving the meaning of sentences [440], focusing on low-frequency words and topics [210, 536]. Cho et al. [67] used WordNet-based [305] synonym substitution, randomly replacing words with synonyms. The Easy Data Augmentation (EDA) strategy [500], synonym replacement, random insertion, random swap, and random deletion, addresses data shortage in audio retrieval [203] and prevents overfitting during fine-tuning [201].

Researchers also convert multi-label audio classifications into text captions. They either concatenate labels [414, 219], describe them with phrases like “this is the sound of” [352], or use their position within the AudioSet ontology hierarchy [190].

Textual augmentation with Large Language Models

Recent advances in LLMs such as GPT-3.5 [47] and ChatGPT improve textual augmentation for training audio-language models by creating diverse datasets. LLMs convert unstructured text or multi-label classification datasets into coherent sentences that are useful for training.

One application is caption mix-up, analogous to audio mix-up (Section 2.3.1) and audio-text mix-up (Section 2.3.3). It blends captions from two audio clips to produce a new grammatical caption. The winning entry in the 2023 DCASE Challenge on Audio Captioning [523] used this technique to increase the quantity and diversity of data. It

also improves the fluency and precision of the captions [529]. Ghosh et al. [132] show that LLMs can generate positive and negative audio-text pairs given acoustic events for modular contrastive learning.

LLMs also assess quality by comparing generated captions with human-written standards. Systems filter captions below a threshold to retain high-quality data [219]. Xiao et al. [530] use GPT-3.5 [47] to judge the precision of the caption against the audio filename tags and then use GPT-3.5 again to review the flagged captions. In bioacoustics, researchers use ChatGPT to generate, review, and correct captions [383].

LLMs are used to transform existing datasets. In audio question answering, LLMs generate contextually relevant question answer pairs using audio metadata [139]. For language-assisted detection of sound events, Wang et al. [481] used text augmentation to generate captions consisting of AudioSet events with start and end times. LLMs can also create new versions of audio captioning and retrieval datasets by rephrasing existing captions with available keywords [367] and by better matching audio files [530].

LLMs power machine-generated datasets such as Auto-ACD and WavCaps. Auto-ACD uses ChatGPT to process labels from multiple modalities in a structured format to generate datasets [435]. WavCaps (Section 2.2.5) uses a multistage approach: ChatGPT refines and redrafts audio captions to ensure precision and relevance, avoids named entities, and removes subjective content [296]. Kim et al. [203] use a similar prompting method to paraphrase captions with ChatGPT.

Bai et al. [24] compared four LLMs, ChatGPT [47], LLaMA 2 [460], Spark⁷, and ERNIE bot [436], to create paragraph-length descriptions of minutes-long audio for AudioLog (Section 2.2.7). They found all four effective and recommend ChatGPT for summaries of long audio that also describe detailed sound events.

Other ML models for text augmentation

Beyond LLMs such as GPT-3.5, LLaMA, and Gemini, other language modeling approaches augment text for audio-language applications. Encoder-decoder models like PEGASUS [578] and T5 [376] transform one textual representation into another and are often used for machine translation [68], but have also been used for audio-language tasks [83, 483].

PEGASUS [578] rephrases the captions [83]. T5 [376] augments datasets such as AudioSet and Epidemic Sound [483] and supports multi-stage pipelines. For example, a four-module pipeline uses T5 to generate questions from a caption and answer candidate, and a second T5 module generates similar questions [33].

⁷<https://global.xfyun.cn/products/SparkLLM>

Several studies use a teacher-student approach [172] for noisy captions [419, 494]. A teacher model trained on noisy and clean datasets identifies more reliable data; then, a student model trains on the refined data. Another technique is back-translation [407] that augments text by translating a sentence into a foreign language and then back to the original, introducing linguistic variation without altering the core meaning [369].

2.3.3 Audio-Text Multimodal Augmentation

This section discusses augmentation techniques that involve both text and audio.

Manual techniques

Huang et al. [182] addressed limited labeled data with a framework that generates descriptive captions aligned with audio content. Their approach has two stages. First, the system converts the audio into candidates using the AAC and ATR models. Next, a heuristic merges text-audio pairs with concepts from a sound event database, creating new training examples with different concept compositions using prepositions. Huang et al. [182] further diversify the training data by concatenating extra audio and text before or after the audio.

Noriy et al. [334] concatenated sound event labels with prefixes and metadata and used a language model to generate coherent sentences. Cousin et al. [78] expanded their dataset with DeepL translations⁸ to create captions in French, German, and Spanish. Ye et al. [568] introduced FeatureCut, which adaptively generates samples using Kullback-Leibler divergence to encourage audio models to learn from entire features.

PairMix

PairMix (also known as AL-MixGen, or Mixup with text), originally introduced in vision-language as MixGen [160], creates new training samples by performing mixup [577] on a pair of audio clips and combining their labels. Researchers apply PairMix to increase the size of the dataset for the training of audio models [83, 67, 203].

PairMix randomly selects N audio-text pairs containing audio clip α and text τ , then mixes each modality separately to generate a new audio-text pair. Implementations mix audio at the waveform level, in mel spectrograms, or by concatenation, with a set probability for each generated sample [200]. They then concatenate the corresponding texts with conjunctions [215, 227, 78]. Some methods mix in the embedding space [210, 218, 228].

⁸<https://www.deepl.com/>

PairMix increases performance when combined with other training techniques such as test-time augmentation [203, 200] and audio augmentation methods in 2.3.1. It also diversifies training data [201]. Kuzmin et al. [225] found that Mixup with simple concatenation of captions and audio outperformed random shuffling and mixing. However, Mixup and MixGen can degrade performance in text-to-audio tasks [256].

2.3.4 Augmentation with other modalities

Beyond text and audio augmentation, researchers incorporate other modalities to create data and context for audio-language datasets. AudioLog generates audio summaries based on audio labels with the corresponding start and end times [24]. Integrating visual features also improves performance in datasets such as AudioCaps (Section 2.2.4) [39].

Unified multimodal models such as UniVAL [417], Video-LLaMA [576], PandaGPT [431], MacawLLM [277], and Audio-Visual LLM [416] achieve strong performance in AudioCaps [199] and Clotho [102], even when trained in audiovisual datasets. These models also perform well on text-audio retrieval [322]. Aligning textual and visual features at the spatial and temporal levels further improves personalized text-to-audio performance [309].

To create an audio difference dataset, Komatsu et al. [211] superimposed Clotho and ESC50 [364] audio clips in the time domain, adding individual ESC50 sound events to Clotho scenes. Other studies show that visual features improve audio caption performance in AudioCaps, although relying solely on video performs poorly [262].

Zero-shot audio captioning, as explored by Shaharabany et al. [408], can rely exclusively on audio input. Their approach introduces an audibility term to ensure that captions reflect audible descriptions. Bridging high-quality audio datasets for multi-class tasks shows that datasets such as UrbanSound8K [394], ESC50 [364], and TUT17 [299] can also serve as captions [513].

VIP-ANT [588] uses images as a pivot modality, connecting audio and text without paired audio-text data. WavJourney [265] uses LLMs to generate structured audio scripts to enable training-free audio generation by issuing commands to multiple models.

Retrieval augmentation also helps. For audio generation, it achieves state-of-the-art performance in AudioCaps and excels at generating rare or unseen sound events [256]. In audio captioning, retrieval augmentation outperforms traditional encoder-decoder networks, whether using a data store with audio alone or in combination with text [129, 209].

2.4 Audio Dataset Evaluation

This section evaluates the primary audio datasets. First, it provides visualizations based on a large-scale principal component analysis of audio and text embeddings. These visualizations illustrate the relative distance between audio and text content in the datasets and their absolute numbers in terms of AudioSet top-level categories. Second, the section presents a quantitative evaluation of data leakage using the cosine similarity between CLAP embeddings of the audio and text data.

Given the large scale of the dataset, the calculation of CLAP embeddings, the calculation of the PCA in section 2.4.1 and the calculation of pairwise cosine similarity in section 2.4.2 presented computational challenges. The calculation of the CLAP embeddings using the `laion/larger_clap_general` model took approximately 2 days on a single H100 GPU. Across all datasets, 18,822,210 audio-text pairs were processed. After excluding datasets that directly overlap or are derived from other datasets (e.g., Clotho-Detail being based on Clotho), 6,903,631 audio-text pairs were retained. After experimentation, the Rapids cuML GPU implementation of PCA [452] was used to efficiently reduce dimensionality. Similarly, pairwise cosine similarity computations for the whole dataset were performed in batches using PyTorch on GPU [358].

2.4.1 Audio Dataset Analysis

Category Distribution

In the first set of visualizations (Fig. 2.3), audios and texts are categorized according to AudioSet’s top-200 categories. This categorization is used to highlight the relative distribution of sounds in terms of general semantics. For this purpose, each text corresponding to a sound was processed using Meta’s Llama 3.2 model (specifically, the *meta-llama/Llama-3.2-3B-Instruct* model hosted on HuggingFace). The model was prompted with each caption and instructed to classify it into exactly one AudioSet category, returning only the exact full category name. The prompts were processed in batches of 64 and took approximately a week on a single H100 GPU. Samples were omitted from the visualization that could not be matched to a valid AudioSet category, which occurred in 3.8% of the cases.

The visualization (Fig. 2.3) shows the top 200 categories grouped by the 7 root-level categories of the AudioSet ontology: “Natural sounds”, “Human sounds”, “Sounds of things”, “Source-ambiguous sounds”, “Animal”, “Channel, environment and background”, and “Music”. Among these root categories, only “Animal” and “Music” have their own direct class, while the others serve as grouping categories. The visualization is limited to the top 200 categories instead of 527 AudioSet categories to avoid visual clutter. As a consequence, categories with less than 2,264 samples were excluded. Although the LLM was given 200 categories to choose from in its classification task, it only returned a total of 177 unique categories in all classifications.

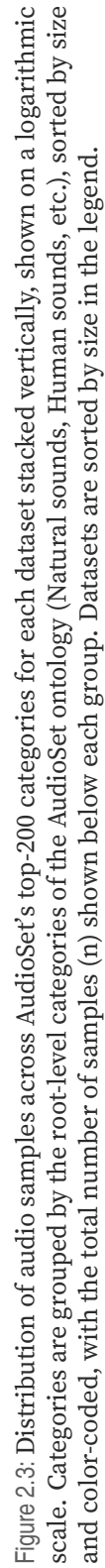


Figure 2.3: Distribution of audio samples across AudioSet’s top-200 categories for each dataset stacked vertically, shown on a logarithmic scale. Categories are grouped by the root-level categories of the AudioSet ontology (Natural sounds, Human sounds, etc.), sorted by size and color-coded, with the total number of samples (n) shown below each group. Datasets are sorted by size in the legend.

Comparing the model’s classifications to AudioSet’s original labels reveals some discrepancies. For example, “Male speech, man speaking” was classified 30.6 times more frequently by the model than it appears on AudioSet’s labels, “Sound effect” was classified 14.5 times more frequently, and “Inside, small room” 10.6 times more frequently. Some examples in the top 200 categories that are not included in the visualization are “Drum kit” with 15169 audios in AudioSet, “Keyboard (musical)” (10473) and “Motor vehicle (road)” (9044).

Based on the analysis of the category distributions in the audio datasets in Figure 2.3, imbalances are observed in the category representation. The most prominent categories show a strong bias toward indoor environments and human-generated sounds, with “Inside, small room” being heavily overrepresented at 21.6% (3.2M samples), followed by “Music” and “Male speech” at 14.1% and 13.6% respectively. These three categories alone account for nearly 50% of all samples, while specific sound categories like “Cluck” (147 samples), “Rimshot” (371 samples), and “Motorboat” (239 samples) are severely underrepresented.

The analysis also shows notable patterns in category concentration within individual datasets. For example, DAQA exhibits a high concentration of male speech (37.8% of its samples), while BBCSoundEffects shows a strong bias toward indoor environments (34.2% “Inside, small room” samples). In general, indoor environmental sounds and human-generated content dominate, while natural sounds and specific sound effects form a long tail in the distribution. Despite the “Animal” and “Bird” categories appearing in the top 10 most common categories (5.5% and 3.5% respectively), most nature-related and specific sound effect categories have minimal representation. The uniformity of the category distribution is also calculated using the formula:

$$H_n(\alpha) = \frac{-\sum_{i=1}^n p_i \log p_i}{\log n} \quad (2.2)$$

$$\text{where } p_i = \frac{\alpha_i}{\sum_{j=1}^n \alpha_j} \quad (2.3)$$

Where n is the number of categories, α_i represents the count in category i , and p_i is the probability of category i . The normalized entropy ranges from 0 (completely concentrated in one category) to 1 (uniformly distributed across categories). A higher value indicates a more uniform distribution of categories within a dataset, something a balanced dataset should have. In this calculation, it is seen that VGGSound shows the most uniform distribution (0.720), followed by FindSounds (0.615) and BATON (0.610). This suggests that these datasets have more balanced representation across their categories. However, the fact that these datasets have a limited number of unique words in their captions influences this observation as well. When examining samples per category, SpatialSoundQA leads with 12,382 samples per category on average, followed by AudioSet (11,773) and DAQA (10,968).

PCA Embeddings

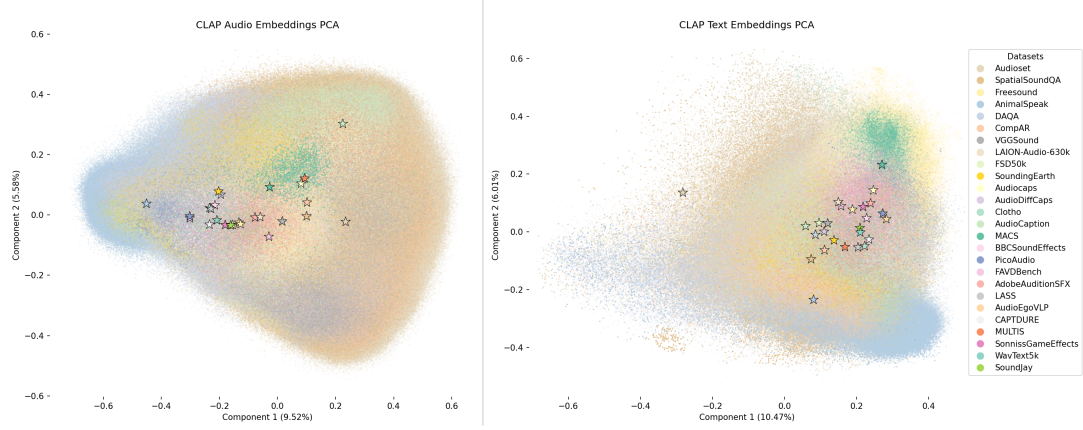


Figure 2.4: PCA decomposition of all the datasets on both the audio and the language. The stars in the visualizations represent the centroid of the dataset.

For the visualization of embeddings, only the original datasets are used (i.e. the ones without a parent dataset; see also the indentation of the datasets in Table 2.1), with a few exceptions. For example, ClothoAQA is based on Clotho and is therefore omitted from this analysis, since this dataset is already overlapping. Similarly, Auto-ACD, ESC50, and PicoAudio are also omitted. In addition, including all derivatives would make the visualizations too cluttered and difficult to interpret meaningfully. Datasets without available data, indicated with empty fields in Table 2.1, are also omitted from this analysis. In total, 26 datasets are analyzed; visualizations of CLAP embeddings are generated for both audio and text (Fig. 2.4).

This set of visualizations (Fig. 2.4) represents the first and second principal components of audio (left) and text (right) embeddings obtained using the *laion/larger-clap_general* CLAP model hosted on HuggingFace. In this set, the data points are color-coded based on the datasets (see the legend) to highlight the acoustic and linguistic variability of each dataset and their centroids. These centroids are indicated by star glyphs and calculated by averaging all CLAP embeddings for that dataset after reducing them to two dimensions. For quantitative variability analysis, distance scores are calculated based on the pairwise Euclidean distance between each individual point in the datasets in the original embedding space. For further analysis, see also Section 2.4.2.

In the audio embeddings PCA (Fig. 2.4, left), it can be observed that the centroid of the datasets AnimalSpeak, AudioCaption, MACS and AudioSet are the most distant from centroids of the other datasets. In the text embeddings PCA (right), the centroid of AudioSet is the most distant from the other centroids, which is likely due to the limited number of unique words in AudioSet’s multi-label classification captions. Notably,

the sounds of AnimalSpeak deviate in both audio and text embeddings, which may be because the dataset contains different animal sounds, which are, to a lesser extent, present in the other datasets.

In the original dimension, before conversion to 2 dimensions with PCA, Freesound (variability: 0.925), FSD50k (0.909) and WavText5k (0.906) show the highest variability in audio embeddings, while AnimalSpeak (0.699), AudioCaption (0.661) and MACS (0.534) show the lowest variability in audio embeddings. This variability is defined by the average distance for all embeddings from its centroid. For the text embeddings, Freesound (0.930), LAION-Audio-630k (0.927) and SonnisGameEffects (0.910) show the highest variability, while AudioCaption (0.688), AudioDiffCaps (0.674) and MACS (0.598) show the lowest variability. Freesound is a large collection of user-uploaded sounds, containing many different sounds, which may explain its high variability in both audio and text embeddings. MACS, on the other hand, represents sounds from three different locations (airport, public square, and park) and is therefore more homogeneous.

2.4.2 Duplicate analysis

Previous research has highlighted the issue of duplicate data in multimodal models, where pretraining datasets often include examples that the model is later expected to predict in a zero-shot context. This overlap can lead to an inadequate evaluation of the generalization capabilities of models [466]. Additionally, studies have shown that a portion of the zero-shot learning accuracy in audio-language models is attributable to the inherent capabilities of their audio and text backbone models, rather than a genuine ability to perform zero-shot learning [447].

Several studies in audio-language datasets have highlighted dataset overlaps, which are particularly relevant for zero-shot inference. For example, when training models on WavCaps [296], researchers excluded data from FreeSound and AudioSet to ensure zero-shot evaluation on Clotho and AudioCaps, due to overlapping samples. Similarly, in LAION-CLAP training [525], samples from various sound event datasets were removed to avoid overlap with Clotho and between the AudioCaps test set and the AudioSet training set. In the HEAR challenge [465], datasets such as FSD50k and Mridingham were excluded from the benchmark to prevent overlap with Freesound [603]. The DCASE challenge 2024 published a CSV file of forbidden Freesound sound IDs to avoid the overlap between Freesound and Clotho.

Duplicate analysis is inspired by two key papers on specific audio-language datasets. Bralios et al. [42] examined memorization in latent diffusion models, identifying numerous duplicated audio clips in the AudioCaps database using CLAP embeddings and mel-spectrogram analysis. Similarly, Weck et al. [497] analyzed the SoundDescs dataset [207], detecting duplicates or overlap recordings and producing a revised version without redundancy by applying an audio fingerprinting algorithm [420].

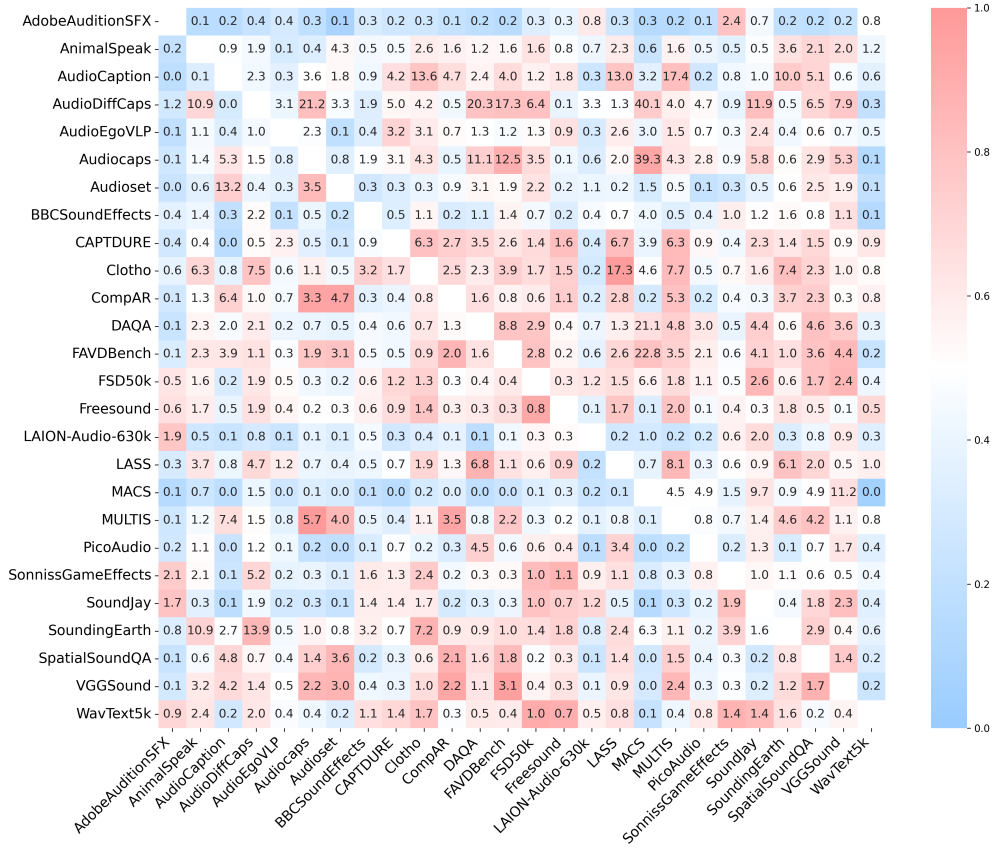


Figure 2.5: Heatmap showing the centroid distances and pairwise similarities between audio and text CLAP embeddings of all datasets. The color intensity represents the centroid distance, while the numerical annotations show the percentage of pairs exceeding 50% similarity threshold between datasets. The lower (left) triangle represents audio similarities, and the upper (right) triangle represents text similarities.

However, an attempt to use this CPU-based single-threaded $O(n^2)$ approach on about 20% of the datasets in this survey took approximately a month, making it not useful for a wider application.

To examine dataset similarities, we generated a heatmap visualization that combines centroid and pairwise similarities for audio and text embeddings (Fig. 2.5). The computation is performed using GPU-accelerated libraries (cuPy and PyTorch) for efficiency. Centroid similarities were calculated using mean embedding vectors for each dataset, with pairwise cosine similarities determining their relationships. In addition, pairwise similarities were computed for all embedding pairs and thresholds of 50% and 99% similarity were used to quantify relationships. The similarity threshold of 50% is used in the visualization to highlight highly similar datasets. The diagonal is masked, since self-similarities are not meaningful for the analysis. The

99% similarity threshold is used to define non-overlapping subsets of datasets, which are provided on our GitHub repository.

In the heatmap, colors represent centroid similarity scores ranging from 0 to 1, where higher scores indicate closer relationships. For pairwise similarities, the percentage of pairs exceeding the 50% threshold provides additional information on the overlap of the dataset. In some cases, higher relative percentages may correspond to a subset of one dataset being highly similar to another, while overall similarity remains low. In contrast, a lower percentage with a high centroid score might indicate broader similarities between datasets.

Analysis of the audio centroid distances in the heatmap’s bottom-left triangle (Fig. 2.5), reveals strong acoustic overlap in pairs like AudioCaps-MULTIS (similarity: 0.993), FAVDBench-VGGSound (0.964), CompAR-AudioSet (0.942), AudioCaps-CompAR (0.942), and MULTIS-CompAR (0.939). For some of these pairs, the overlap can be explained by known relationships - for instance, MULTIS and CompAR are both built upon AudioSet, which also includes AudioCaps samples. In contrast, FAVDBench’s overlap is less clear, but may originate from the inclusion of YouTube sources, which may contain VGGSound samples [412]. Scores close to 0 indicate that datasets exhibit minimal acoustic overlap, suggesting that their combination may be beneficial in training audio-language models. In particular, the lowest scores are found between DAQA-LAION-Audio-630k (0.176), MACS-CAPTDURE (0.167), CAPTDURE-AudioCaption (0.165), SoundJay-MACS (0.148), and AdobeAuditionSFX-MACS (0.148).

The upper-right triangle of the heatmap (Fig. 2.5) shows high text similarity, assessed through the centroids, in pairs like LASS-Clotho (similarity: 0.980), MACS-AudioCaps (0.935), FAVDBench-AudioCaps (0.887), VGGSound-FSD50k (0.859), and LASS-MULTIS (0.847), reflecting shared vocabulary and captions. Meanwhile, the lowest similarities occur between AdobeAuditionSFX-MULTIS (similarity: 0.148), AdobeAuditionSFX-MACS (0.146), BBCSoundEffects-WavText5k (0.144), AudioCaps-WavText5k (0.137), and WavText5k-MACS (0.056). In general, AdobeAuditionSFX (average similarity: 0.280) and WavText5k (0.375) have the lowest text similarity scores with all other datasets. FSD50k (0.631) and SpatialSoundQA (0.623). For audio, this is MACS (0.273) and AdobeAuditionSFX (0.376), and Freesound (0.629), VGGSound (0.606).

Analysis of pairwise audio embedding similarities, as indicated by the numerical value in each cell, shows that AudioDiffCaps-SoundingEarth (266M pairs, 13.89%), AudioSet-AudioCaption (5.9B pairs, 13.24%), and AnimalSpeak-SoundingEarth (4.8B pairs, 10.93%) have the highest number of pairs exceeding the 50% similarity threshold. For the 99% similarity threshold, SonnissGameEffects-VGGSound (22,880 pairs, 0.21%), MULTIS-AudioCaps (5,044 pairs, 0.16%), and Freesound-SonnissGameEffects (29,016 pairs, 0.05%) show the highest number of relative similar pairs. The lowest number of relative pairs exceeding the 50% similarity threshold is found between

MACS-PicoAudio (11,918 pairs out of 258M total comparisons, 0.0046%), PicoAudio-AudioCaption (17,280 pairs out of 286M, 0.0060%), and CAPTDURE-AudioCaption (11,390 pairs out of 186M, 0.0061%).

The analysis reveals patterns across the datasets. AudioDiffCaps shows high acoustic similarity with several datasets, particularly SoundingEarth (13.89% of pairs >50% similar), AnimalSpeak (10.88%) and Clotho (7.53%). Similarly, AudioCaption exhibits strong acoustic relationships with multiple datasets, having more than 5% similarity to MULTIS (7.35%), CompAR (6.40%) and AudioSet (13.24%). In contrast, specialized datasets like MACS and AdobeAuditionSFX show minimal acoustic overlap - MACS has less than 0.5% similarity with all other datasets except SoundingEarth (6.29%), while AdobeAuditionSFX's highest similarity is only 1.94% with LAION-Audio-630k. This suggests that these datasets contain unique acoustic content that is not well-represented in other datasets.

The analysis of pairwise text embedding similarities shows that AudioDiffCaps-MACS (312.8M pairs, 40.07%), Audiocaps-MACS (335.1M pairs, 39.31%), and FAVDBench-MACS (45.8M pairs, 22.80%) have the highest number of pairs exceeding the 50% similarity threshold. For the 99% similarity threshold, SoundJay-MACS (420 pairs, 0.0018%), SoundJay-VGGSound (1,502 pairs, 0.00068%), and Audiocaps-VGGSound (37,199 pairs, 0.00046%) show the highest number of relative similar pairs. The lowest number of relative pairs exceeding the 50% similarity threshold is found between MACS-WavText5k (44,783 pairs of 94.2M total comparisons, 0.048%), AnimalSpeak-AudioEgoVLP (4.2M pairs of 7.7B, 0.054%) and AudioEgoVLP-AudioSet (9.9M pairs of 17.5B, 0.057%).

In general, datasets such as WavText5k consistently show a low text similarity to other datasets, with its highest similarity only 1.16% with AudioCaption. Similarly, AdobeAuditionSFX shows limited textual overlap, with its maximum similarity being 2.39% with SonnissGameEffects. However, AudioDiffCaps again shows high textual similarity with several datasets, particularly AudioSet (21.2%), DAQA (20.3%) and FAVDBench (17.3%). Particularly noticeable are the 39.3% and the 40.1% similarities in text between LASS and AudioCaps and AudioDiffCaps respectively.

For all datasets, our GitHub repository provides subsets of datasets that do not overlap with other data, according to the 99% threshold. The GitHub repository also contains instructions and easy-to-use bash scripts for downloading each dataset in our analysis.

2.5 Discussion

2.5.1 Data quality

The quality of training data influences the performance of machine learning models. Although high-quality training datasets are limited in size [603], relying on large-scale

noisy datasets remains the only scalable option for many tasks [104]. However, as noted in the AudioCaps paper, such datasets may induce distributional changes [199], particularly in the generation of audio samples with user-generated free-form text input instead of training set prompts [53]. Mei et al. [296] showed that filtering noisy data and rewriting raw descriptions into high-quality captions can achieve state-of-the-art performance with smaller datasets. Similarly, the addition of AudioCaps with ChatGPT resulted in better audio-text retrieval performance [339].

The combination of datasets poses additional challenges. For example, Mei et al. [296] observed a performance drop in the AudioCaps dataset when the LAION-Audio-630K dataset, which is larger but presumably of lower quality, was included during training. However, integrating the WavCaps dataset reversed this decline and led to improved performance in the Clotho and AudioCaps datasets despite the smaller dataset size. These findings emphasize the importance of data quality over quantity.

LLMs offer a promising solution to improve the quality of the dataset. Rewriting existing datasets to produce more elaborate and diverse captions has been shown to improve training outcomes and mitigate the adverse effects of noisy data [523, 367, 530, 573, 214]. However, a key challenge is the inconsistency in the style of the caption in audio-language datasets [603]. Proposed solutions include the introduction of tokens specific to the dataset [228] or standardizing the style and length of the caption using LLM-assisted rewriting (e.g., [203, 129, 435]).

2.5.2 Bias in Audio Datasets

Audio-language models face multiple biases that can affect their fairness and performance. Key sources of bias include cross-contamination of train/test splits, the bias related to the visual-grounding of datasets, diversity bias and the bias related to dataset augmentation (model collapse).

Cross-contamination

Cross-contamination occurs when training and testing datasets overlap, leading to overly optimistic performance metrics, as the model is essentially being tested on information it has already seen and masking a model’s generalizability. This issue is prevalent in the context of Audio-Language Learning, especially when combining datasets. For example, 17.6% of AudioCaps overlaps with WavCaps, while Clotho shares an even larger overlap of 89% [228]. Such overlaps distort evaluations and undermine real-world applicability. The dataset overlap analyses provided in this survey can help avoid such cross-contamination.

Visually-grounded datasets

Many prominent audio-language datasets, such as AudioCaps, were annotated with access to the corresponding video content [199] (see Section 2.2.3 and Appendix 2.7.1). This introduces a risk of visual bias, where captions include information not discernible from audio alone (e.g., “A female instructor giving a speech in front of a live audience”). Xu et al. [554] found that the noise introduced by visual context limits the improvements in audio-text alignment studies [588, 151, 521]. Efforts to address this include filtering visually-influenced captions using LLM (GPT4o), though such a method resulted in an audio-language dataset that contained only 0.65% of the original data (70,000 of 10.8 million) [222].

Several datasets such as Auto-ACD [435] and Sound-VECaps [573] have been created by augmenting audio datasets with visual information to enrich their content. However, research by [333] showed that visual language models tasked with processing both video and audio tend to generate audio descriptions predominantly based on visual information rather than actual audio content. Their analysis of 1000 cases showed that Video-LLaMA produced audio descriptions containing visual-only information in 32.3% of the cases. To mitigate this, several works proposed to filter out visual-only information from the caption [573, 435, 154].

Diversity bias

Audio-language models trained on unbalanced, non-diverse datasets may perform poorly on underrepresented groups and cultures [388]. Similarly, audio retrieval systems may exhibit lower accuracy when searching for content related to cultures or languages poorly represented in the dataset. Most existing datasets focus on English, with only a few exceptions, such as AudioCaption [522, 549] and FAVDBench [412], both of which have original Mandarin captions translated into English, and CAPTDURE, translated from Japanese to English [340]. Efforts to expand linguistic diversity include translating datasets into multiple languages. Cousin et al. [78] used automatic translation to translate AudioCaps and Clotho to German, French, and Spanish. Manakul et al. [284] translated AudioCaps from English to Thai. For audio question answering, Behera et al. [34] translated into ClothoAQA in 8 languages. Kang et al. [193] built a Korean audio captioning system, first training the system in English and then translating the captions from English to Korean. This approach has also been scaled to 181 languages by Noriy et al. [334]. The authors found that this multilingual data increase enriches the variability of languages and accents within their training set, enhancing the model’s generalization capabilities. However, relying on automatic translation can introduce its own bias. Annotator proficiency in dataset languages also raises concerns, particularly when using platforms such as Mechanical Turk, known for demographic biases [167, 92].

Model collapse

Augmentation techniques have become increasingly important in audio-language research to address data scarcity and improve model performance. Research has shown that the use of ChatGPT and other LLMs for text augmentation leads to higher-quality datasets for training audio language models by filtering captions [219, 383, 530] or generating new captions [383, 367, 203, 435, 129, 145]. However, attempts to augment audio data using text-to-audio generation models have shown limited success due to the insufficient diversity and realism of the generated samples [265, 127, 132]. For example, [132] attempted to create compositional audio using cutting-edge text-to-audio generation models [265, 256, 127] to enhance training for audio-language models, but found limitations in generating diverse and cohesive audio for acoustic scenarios and concluded that a gap in the capability of current text-to-audio models is still present. This limited success is a consequence of “model collapse” [418], a phenomenon common in many areas of machine learning where models relying on synthetic data perform well on such data, but fail to generalize to real-world scenarios.

2.5.3 Dataset Accessibility

Dataset accessibility and reproducibility remain challenges in audio-language research. Many datasets, such as those derived from AudioCaps, require researchers to scrape audio content manually using command-line tools such as *yt-dlp*, or *yt-dl* for YouTube-based content. Some others follow potentially cumbersome naming conventions [212], such as adding a ‘Y’ prefix to filenames - a practice seen in [212] and subsequent works - to handle YouTube IDs that might contain symbols problematic for programmatic parsing. For other datasets, such as AnimalSpeak, SoundingEarth, and Zapsplat, obtaining the data is a time-consuming and complex process. Instructions for accessing datasets like DAQA and AudioEgoVLP are often unclear, while others, such as SoundScaper, present barriers that hinder use.

Reusing existing datasets as building blocks for new ones further complicates access and reproducibility, as acquiring a single dataset may necessitate downloading multiple prerequisites. To improve this situation, researchers should prioritize hosting datasets on accessible platforms such as Zenodo or HuggingFace, to provide straightforward access and facilitate reproducibility. Our GitHub repository provides scripts for easy acquisition of datasets.

2.6 Conclusion

This survey provides an overview of audio-language datasets and their role in developing audio-language models. The field has seen advances in computing power and

the creation of large-scale datasets like AudioCaps and Clotho. This enabled the training of more sophisticated models capable of understanding and describing diverse sounds. This survey examined various aspects, including dataset design, availability, training techniques, and challenges in the field.

A key finding is the increasing use of language models for dataset augmentation and caption generation. Currently, 35 out of the 69 analyzed datasets have used LLMs to generate captions for their datasets. While LLMs can help tailor datasets for specific model purposes, our analysis reveals potential redundancy as many LLM-generated variations of the same datasets may already exist (e.g. several similar versions of Clotho). Section 2.5 highlighted how this promising approach needs consideration. Both bias and hallucination issues could occur when using LLMs to generate captions.

This survey emphasizes the role of data quality and quantity, particularly for contrastive learning approaches such as CLAP. While larger datasets enable better model performance, the literature shows that simply increasing dataset size is not sufficient. In this regard, the quality and diversity of the data are equally important. Studies show that incorporating lower-quality datasets can actually decrease model performance, while smaller but higher-quality datasets can lead to improvements.

The analysis in Table 1 reveals that YouTube-based datasets dominate the field (23 datasets), followed by Freesound-based datasets (21 datasets). Sound-effects libraries and field recording datasets also form a large portion of datasets (18 and 15 datasets, respectively). However, sound effect libraries sometimes require payment or special access for their use. Other datasets, particularly AudioSet-derivatives, are difficult to access since AudioSet’s YouTube-based constraints prevent direct resharing of audio data, requiring a manual download. 49.2% (34 out of 69) of the analyzed datasets use LLM-generated captions, indicating a clear trend toward synthetic data generation. This trend is particularly strong in recent years - in 2024, 80% of new datasets (20 out of 25 datasets released that year) used LLM-generated captions, up from 69.2% (9 out of 13) in 2023.

Analysis of sound content distributions in audio datasets reveals imbalances in sound category representation, with indoor environments and human-generated sounds dominating. Our analysis shows that “Inside, small room” (21.6%), “Music” (14.1%), and “Male speech” (13.6%) account for nearly 50% of all samples, while specific sound categories remain severely underrepresented. Dataset-specific patterns emerged, such as the high concentration of male speech of DAQA’s (37.8%) and BBCSoundEffects’ bias toward indoor environments (34.2%). When examining the uniformity of the datasets, VGGSound showed the most balanced distribution (0.720 normalized entropy), followed by FindSounds (0.615) and BATON (0.610). PCA analysis of CLAP embeddings revealed that AnimalSpeak, AudioCaption, MACS, and AudioSet had the most distinct centroids. At the same time, Freesound demonstrated the highest

variability in both audio (0.925) and text (0.930) embeddings, reflecting its diverse user-uploaded content.

Duplicate analysis as shown in Section 2.4.2 revealed overlap between various datasets. Practitioners should account for data leakage when training and evaluating models, particularly for zero-shot evaluation. The heatmap analysis showed great similarities between certain pairs of datasets, such as VGGSound-FAVDBench and Freesound-FSD50k. Large audio-language models may have seen similar data during pretraining, which can lead to over-optimistic performance metrics. It is not recommended performing strong zero-shot evaluation on datasets that share similar origins (e.g., training on Clotho and evaluating on FSD50k, or training on Auto-ACD and evaluating on AudioCaps). This evaluation is better described as a weak zero-shot evaluation.

Some overlaps are more obvious than others. The following are some examples of overlaps: the VGGSound and AudioSet overlap. ProSoundEffects and BBC Sound Effects also overlap, as Pro Sound Effects operates the commercial licensing of BBC Sound Effects. Auto-ACD is based on VGGSound and AudioSet, and WavCaps is based on FreeSound, BBC, SoundBible, and AudioSet. AnimalSpeak also contains AudioCaps as a subset. The analysis in this survey also deduced that the overlap between VGGSound and FAVDBench Audio is quite evident. For each dataset, specific datasets without overlaps are provided in our GitHub repository.

A relevant trend is the increasing amount of different tasks and subtasks that the field of Audio-Language Learning is exploring. This survey shows that 8 different tasks are currently being explored, with a further division of 11 subtasks. As Audio-Language models become more usable for real-world applications, it is likely that more tasks will be explored. In particular, various tasks that originally involved sound classification are now being extended to broader language involvement, such as text-queried sound event localization, language-queried audio source separation, and audio difference captioning.

The analysis of dataset statistics reveals varying approaches to caption length and detail. While some datasets focus on single-sentence captions (e.g., AudioCaps with an average of 49.22 characters per caption or MACS with an average of 54.26), others contain more elaborate captions (e.g., Clotho-Detail with an average of 324.83 characters per caption). This trend of more elaborate captions is likely to continue, particularly for instruction-tuning large audio-language models, as it allows for more detail and context to be captured in the training data.

2.6.1 Future Research Directions and Recommendations

Looking ahead, this survey identifies several areas for future dataset development. Most existing datasets are heavily skewed toward English-language content. While some datasets like AudioCaption and FAVDBench have original Mandarin captions, and CAPTDURE has Japanese captions, the vast majority of audio-language datasets

are English-only. Current approaches largely rely on automatic translation of English datasets to other languages, as seen in efforts to translate AudioCaps and Clotho to languages such as German, French, Spanish, and Thai. Although translation can help address immediate needs, it may not capture the full linguistic and cultural nuances of different languages. The development of original high-quality datasets in diverse languages would be valuable to advance truly multilingual audio-language models.

A gap also exists in specialized domains, with some noticeable exceptions: Animal-Speak in the bioacoustics domain, AudioCaption Hospital in the medical domain, SoundScaper dataset for soundscapes and MIMII-Change for malfunctions in industrial machines. This survey shows that most datasets focus on human activities, speech, and urban sounds, leaving room for expansion to other underexplored acoustic domains, such as environmental audio, niche industrial applications, and natural soundscapes.

YouTube-based datasets, while numerous, face challenges related to link rot and content availability, affecting research reproducibility. These findings suggest that future efforts should focus on quality, accessibility, and robustness. Emphasis should be placed on reducing redundancy and addressing biases and representation gaps to ensure that the datasets remain reliable and inclusive.

The collapse of models in audio-language datasets also requires attention. Although LLMs have shown promise in text augmentation and caption filtering, attempts to use audio generation models for data augmentation have had mixed results. Current text-to-audio generation models still show limitations in generating diverse and cohesive audio for acoustic scenarios. Future research needs to focus on developing more robust audio generation techniques that can produce high-quality synthetic data while avoiding the pitfalls of model collapse, where models rely on synthetic data and perform poorly on real-world examples.

To improve pretraining for audio-language models, researchers should explore alternative sources beyond Freesound-based datasets like Clotho or YouTube-based datasets such as AudioCaps. This can diversify the dataset and result in a more robust model. Some sound effect datasets and field recording datasets show the least overlap in this survey's analysis, offering promising alternatives for dataset diversification. WavText5K, BBC Sound Effects, Adobe Audition SFX and MACS are particularly recommended.

Dataset accessibility remains a pressing concern. Researchers should prioritize making their datasets directly available through established platforms such as Zenodo or HuggingFace to ensure easier access and better reproducibility. Researchers should also be transparent about their dataset acquisition process to avoid confusion and ensure that their datasets are accessible to others. Our GitHub repository provides scripts for easy acquisition of datasets.

This survey identifies a need for more sophisticated duplicate detection and dataset curation methods. Our analysis revealed complex patterns of overlap between datasets, with some similarities being more subtle than others. Although current methods, such as audio fingerprinting and embedding-based similarity detection, provide insight, they are often computationally intensive and limited in their ability to detect subtle overlaps. Future research should focus on developing more efficient and accurate methods for detecting duplicates and overlaps in audio-language datasets. This would enable better dataset curation and more reliable evaluation of zero-shot learning capabilities, while also addressing current challenges in assessing true model generalization versus performance inherited from overlapping training data.

This survey provides detailed information on the current state of datasets, their applications, and challenges. The analysis reveals trends in the development of datasets, including the growing use of LLMs for caption generation, differences in the ease of dataset acquisition, and dataset overlap issues. These findings can guide researchers in making informed decisions about dataset selection and future development efforts in audio-language learning.

2.7 Appendix

2.7.1 List of audio-visual datasets

- Video Captioning and Dense Description: ActivityNet-Captions [220], MSVD [56], MSR-VTT [545], VATEX [488], TaCos [385].
- Instructional and Procedural Content Understanding: YouCook [595], HowTo100M [304], CrossTask [605], How2 [398].
- Visual Instruction Tuning: VideoChat-instruct-11K [246] and Valley-Instruct-65k [275]
- Event Localization and Action Recognition: AVE [457], TVC [235], LLP [456]. DiDeMo [169], Charades-STA [122].
- Large-Scale Multimodal Integration: HD-VILA-100M [555], Webvid10M [26], VIDAL-10M [601], VALOR-1M [62], VAST-27M [63], ViT-Lens-2 [236].
- Specialized Domains: LSMCDC/MPII-MD [387, 386], M-VAD [459], Sound Of Story [20], SoundNet [19], ASIW [54]

2.7.2 Example caption per dataset

Table 2.3: Example caption per dataset: the first occurrence of the word dog in each dataset.

Dataset	Example Caption
ACalt4	A dog whimpering and yipping, possibly in discomfort or seeking attention
AFAudioSet	The audio contains a variety of sounds including a dog growling, the whirring sound of mechanisms, and a man speaking.
AdobeAuditionSFX	Animal Mammal Carnivore Dog Bulldog Jowl Flap 01
AnimalSpeak	The sound of a Magnolia Warbler singing from dense dogwood foliage.
AudioAlpaca	A female’s speech followed by a dog barking and whimpering.
AudioCaps	A dog whimpers quietly
AudioCaption	A woman nurse and a female doctor talk about dogs.
AudioCondition	Mechanisms from 0.0 to 10.0 and Generic impact sounds from 0.128 to 0.259 and Clicking from 0.467 to 0.543 and Whimper (dog) from 0.688 to 1.152 and Dog from 1.124 to 4.817 and Clicking from 4.775 to 4.824 and Surface contact from 4.886 to 5.066 and Dog from 5.149 to 6.912 and Clicking from 5.17 to 5.336 and Generic impact sounds from 5.979 to 6.186 and Generic impact sounds from 6.435 to 6.629 and Generic impact sounds from 6.857 to 7.002 and Generic impact sounds from 7.327 to 7.5 and Human voice from 7.68 to 8.178 and Generic impact sounds from 7.791 to 7.998 and Generic impact sounds from 8.288 to 8.489 and Generic impact sounds from 8.745 to 8.966 and Generic impact sounds from 9.236 to 9.409 and Human voice from 9.277 to 9.838 and Generic impact sounds from 9.713 to 9.927
AudioDiffCaps	remove sound of dog barking
AudioEgoVLP	The hotdog display counter is closed, producing a sliding and locking sound.
AudioHallucination	Is there a sound of dog in the audio? No
AudioTime	A dog barked for 1.385 seconds, followed by a steam whistle lasting 5.25 seconds.
AudioSet	Speech, Whimper (dog)

Continued on next page

Dataset	Example Caption
AudioSetStrong	Whimper (dog)
Audiostock	Medium-sized dog barks (one!)
AutoACD	A dog howls loudly followed by a person screaming, creating a chaotic atmosphere in the vicinity.
BATON	A dog barks, bees buzz, and rail transport moves
BBCSoundEffects	Ext. aerobatics as in dog fight - 1967 (104B)
BigSoundBank	barking dogs.
CAPTDURE	The sound of a dog on a leash “roaming” on an inorganic floor.
CHiMEHome	-
Clotho	A dog jumped into the water and swim and then came out panting, it shook its body
ClothoAQA	Is it a dog making the noise? no
ClothoChatGPTMixup	a couple converses as a dog barks intermittently while coins are dropped into a jar and a hand searches through a packet
Clotho-Detail	In this audio, we hear a series of creaking and squeaking sounds. These noises could be attributed to someone leaning back and forth on an office chair, or perhaps to a slowly creaking door opening. Additionally, there are intermittent popping sounds that eventually taper off. Background noises, such as a dog barking, add additional context and atmosphere to the scene.
ClothoEntailment	Entailment: A dog engages in a water activity and exhibits behaviors such as panting and shaking.
ClothoMoment	A dog is barking at the heavy downpour of rain. [13s, 38s]
ClothoV2GPT	A door creaks while a dog barks in the distance.
CompAR	Based on the audio elements provided, infer the ambiance or mood conveyed by this outdoor setting. Be sure to consider the impact of the specific animal and human sounds. The combination of bird chirping, insect sounds, and a human voice suggests a serene, possibly pastoral setting. The barking dog adds to the perception of a lively yet relaxed ambiance.

Continued on next page

Dataset	Example Caption
DAQA	Was the sound of the car passing by louder than the sound of the dog making noise? yes
ESC50	dog
EpidemicSoundEffects	the sounds of ambience, subdivision, night, insects, distant, dog, loop, bugs, and distant dog.
EzAudioCaps	A woman speaks while a dog barks in a domestic setting, creating a lively atmosphere.
FAVDBench	A man commands a dog. Occasionally, you may hear a dog barking in the distance
FSD50k	20061012. a distant ambulance siren along with some dog barks. The noise of the siren ends abruptly, presumably upon arrival to destination. ME62+AKGCK94 into iRiver H120 decoded to midside stereo / una abulancia lejana con algun ladrado de perro
FindSounds	dog
FreeToUseSounds	animals dog bark med 004
Freesound	Creek dogs.
LAION630k	Village quiet insects birds dogs and light activity away 140301 04
LASS	a lot of insects are chirping and a dog is barking.
MACS	adults speaking while a dog barking
MULTIS	What is happening to the dog? The dog sounds like it is upset, possibly whimpering because of hunger, pain, or fear.
NonSpeech7k	-
Paramount	-
PicoAudio	dog barking at 3.038-5.038, 5.585-7.585
ProSoundEffects	Welsh Hillside With Occasional Creaking Trees, Birds & Distant Sheepdog
RDATS	A loud wheel rolling precedes a dog barking, amidst the background of crickets chirping.
SonnissGameEffects	Village quiet insects birds dogs and light activity away 140301 04

Continued on next page

Dataset	Example Caption
SonycUST	Manhattan, medium engine, large engine, impact machinery, non-machinery impact, people talking, large crowd, dog barking/whining
SoundBible	A dog is barking.
SoundDescs	Common (Northern) Flicker (Colaptes Auratus) - medium close-up calls. Calls from Killdeer flying over lake, dogs barking in part and insects in background.
SoundJay	-
SoundVECaps	A festive atmosphere fills the indoor setting as a person holds a small dog, and then music plays in the background while a man speaks.
SoundingEarth	Dandeli jungle camp, Pradhani, Karnataka 581363, India - Dogs howling. India, Karnataka, Dandeli, Pradhani, jungle, camp, night, dogs, insects, Binaural, Fostex FR-2LE, MS-TFB-2 microphones
SpatialSoundQA	Identify the sound events in the audio clip. animal; dog; domestic animals, pets; grunt; pig; speech
Syncaps	Helicopter flying before dog barking
TextToAudioGrounding	a dog is growling and barking
VGGSound	dog growling
WavCaps	People are speaking and dogs are barking, with ticking sounds in the background.
WavText5K	barking dogs.
WeSoundEffects	-
Zapsplat	Cartoon dog grunt 1

The Data Advantage: Balanced Curriculum Learning for Enhancing Audio Question Answering



ADAPTED FROM

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. *Data-Balanced Curriculum Learning for Audio Question Answering*. July 2025. doi: 10.48550/arXiv.2507.06815

Abstract

Data quality influences the performance of Audio Question Answering (AQA) models. This work introduces a data-balanced curriculum learning pipeline that addresses challenges in data quality through three mechanisms: eliminating training-evaluation data overlap, filtering overrepresented data categories, and stabilizing reinforcement learning through curriculum design. The pipeline uses Phi-4-mini to categorize audio content and applies statistical thresholds to identify imbalanced classes, creating a reproducible curation strategy accessible to researchers with limited resources. Experiments demonstrate that filtering and curriculum learning provide a performance boost in the DCASE 2025 Task 5 development dataset. The proposed model achieves a 3.3% improvement over standard supervised fine-tuning, indicating that data curation is a key factor in AQA model performance under the studied conditions.¹

Preface

This chapter answers Research Question 2 (Curriculum Learning and Data Balancing for AQA) by evaluating how clean training-evaluation separation, statistical category balancing, and curriculum-stabilized reinforcement learning affect audio question answering performance. It contributes to the data-quality theme by implementing low-overlap splits and category balancing as a reproducible pipeline and by quantifying the effect of reducing contamination and imbalance. The resulting procedure motivates how later chapters treat data preparation when comparing models and methods.

¹Code and data are available at <https://github.com/GLJS/balanced-curriculum-aqa>

3.1 Introduction

Audio Question Answering (AQA) is the task of answering questions about audio content. It demands capabilities at the intersection of acoustic understanding and natural language processing. Unlike simple sound classification, AQA requires audio-language models (ALMs) to comprehend complex acoustic scenes, identify temporal relationships between sounds, and generate precise answers to various questions [241, 34, 110]. The field of AQA has evolved from early models with limited set of labels [1, 432, 254] to advanced systems capable of generating natural language responses [389, 541]. Audio Question Answering splits into two categories: multiple-choice and open-ended. Most benchmarks use multiple-choice formats to enable direct verification against ground truth [393, 280]. Under these benchmarks, AQA models face three challenges:

Challenge 1: Data impurity. Audio-language models (ALMs) suffer from data contamination from training-evaluation (see Chapter 2). Models often train on datasets derived from common sources such as AudioSet [126] or Freesound [117], and evaluation benchmarks often contain data from these same sources. This contamination is widespread; examples include MMAU [393] containing 5,630 audio examples (out of 10,000) from AudioSet derivatives [199, 296, 170] and AIR-Bench [565] including 1.9k audio samples from common sources [102, 114, 296]. This overlap fundamentally undermines the performance gains reported.

Challenge 2: Data imbalance. Audio datasets exhibit category imbalances, with certain classes like Music and Speech overrepresented [126, 510], which limits model generalization to more diverse categories [161].

Challenge 3: Training instability. Reinforcement learning could introduce training instabilities, which include reinforcement of unhelpful reasoning patterns [79] and high gradient variance [332].

This chapter makes the following contributions:

1. Addressing data impurity through a clean training-evaluation separation that eliminates all overlap between training and evaluation data. Before training, the training and evaluation sets were verified to be disjoint.
2. Mitigating data disproportions by statistical data balancing using Phi-4-mini [2] to categorize audio questions. Statistical thresholds are used to filter over-represented classes to create a reproducible curation strategy that improves performance.
3. Resolving training instability through curriculum-stabilized reinforcement learning where the approach stabilizes Group Relative Policy Optimization (GRPO) by introducing easy samples only. This establishes reliable reward signals [35, 106].

Experiments on six training datasets with evaluation on the benchmark dataset, the DCASE 2025 Task 5 Challenge development dataset, validate this approach. The results show that intelligent data filtering provides the best increase in performance without requiring massive proprietary datasets or exotic computational resources, making the approach accessible to researchers with limited resources.

3.2 Related work

Reinforcement learning has become an important technique in post-training and fine-tuning large language models that leads to improvements [559, 142]. A notable recent advancement in this area is Group Relative Policy Optimization (GRPO) [409], a reinforcement learning algorithm designed to improve the reasoning skills of LLMs. GRPO distinguishes itself by being more memory-efficient than traditional methods, as it does not require a separate critic model [409].

Recently, reinforcement learning has also been applied to audio language models with promising results [240, 503]. GRPO [409] has been shown to be effective for audio language models. Fine-tuning Qwen2-Audio-7B-Instruct with reinforcement learning shows improvements over supervised approaches [240], especially when fine-tuning with reinforcement learning on a foundational model [389]. The SARI framework [503] demonstrates that curriculum-guided GRPO with structured reasoning improves performance on complex audio tasks.

The method proposed in this chapter differs from previous approaches by uniquely combining three elements: (1) rigorous data separation between training and evaluation, (2) statistical data balancing to address imbalances of the datasets, and (3) curriculum learning to stabilize reinforcement learning.

3.3 Method

3.3.1 System Overview

In audio question answering (AQA), an audio-language model processes audio input along with a question and optionally multiple choices (four in the evaluation set), and returns the correct choice. This work uses Qwen2-Audio [72] as the foundation model, a unified architecture capable of processing diverse audio inputs. Qwen2-Audio is a large-scale audio-language model that combines a Whisper Large V3 [374] audio encoder with a Qwen-7B [22] language model.

3.3.2 Datasets

The following six datasets are used for in this research for training the different models. Together, they do not overlap with any of the examples from the evaluation set. The training dataset consists of 373,700 examples in total over 1102.09 hours of audio.

- **DCASE 2025 Task 5 Development Set**²: The training data from a public audio question answering challenge of DCASE 2025. This dataset consists of 8221 question-answer pairs over 33.58 hours of audio
- **AVQA** [563]: Contains 56,369 question-answer pairs from VGGSound (originally 57,335 before YouTube deletions). The questions span eight semantic types. The dataset includes audio and visual information. AVQA requires models to discover causal correlations between modalities.
- **ClothoAQA** [254]: Contains 35,838 questions from 1,991 environmental audio samples (15-30 seconds). Each audio has 6 questions with 3 annotator responses per question. Four questions require binary answers; two require single words. Different annotators create questions and answers.
- **CompA-Order** [132]: Contains 900 examples of compositional reasoning. The dataset tests the temporal understanding between sounds.
- **TACOS** [368]: Contains 12k audio recordings with 61,137 temporally-aligned captions. The dataset is designed to train temporal reasoning.
- **AudSem** (see Chapter 4): Synthesized from YouTube closed captions. Contains 210,288 multiple-choice questions in various acoustic scenes and events.

Evaluation Dataset. The evaluation dataset used in this chapter is the *development* set from the DCASE 2025 challenge, which is different from the *training* set. The evaluation metric is standard 4-way multiple-choice accuracy, where chance level performance is 25%. The development set consists of three parts: (1) Bioacoustics QA: 0.2K development examples from 31 marine mammal species [518]. (2) Temporal Soundscapes QA: 0.6K development examples covering 26 sound classes for temporal reasoning [462, 287, 4]. (3) Complex QA: 1.6K development examples from AudioSet and Mira datasets [126, 188] that require multi-faceted reasoning in temporal, acoustic and contextual dimensions to answer. Before training, overlap between training and evaluation sets was verified to be absent. This chapter uses only the DCASE 2025 benchmark to guarantee a contamination-free evaluation setting. Other AQA benchmarks such as MMAU [393] contain AudioSet examples that overlap with AVQA (VGGSound) in the training data, which would undermine the validity of the evaluation.

²<https://dcase.community/challenge2025/task-audio-question-answering>

3.3.3 Experimental Setup

The proposed approach combines three components: (1) Curriculum-based filtering, (2) Statistical balancing, and (3) Hybrid training. These components work together to address the challenges of impurity, imbalance, and training instability of the data.

Curriculum-Based Data Filtering

Traditional data curation approaches often discard outliers as noise. In contrast, the curriculum approach presented here uses these challenging examples through the gradual introduction of complexity. Curriculum learning, training in easy examples before difficult ones [35, 106], has proven to be effective in machine learning domains [422].

The curriculum implementation uses Phi-4-mini-instruct [2] to score the difficulty of the question (see Box 3.3.3), to demonstrate that curriculum design requires neither proprietary models nor massive computation:

Prompt to Phi-4-mini-instruct for difficulty assessment

Task: Rate the difficulty of this audio-based question on a scale from 0.0 (very easy) to 1.0 (very difficult).

Consider the following factors:

- Complexity of the question
- Required knowledge/understanding
- Ambiguity or clarity of the question
- Number of concepts involved

Question: {question}

Choices: {choices}

Correct Answer: {answer}

Following difficulty assessment, samples undergo filtering. Easy samples (difficulty < 0.3) establish reliable GRPO reward signals, which prevent instability caused by ambiguous examples in reinforcement learning. Later training stages may incorporate all samples or focus on difficult examples (difficulty > 0.7). Experimental results demonstrate that training only on easy samples achieves optimal performance (see Section 3.4). The reason for this finding is that GRPO relies on receiving usable reward signals to update the model. Restricting training to easier items increases the probability of correct responses early in training, which stabilizes gradient updates and improves downstream accuracy. When questions are too difficult, the model rarely produces correct answers, resulting in sparse or uninformative reward signals that hinder learning.

Statistical Category Balancing

Dataset imbalance presents a fundamental challenge to audio-language models, limiting their generalization capabilities and real-world applicability. Analysis of AudioSet reveals category imbalances, with classes such as Music and Speech overrepresented [126]. The proposed statistical balancing approach addresses this problem and maintains computational efficiency.

The procedure first categorizes audio content into 24 categories using Phi-4-mini-instruct, partially derived from the AudioSet ontology and supplemented with 14 additional categories identified through manual evaluation of 100 examples (see Figure 3.1). Phi-4-mini-instruct was selected as the best-performing model on standard benchmarks at its size at the time of writing. The model is used only to assign coarse category labels and scalar difficulty scores (0-1), not to generate new captions or text, which reduces the risk of hallucination. Manual inspection of 100 randomly sampled examples confirmed that the category tagging is sufficiently consistent for filtering purposes. Statistical thresholds then identify and filter overrepresented categories based on their deviation from the mean distribution.

The statistical threshold procedure calculates the mean count μ in all categories and filters examples where the category count c_i exceeds $\theta \cdot \mu$. Empirical testing identified the optimal threshold at $\theta = 0.7$ (see Section 3.4) to filter the overrepresentation of human sounds and mixed environment sounds and preserve sufficient training examples. The thresholds 0.3, 0.5, and 0.7 were chosen to cover the range of filtering strengths distributed between 0 and 1, providing a systematic sweep from aggressive ($\theta = 0.3$) to mild ($\theta = 0.7$) filtering. This approach directs reinforcement learning toward diverse examples and generalizes to rare acoustic events without requiring manual annotation or specialized tools.

As shown in Figure 3.1, this approach prevents overfitting to dominant categories without requiring complex weighting schemes or specialized architectures. The methodology is generalized to other audio-language tasks affected by the imbalance of the dataset.

Hybrid Training Pipeline

The training methodology uses a two-stage hybrid approach that integrates supervised fine-tuning (SFT) with the GRPO reinforcement learning technique, which eliminates the separate value model required in traditional reinforcement learning from human feedback (RLHF) approaches and maintains training stability. The SFT stage warms the model, which is beneficial for training stability [351, 403] and adapts the model to the format and style of the task.

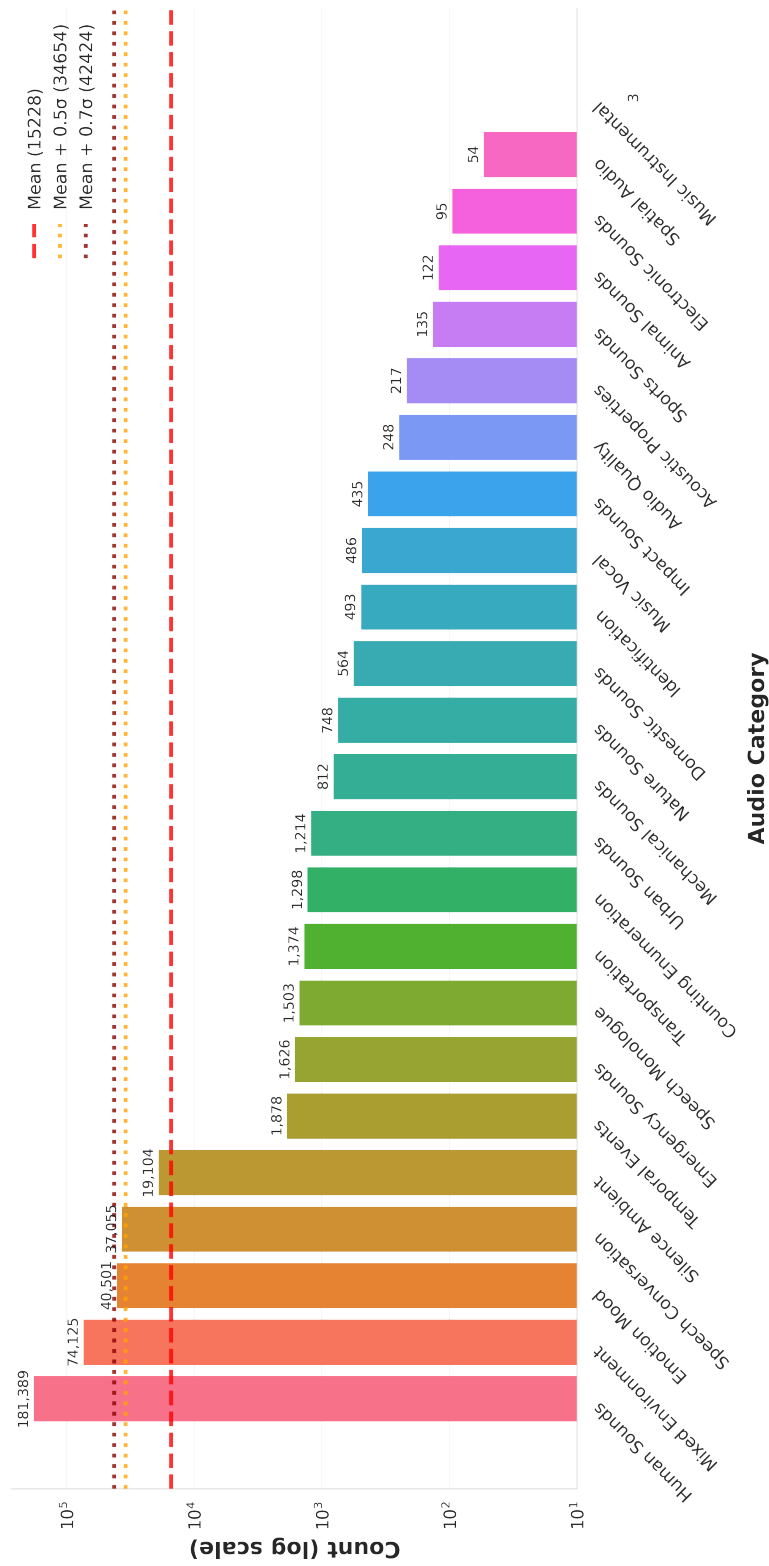


Figure 3.1: Distribution of audio categories in the training sets. Dotted lines represent the threshold during diversity filtering, with example cut-offs of 0.5σ and 0.7σ above the mean.

Stage 1: Supervised Fine-Tuning (SFT). In the initial stage, Qwen2-Audio-7B-Instruct [72] undergoes fine-tuning using cross-entropy loss with Low-Rank Adaptation (LoRA) [179]. This parameter-efficient technique modifies only query and value projection matrices with rank $r = 8$, scaling factor $\alpha = 16$, and dropout 0.05, reducing computational requirements. Optimization uses AdamW [268] with learning rate 2×10^{-5} , cosine annealing schedule, and gradient clipping at 0.5.

Stage 2: Group Relative Policy Optimization (GRPO). Following SFT, the second stage applies GRPO to refine performance through reinforcement learning with verifiable rewards (RLVR) [409]. The reward function decomposes into the accuracy and format components:

$$R(y, \hat{y}) = R_{\text{acc}}(y, \hat{y}) + R_{\text{format}}(\hat{y})$$

The accuracy component R_{acc} allocates 0.5 points for exact matches, with partial credit of 0.25 for correct identification of the answer letter (A, B, C, or D) and 0.25 for the matching of the content. The format component R_{format} contributes 0.5 points for the adherence to the required response formatting.

Training stability is maintained through conservative KL-regularization ($\beta = 0.01$) with a 50-step warmup period. Each training example generates 4 candidate responses. The implementation disables dropout during this phase and uses a reduced learning rate of 1×10^{-6} with AdamW optimization and cosine annealing.

Both supervised fine-tuning and reinforcement learning train for 2 epochs. Both models were trained on a single Nvidia H100 GPU. On average, SFT requires 1 hour per epoch, and GRPO requires 2.5 hours per epoch. During evaluation, the audio together with the question and 4 options pass to the model. Evaluation uses exact matching of the correct answer with the response. The predicted answer is extracted from the fragment of the response that contains the answer.

3.4 Results

3.4.1 Ablation Analysis

Table 3.1 presents two baselines. The *No training* baseline is a default Qwen2-Audio-7B-Instruct model without any training on the datasets in Section 3.3.2. The *SFT + GRPO* baseline is a default Qwen2-Audio-7B-Instruct model fine-tuned with supervised fine-tuning (SFT) and reinforcement learning (GRPO) without filtering or curriculum learning.

The analysis of Table 3.1 shows two conclusions about the performance of the audio question answering model. First, statistical diversity filtering emerges as the

performance driver. Configurations with weaker filtering parameters ($\theta = 0.7$) consistently outperform those with stronger filtering ($\theta = 0.3$), showing performance improvements of +3.1 to +3.3% versus +2.2 to +2.6%. This shows that both no filtering and overly aggressive filtering are counterproductive, so a balanced approach is necessary.

Second, curriculum design functions as a complementary factor that focuses on the benefits of data filtering. Easy-only curricula (< 0.3 difficulty) outperform hard-only curricula (> 0.7 difficulty), with an average 0.2 - 0.3% discrepancy (see the curriculum part in Table 3.1).

Per-subset analysis of the effects of diversity filtering and curriculum learning differs across evaluation subsets. Bioacoustics QA shows the largest gains: the easy-only curriculum (< 0.3) improves performance by approximately 5 percentage points, while diversity filtering ($\theta = 0.7$) contributes an additional 3 percentage points. This benefit on bioacoustic data supports the hypothesis that underrepresented categories (such as marine mammal sounds) gain more from balanced training, as diversity filtering places more weight on less common acoustic events. Temporal Soundscapes QA exhibits gains roughly similar to the overall average. Complex QA shows more modest improvements of 0.5-0.6 percentage points, which is expected given that this subset already constitutes the largest portion of the evaluation data and is better represented in the training distribution.

3.4.2 Alternative Approaches Tried

Several other configurations were tried and underperformed; these are discussed in more detail below.

Guided decoding [514] is a technique to constrain the model’s output to a valid multiple-choice format using a regular expression. By compiling the expression into a finite state machine, a logit mask forces the model to select only tokens that conform to the required syntax, such as “A, B, C, or D”. This is useful for restricting the model to answer multiple-choice questions and to be able to extract the answer afterward. Guided decoding variants performed noticeably worse (-11.2%). This is in line with findings in [443] that strong format constraints can hinder reasoning ability.

Training without LoRA during SFT collapsed performance across all tested diversity and curriculum settings to 30.7pp decrease. Low-Rank Adaptation (LoRA) [179] is a parameter-efficient fine-tuning technique that adapts a small subset of the ALM’s weights. Removing it means performing full fine-tuning, which is computationally expensive and can also lead to catastrophic forgetting. This reinforces that SFT with LoRA is necessary and that the RL benefits are contingent on curated data.

	SFT div.	GRPO curr.	GRPO div.	Gain
<i>Baselines</i>				
No training	-	-	-	-15.9%
SFT + GRPO	0.0	1.0	0.0	+0.0%
<i>Diversity filtering and curriculum learning effects</i>				
	0.3	> 0.7	0.3	+2.3%
	0.5	> 0.7	0.5	+2.2%
	<u>0.7</u>	> 0.7	<u>0.7</u>	+3.1%
	0.3	< 0.3	0.3	+2.6%
	0.5	< 0.3	0.5	+2.4%
	<u>0.7</u>	< 0.3	<u>0.7</u>	+3.3%
<i>Other configurations</i>				
Guided decoding				-11.2%
Training without LoRA				-30.7%
Multi-Stage Pipeline				-0.6%
Reasoning				-10.6%
Only diversity filtering				+2.1%
Only curriculum learning				+2.0%

Table 3.1: Overall results and ablation on the benchmark dataset. (gain in accuracy in %). Best performing configurations are indicated with dashed lines for diversity filtering, the best ones are bold, and dotted lines for curriculum learning, the best ones are underlined.

Multi-Stage Pipeline, where the ALM is trained in three stages: (1) SFT on all datasets with LoRA. (2) GRPO in easy samples with diversity balancing. (3) GRPO in the entire dataset with diversity balancing. Each stage builds on previous checkpoints. The pipeline combines SFT initialization with progressive GRPO improvements and did not return any further gains. This pipeline failed with -0.6% performance.

Reasoning to let the ALM generate intermediate reasoning steps before producing final answers resulted in a lower performance (46.2-51.6% accuracy). This limitation stems primarily from the dataset characteristics: among the six datasets used, only AudSem contains reasoning annotations. Although the theoretical capability of reinforcement learning to develop reasoning abilities by evaluating only final responses helps, having no reasoning in supervised training data results in a barrier in the training process.

3.5 Conclusion

This chapter presents a systematic training pipeline that integrates programmatic data balancing and curriculum-guided reinforcement learning to address Audio Question

Answering challenges. The approach achieves performance gains without requiring extensive computational resources or proprietary datasets.

The primary finding demonstrates that intelligent data filtering through statistical balancing delivers performance improvements. Another finding is that both no filtering and overly aggressive filtering are counterproductive, so a balanced approach is key. The methodology uses Phi-4-mini for content categorization and applies statistical thresholds to filter overrepresented classes, establishing data quality as a fundamental driver of progress in audio-language tasks.

The curriculum learning component stabilizes reinforcement learning by establishing reliable reward signals before introducing complex examples. This approach makes reinforcement learning useful for complex reasoning tasks and maintains training stability.

The best performing configuration had a weak filtering parameter ($\theta = 0.7$) and a curriculum learning strategy of easy-only. This configuration achieved +3.3% over the baseline. Various other configurations were tried and did not work, including guided decoding, training without LoRA, multi-stage pipeline, and reasoning analysis.

These results show the importance of data strategy in model training. This chapter shows that quality should be prioritized over quantity. Researchers should focus on curating balanced, high-quality training data through systematic filtering and statistical analysis to improve model performance and reduce computational requirements. The code to reproduce these experiments is available at <https://github.com/G1JS/data-advantage-aqa>.

3.6 Acknowledgment

This work was supported by the Dutch Research Council (NWO 406.20.GO.030 to Prof. Elia Formisano), the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-12157, Data Science Research Infrastructure (DSRI; Maastricht University) and the Dutch Province of Limburg.

AudSemThinker: Enhancing Audio-Language Models through Reasoning over Semantics of Sound



ADAPTED FROM

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. “AudSemThinker: Enhancing Audio-Language Models through Reasoning over Semantics of Sound”. In: *Neural Information Processing Systems (NeurIPS)*. San Diego, Dec. 2025

Abstract

Audio-language models have shown promising results in many sound understanding tasks, yet they remain limited in their ability to reason over the fine-grained semantics of sound. This chapter presents AUDSEMT THINKER, a model whose reasoning is structured around a framework of auditory semantics inspired by human cognition. To support this, we introduce AUDSEM, a dataset specifically curated for semantic descriptor reasoning in audio-language models. AUDSEM addresses the persistent challenge of data contamination in zero-shot evaluations by providing a filtered collection of audio samples paired with captions generated through a multi-stage pipeline. Experiments on MMAU, AudioBench, and OmniBench show that AUDSEMT THINKER matches or exceeds the best reported results for several tasks and training settings. Both AUDSEMT THINKER and the AUDSEM dataset are publicly released¹.

Preface

This chapter answers Research Question 3 (Semantic-Descriptor-Based Human Inspired Reasoning) by introducing a reasoning framework that decomposes audio into semantic descriptors before producing an answer or caption. It contributes to the human-inspired structure theme by showing that descriptor-based reasoning yields more explicit, inspectable outputs and improves benchmark performance, and by introducing a curated dataset designed to reduce overlap with existing benchmarks. The semantic descriptor framework introduced here is also used in the evaluation metric in Chapter 5.

This chapter was accepted at Neural Information Processing Systems (NeurIPS) conference 2025 [510]. The NeurIPS version included a brief summary of the data creation process in the main body and a detailed description in the appendix, constrained by the 10-page limit (excluding appendix). In this dissertation, the two descriptions are merged into a single, comprehensive account in the main text. Other appendix material (reasoning length analysis, prompts, etc.) remains in the appendix.

¹<https://github.com/GLJS/AudSemThinker>

4.1 Introduction

Recent foundation models, such as OpenAI’s o1 [349] and DeepSeek R1 [148], have shown that incorporating explicit reasoning phases before generating answers improves performance across tasks. Despite these advances, audio-language models have rarely implemented such structured reasoning approaches. Although general foundation models increasingly separate generation into explicit “thinking” and “answering” stages, audio understanding models still rely largely on end-to-end architectures that directly map audio into text without intermediate reasoning steps. This gap represents a promising opportunity to improve audio understanding capabilities through more deliberate reasoning mechanisms.

Several techniques have emerged to extend test-time reasoning in a controlled and scalable manner. One such method is budget forcing, which inserts “wait” tokens during generation to lengthen the model’s reasoning process [316]. Other approaches rely on reinforcement learning to manage the model’s thinking budget (the computational effort allocated during inference). For example, Length-Controlled Policy Optimization (LCPO) [6] jointly optimizes for accuracy and reasoning length, while cosine-shaped reward functions have been used to penalize overly long yet incorrect outputs [82]. In particular, scaling test-time compute has shown benefits, with smaller models such as LLaMA 3B outperforming much larger ones, such as LLaMA 70B, when allowed for more extensive reasoning during inference [32].

The limited diversity of training data is a parallel challenge in audio-language modeling. The field has become highly compartmentalized, with most models trained on overlapping datasets sourced primarily from AudioSet [126] and Freesound [117]. This redundancy is reflected in widely used benchmarks such as AudioCaps [199], WavCaps [296], and various derivatives. For example, studies report that 17.6% of AudioCaps overlap with WavCaps, while Clotho [102] exhibits even higher overlaps - up to 89% - with other datasets [228]. Such cross-contamination makes datasets more similar and reduces the degree to which audio-language models handle new data (see Chapter 2).

To address the dual challenges of limited reasoning in audio-language models and the lack of diverse training data, we introduce AUDSEMTHINKER (Audio Semantics Thinker), a model that grounds its reasoning process in a structured framework of auditory semantics inspired by human cognition. Specifically, the model analyzes audio in terms of functional components, including sound-generating agents (who), physical sound sources (what), generation mechanisms (how), and contextual cues (when/where). This framework is based on theoretical foundations [123] and experimental research [152, 40] that investigate the perceptual and cognitive mechanisms that drive human perception of complex sounds [133]. To support the development of AUDSEMTHINKER, we also construct AUDSEM, a new dataset synthesized from YouTube closed captions, designed to enhance training diversity and reduce overlap with existing resources. Using the AUDSEM dataset, we train AUDSEMTHINKER using

two paradigms, Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO), and evaluate its performance on a range of established benchmarks that span both general audio foundation tasks and audio-based dialog.

The main contributions are as follows.

- **AUDSEMTHINKER** — a novel audio-language model that integrates structured reasoning with semantic descriptor analysis, achieving state-of-the-art performance across multiple tasks.
- **AUDSEM** — a newly curated dataset built from YouTube closed captions, specifically designed to reduce overlap with existing datasets. Sound descriptions in AUDSEM are generated via a multi-stage synthetic pipeline, to provide both diversity and high quality.

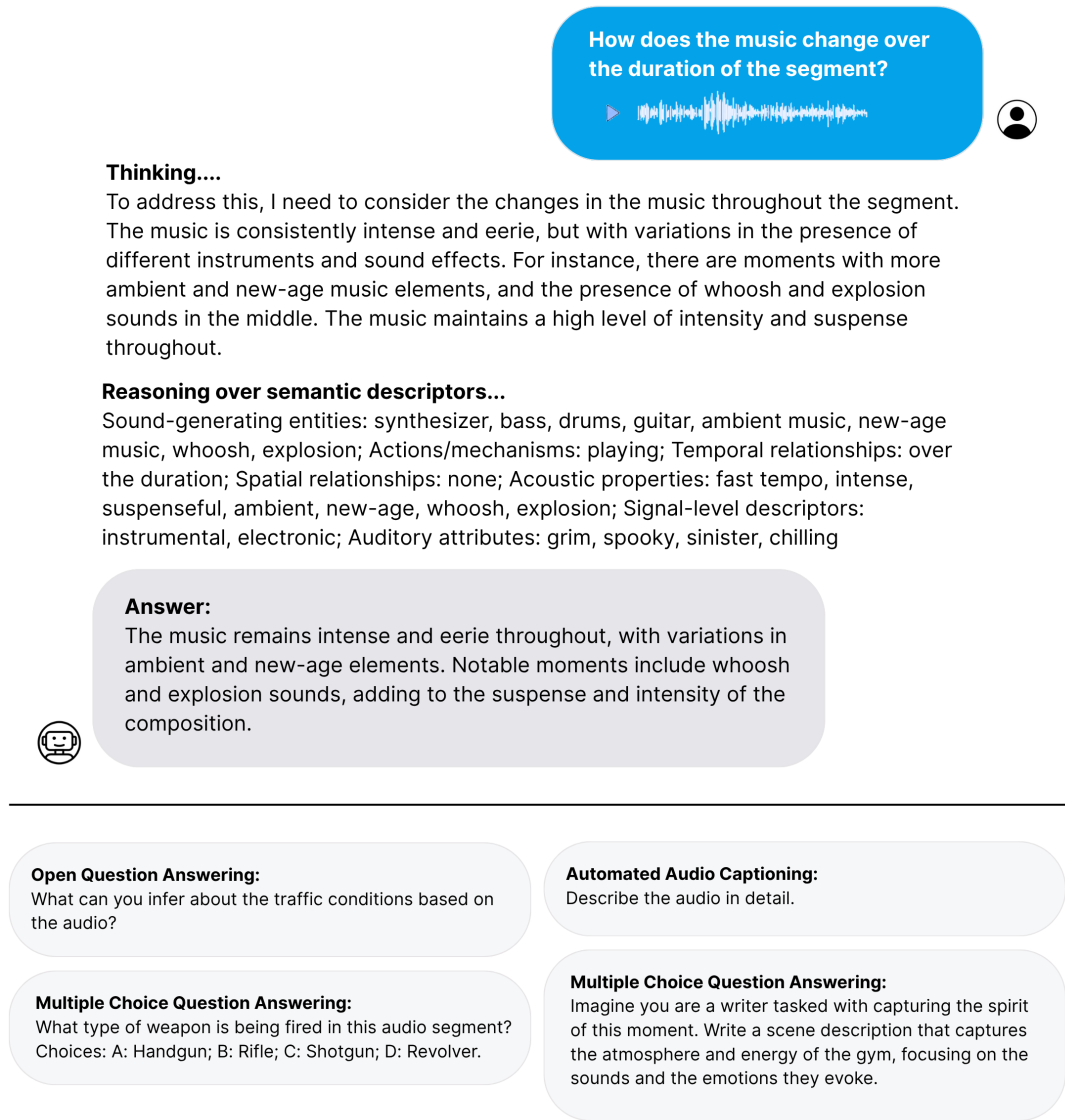


Figure 4.1: **Schematic overview of the AUDSEMTHINKER model and the AUDSEM dataset.** Top: The AUDSEMTHINKER model’s reasoning process, involving a thinking phase, semantic descriptor analysis, and answer generation. Bottom: Example of four types of tasks in the AUDSEM dataset: open-ended question answering, multiple-choice question answering, audio captioning, and creative writing.

4.2 Related Work

4.2.1 Audio Datasets with Synthetically Generated Captions

Early audio-language datasets relied primarily on manual annotation or existing meta-data. For example, AudioCaps [199] employed Amazon Mechanical Turk workers to generate captions for AudioSet clips. Subsequently, the SAM-S dataset [163] introduced an innovative approach by extracting sound-specific data from closed caption transcripts of Hollywood movies, which yielded 116,000 captions. A shift occurred with the advent of AI-assisted dataset creation for audio models. In the audio domain, 80% of new datasets created in the first half of 2024 incorporated LLM-generated captions, a marked increase from 69.2% in 2023 (see Chapter 2). One notable example is Auto-ACD [435], which employed multiple AI models to process both audio and visual information from VGGSound and AudioSet. The dataset creation pipeline integrated seven different models, including BLIP-2 [244] for image captioning and the CLIP model [373] for categorization. Other examples of AI-assisted dataset creation for audio models include Sound-VECaps [573], which employed CogVLM [486] for visual captioning and LLaMA 3 [142] for generating audio descriptions. This approach produced two complementary variants of the dataset: Sound-VECaps_F, which incorporates detailed visual features, and Sound-VECaps_A, which focus solely on audible content. Building on this, AudioSetCaps [23] introduced a more refined three-stage pipeline that combines large audio-language models for content extraction, LLMs for caption generation, and CLAP-based refinement [525], resulting in 1.9 million high-quality audio-caption pairs.

Our AUDSEM dataset creation methodology advances beyond prior work by integrating closed caption audio, video, and YouTube data through a novel pipeline. This approach employs an ensemble of specialized AI models for multimodal analysis. Table 4.1 compares our pipeline with other synthetic audio caption datasets.

4.2.2 Audio Models with Reasoning

Recent research has focused on integrating reasoning capabilities into audio-language models. An example is Mellow [87], a lightweight audio-language model that demonstrated strong reasoning abilities. Despite having only 167 million parameters and being trained on 1.24 million examples, Mellow outperformed larger state-of-the-art models across various domains. Another prominent model, Audio-Reasoner [541], was specifically designed for deep reasoning in audio tasks. It introduced a structured reasoning process using a large-scale dataset (CoTA) and a multi-phase “thinking” architecture consisting of planning, captioning, reasoning, and summarization before generating its final response. Audio-CoT [279] explored the integration of Chain-of-Thought (CoT) prompts within large audio-language models, revealing a positive correlation between the length of reasoning paths and model accuracy.

Model	Video	Audio	Texts	Image	No AudioSet	Reasoning Phase	Text Input	Models	Length
WavCaps [296]	✗	✓	✓	✗	✗	✗	Data	1	44.87
AudioSetCaps [23]	✗	✓	✓	✗	✗	✗	Labels	4	179.20
AF-AudioSet [214]	✗	✓	✗	✗	✗	✗	-	1	143.57
Auto-ACD [435]	✗	✓	✓	✓	✗	✗	Labels	7	103.87
Sound-VECaps _F [573]	✓	✓	✓	✗	✗	✗	Labels	3	180.72
AUDSEM (<i>ours</i>)	✓	✓	✓	✓	✓	✓	CC	9	852.63

Table 4.1: **Comparison of synthetic audio caption dataset creation pipelines.** The AUDSEM pipeline uses all four modalities (video, audio, text, and image) for its data across nine models. Instead of AudioSet, closed captions were used in its creation. The AUDSEM dataset contains thinking and semantic descriptor steps, and has a higher average character length compared to other datasets.

Our work builds on these advances, but introduces three key innovations. AUDSEM-THINKER applies a novel semantic analysis framework that decomposes audio into interpretable components—who, what, how, and when/where—enabling a more detailed and human-aligned understanding of audio scenes. Furthermore, it is trained on AudSem, a curated dataset designed to mitigate the data contamination and redundancy issues prevalent in existing audio-language resources.

4.3 AudSem Dataset

The AUDSEM dataset was created through a multi-stage pipeline designed to generate high-quality, diverse audio-language data while minimizing overlap with existing resources. The pipeline, visualized in Figure 4.2, involved the stages described below. A distinctive strength of AUDSEM is its fully automated creation process. While the initial closed captions are manually created by the uploaders of the video on the YouTube platform, the subsequent stages of dataset generation require no additional human intervention. Using YouTube’s closed caption subtitles as ground truth for audio events, automatic consistency checks are implemented between the generated audio descriptions and the original captions. Misaligned entries are filtered out, providing an implicit quality control mechanism without manual review. This approach enables a scalable and resource-efficient pipeline capable of producing high-quality descriptions for millions of audio examples.

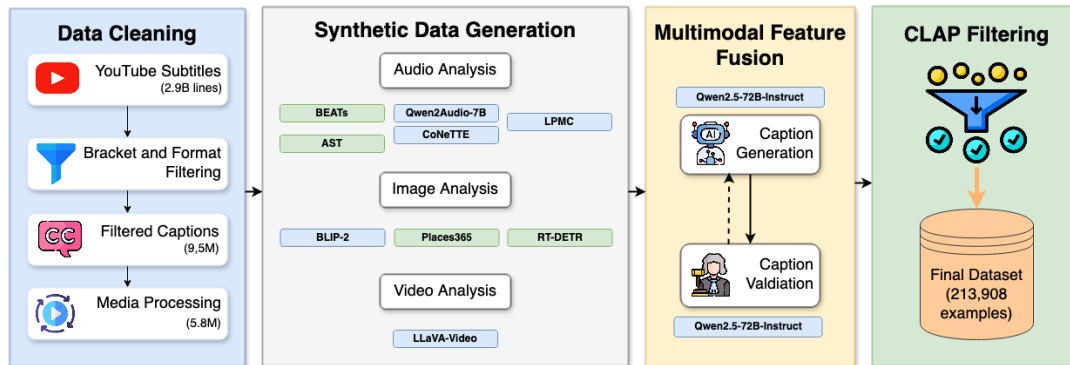


Figure 4.2: **Pipeline visualization of the creation of AUDSEM.** Models with blue background are generative language models, while models with green background are classification models that predict outputs from a fixed set of possible labels.

4.3.1 Dataset Creation

The dataset creation process begins with a large-scale collection of manually annotated English subtitle data from YouTube videos up to 2024, provided by Filmot,² a website specializing in searching through YouTube’s subtitle repository. This collection contains approximately 2.9 billion subtitle lines (288 GB of text data) from 18,641,150 videos through 31 December 2023.

From these, 75,794,563 lines (2.57% of the total) start and end with brackets, parentheses, or curly braces, excluding dialog and non-audio related text. This filtering uses regular expressions to identify genuine sound descriptions, following a common convention for Subtitles for Deaf and Hard of Hearing (SDH) captions where sound effects and audio descriptions are enclosed in brackets.

The preprocessing stage incorporates two primary filtering steps to ensure proper formatting of caption text: (1) Captions longer than 10 seconds and shorter than one second are removed. (2) New lines are removed, quotation marks are standardized, and additional filtering removes non-ASCII characters. The dataset still contained subtitles at this point that are not sound descriptions, such as lyrics to songs and narrator descriptions.

Filtering is also performed to ensure the quality of the dataset using two different language models. First, a BERT-based classifier [91] is trained on a manually annotated subset of captions to identify genuine sound descriptions. The training data consists of balanced positive and negative examples, where positive examples represent actual sound effects, and negative examples include dialog, lyrics, and other non-sound descriptions.

²filmot.com

A second classification pass uses the Mixtral-8x7B-Instruct model [187] in a zero-shot setting to verify the sound descriptions. The model is asked to identify whether captions represent genuine audio events based on specific criteria, including sound-related verbs, impact descriptions, and audio property descriptions (see the prompt in Appendix 4.8.4). Next to a Mixtral-based model, a small sample of 1000 examples of closed caption subtitles from movies and TV shows is used to train a simple BERT sentence classification model, which is used to filter out the captions as well. Only captions that are classified as sound descriptions by both models are kept. The remaining dataset contained 9,464,882 captions.

The video segments corresponding to these captions are downloaded using `yt-dlp`, a YouTube video downloader. For each caption, only the relevant audio segment is extracted based on the subtitle timestamps. The download process includes extracting precise time segments using `yt-dlp`'s `--download-sections` parameter before specifying the video ID, to bulk download only the specific segments for each video in a fast manner. The trade-off here is that the resulting extracted video may contain black frames at the beginning of the video. This is a known issue with `yt-dlp` and related tools due to the way YouTube videos are encoded. A part of the dataset contains no video or audio at all, which was removed. Another issue is that some videos are not available for download or are restricted. The remaining dataset contains 6,965,224 video clips, approximately 13.53 TB.

Using `ffmpeg`, a tool for video processing, each video is resized to 360p with two frames per second and converted to MP4 format. The audio is converted to WAV format, in 32kHz sample rate with 16-bit depth and one channel. This is done to ensure that the audio and video data are compatible with the models used in the experiments, and also to reduce the size of the dataset and improve the efficiency of the data loading process.

To further improve the data loading process, the dataset is packaged into WebDataset tar files for efficient data loading and processing. In WebDataset, each dataset entry contains paired files, an MP4 audio file pairs with corresponding JSON metadata. The Tar files contain 4,096 samples each. The corresponding JSON metadata are uploaded to the HuggingFace Dataset Repository³.

4.3.2 Dataset Processing

The dataset processing pipeline employs three categories of models for video analysis: audio models, image models, and a video model. Each model extracts different modalities of information from the video segments.

The audio analysis pipeline uses a multi-model ensemble approach to analyze the audio content. The audio is first preprocessed by resampling to 16kHz and 32kHz to

³<https://huggingface.co/datasets/gijs/audsem>

accommodate different model requirements. For each audio segment, five specialized models are used:

1. **Qwen2Audio-7B-Instruct** [72], a fine-tuned large language model for audio understanding, generates detailed descriptions in natural language of the audio content. The model processes the 16kHz audio waveform along with a prompt requesting a detailed audio description. The model outputs are decoded using a beam search with a maximum length of 1024 tokens.
2. **BEATs** (Bidirectional Encoder representation from Audio Transformers) [60], pretrained on AudioSet-2M, performs multi-label audio tagging. The model analyzes the padded audio waveforms and outputs probability scores for audio events. The top five predicted tags and their confidence scores are retained for each segment.
3. **AST** (Audio Spectrogram Transformer) [137], fine-tuned on AudioSet and ESC-50, provides audio event classification. The model processes audio spectrograms and classifies them into predefined categories from the AudioSet ontology. This provides an additional layer of audio event recognition.
4. **CoNeTTE** [228] model analyzes the 32kHz audio to generate contextual audio event descriptions. The model uses contrastive learning to map audio segments to semantic descriptions of audio events.
5. **LP-MusicCaps Model** [96], a specialized BART-based music captioning model, trained on the LP-MusicCaps dataset, is used for music-containing segments. The model generates natural language descriptions of musical characteristics when music is detected in the audio.

The output of all models is combined into a structured format containing the general audio caption, confidence scores-based audio tags, event classifications, contextual descriptions, and music-specific captions when applicable. This analysis captures different aspects of the audio content, from general descriptions to specific event detection and musical characteristics.

For visual analysis, the image processing pipeline employs multiple vision models that work in parallel. First, video frames are extracted at regular intervals; for videos under two seconds, frames are sampled at the 1/3 and 2/3 points, while longer videos have frames spread evenly across their duration up to a maximum of four frames. Black frames are filtered out by checking the mean pixel intensity.

The extracted frames are processed through four specialized models:

1. **BLIP** [245], a vision-language model, generates captions in natural language for each frame. The model uses a beam search with four beams and a maximum length of 20 tokens to generate descriptive captions.

2. **CLIP** [373] performs zero-shot classification against a predefined set of object categories. Each frame is evaluated against text prompts in the format “a photo of a [category]”, with the model’s logits being converted to probabilities through softmax normalization.
3. **RT-DETR** [276], a real-time object detection model, identifies objects in each frame with a confidence threshold of 0.3. The model’s outputs are post-processed to account for varying image sizes while maintaining accurate bounding box coordinates.
4. **A Places365-trained ResNet-50** [593] classifies the scene/environment in each frame. The frames are resized to 256x256, center-cropped to 224x224, and normalized using ImageNet statistics before being processed by the model.

The results of all frames are aggregated using a deduplication strategy that combines identical predictions while preserving their confidence scores. For each unique prediction, scores are averaged across all occurrences using scatter operations on GPU for efficiency. The final output for each video segment contains lists of object detections that are removed from duplications, as well as scene classifications, and generated captions, along with their associated confidence scores.

Video-specific features are extracted using the **LLaVA-Video-7B-Qwen2** model [582], which processes the temporal aspects of the content. For each video segment, eight frames are uniformly sampled throughout the duration, with black frames filtered out by checking mean pixel intensity. The frames are preprocessed using the model’s image processor. A prompt template is constructed for each video that includes the duration of the video, the number of frames sampled, and their precise timestamps. The model processes these inputs and generates detailed captions up to 2048 tokens long. Generation uses a deterministic decoding strategy to ensure consistent outputs. The model is compiled using PyTorch 2.0 for optimized performance and processes videos in batches of 32 using multiple worker processes for efficient data loading. The generated captions capture both the temporal progression and spatial relationships between objects and actions in the video segments, providing context that might be missed by frame-by-frame analysis.

The dataset undergoes multiple filtering steps to ensure quality. From the initial 5,332,211 results, outliers are removed on the basis of embedding distances. The method computes the average embedding using CLAP and measures the cosine distance between this embedding and all elements, for both audio and text embeddings. Items with a cosine distance greater than 0.9 are filtered out. Additionally, the method filters out items where the cosine similarity between the generated audio caption and the original YouTube closed caption is below 0.5. This ensures that the audio content aligns with its textual description. Audio samples shorter than three seconds are also omitted to improve model training on substantive audio content. After applying these filtering criteria, the final dataset consists of 213,908 results.

4.3.3 Dataset Generation

The dataset generation process combines the outputs from the three modality-specific pipelines into a structured format. Initial evaluations of five large language models (Qwen 2.5 [371], QwQ [372], Llama 3.3 [142], Mistral-Small 3 [307] and Deepseek-based models R1 Llama and R1 Qwen [148]) show that Qwen 2.5 performs better, particularly in maintaining consistent JSON output formats. Based on these results, the final dataset generation pipeline uses Qwen2.5-72B-Instruct, implementing structured generation with a schema through the xgrammar library [99] with the help of vLLM [226]. This model with a high number of parameters was chosen due to the tendency that larger models often function as teacher models for smaller student models with knowledge distillation and compression [178]. The generation process follows a two-phase prompt and three-phase reasoning: a thinking phase for reasoning, a semantic descriptor phase for the key semantic descriptors, and an answer phase for the final caption, both enforced through structured JSON output (see Appendix 4.8.4).

Initial testing identified challenges with generation consistency and thought trace length. The thinking phase requires detailed reasoning about the audio scene, analyzing primary and background sounds, key events, activities, and the acoustic environment. For this, the prompt requests at least 50 words for the thought phase and a maximum of 50 words for the audio caption. A separate judging model validates the results by assessing adherence to the generation guidelines. The judge verifies that outputs do not incorporate pre-existing model outputs and maintain focus on audio-relevant information. If an output fails validation, the system attempts regeneration up to five times before skipping the problematic input.

With these 212k examples, four different datasets are created for both with and without added semantic descriptors:

1. An audio captioning dataset, where the instruction is “Describe the audio in detail”.
2. A multiple-choice audio question answering dataset, where the instruction is generated together with four choices of answers.
3. An open-ended audio question answering dataset, where the instruction is generated together with a text field for the answer.
4. An audio captioning dataset based on creative writing and story-generation.

Audio captioning generates one caption per example. All other datasets generate 2-3 examples per input. The full dataset with added semantic descriptors contains 797k examples. The full dataset without semantic descriptors contains 873k examples. The split is approximately 20% for (1), 25% for (2), 50% for (3) and 5% for (4). The total GPU compute time for the dataset generation, including preliminary and failed experiments, was about 10k hours. When comparing against other datasets, there

is an overlap of only 12 examples with AudioSet, with one example also being in AudioCaps, and 0 in VGGSound.

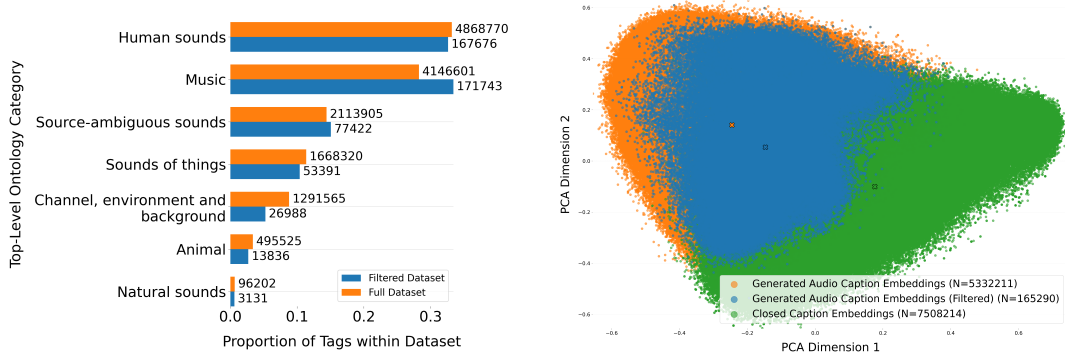


Figure 4.3: **Visual analysis of AUDSEM dataset characteristics**, showing distributions of sound categories (left) and PCA projection of caption embeddings (right). The filtered dataset was created by filtering out audio caption embeddings with less than 0.5 similarity with the closed captions.

Both configurations use XML-style tags to distinguish between phases. In the two-phase approach, the model’s output is structured with `<thinking>` and `<answer>` tags. The three-phase approach introduces an additional `<semantic_elements>` section between thinking and answering, allowing the model to explicitly represent key semantic descriptors before generating the final caption. See also Appendix 4.8.3 with Figure 4.5 for an example of an output from the two-phase approach, and Figure 4.6 for an example of an output from the three-phase approach.

Figure 4.3 provides information on the composition and refinement of the AUDSEM dataset. The left panel details the distribution of broad sound event categories, derived by mapping specific AudioSet labels to their top-level taxonomic parents. Music labels in particular are kept in the filtered dataset. The right panel visualizes the semantic space of captions through PCA-reduced embeddings. The figure displays the original closed captions sourced from YouTube (green), the audio captions generated by the Qwen2-Audio-7B model (blue), and the resultant filtered samples that constitute AUDSEM (orange). The filtering criterion, a cosine similarity score greater than 0.5 between the original and generated caption embeddings, aims to select for high agreement and relevance.

A systematic evaluation was conducted on a randomly selected subset of 100 examples from the dataset, with three annotators assessing both the correctness and depth of responses on a Likert scale from 1 to 5. For caption correctness, defined as how accurately the captions describe the events occurring in the audio, the evaluation achieved moderate to substantial inter-annotator agreement across multiple metrics: Krippendorff’s Alpha of 0.538, ICC(2,1) of 0.589, and weighted Kappa values ranging from 0.464 to 0.529 using linear weighting. The rating distributions demonstrated

strong consistency, with 70-77% of the samples receiving the highest rating (5/5) in all three annotators. For response depth, although inter-annotator agreement was lower (Krippendorff’s Alpha of 0.207), the ratings remained consistently high, with more than 85% of samples receiving the highest score.

4.3.4 Semantic Descriptors

The core idea behind incorporating semantic descriptors into model reasoning is to capture the detailed aspects of audio that a human listener would naturally perceive (see also [133, 512]). Before generating a final caption, the model first engages in a structured “thinking” phase, during which it attempts to identify and describe specific components of the audio signal. This process enables the model to move beyond surface-level recognition toward a deeper understanding of what it is “hearing”. The descriptors cover both objective and subjective dimensions of auditory perception. Objective descriptors include elements that produce the sound, such as sound-generating agents and physical sound sources, while subjective descriptors address how the sound is perceived, including auditory attributes and non-auditory sensations. This structured reasoning approach allows the model to decompose complex audio scenes into their core semantic descriptors, resulting in more informative and interpretable outputs. See Table 4.2 for definitions of the semantic descriptors used in our approach.

Semantic Descriptor	Description
Sound-Generating Agents (Who)	Denotes animated beings that generate sounds, such as people, birds, or animals. Only includes living beings, not pronouns or articles.
Physical Sound Sources (What)	Physical objects and substances that generate sounds, acting as subjects or objects in sound-related actions (e.g., bells, cars, shoes).
Sound Generation Mechanisms (How)	Actions and mechanisms of sound generation, including abstract nouns and verbs describing specific actions or events (e.g., chirping, walking, ringing).
Temporal Context (When)	Specific time periods or events providing temporal context for sound generation (e.g., morning, holidays), excluding temporal conjunctions.
Spatial Context (Where)	Locations relative to the listener or specific environments where sounds occur (e.g., background, train station, room).
Acoustic Surfaces (What/Where)	Physical surfaces and materials that contribute to the acoustic properties of the sound event.
Signal Descriptors (Sound Type)	Signal-level acoustic descriptors and basic sound classifications (e.g., noise, chord) without reference to production method, including adverbs that describe sound actions like <i>loudly</i> , <i>softly</i> , <i>rhythmically</i> , and words that characterize the acoustic signal itself such as <i>buzzing</i> , <i>humming</i> , or <i>ringing</i> .
Auditory Attributes (Sound Property)	Descriptors of the auditory sensation itself (e.g., loud, soft, steady), excluding source-related adjectives. This includes words that characterize how a sound is heard like <i>dull</i> , <i>hard</i> , <i>steadily</i> , <i>noisily</i> , <i>continually</i> , <i>with a ding</i> , or quantifiers like <i>repeated 8 times</i> .
Non-auditory Sensation	Non-acoustic attributes and emotional descriptors of perceived sounds (e.g., beautiful, relaxing, calm), including subjective impressions of the sound scene like <i>quiet in quiet bus stop</i> or <i>calm in calm manner</i> .

Table 4.2: **Semantic descriptors used in the three-phase reasoning approach.**

4.4 AudSemThinker Model

The proposed approach uses the Qwen2.5-Omni [544] model with 7B parameters. Qwen2.5-Omni is an end-to-end multimodal model that processes diverse inputs including text, images, audio, and video, and generates both text and speech outputs through its Thinker-Talker architecture. For AUDSEMTHINKER, the “Thinker” component, the large language model responsible for text generation, is specifically fine-tuned. This architecture processes multi-modal inputs and generates textual outputs. Fine-tuning occurs exclusively on the AUDSEM dataset (see Section 4.3); no other fine-tuning is performed. The core idea involves guiding the model to first engage in a “thinking” process, describe the semantic descriptors of the audio, and then produce a final “answer”. Two primary training paradigms are explored: SFT and GRPO, both using the AUDSEM dataset, demonstrating the dataset’s effectiveness in both scenarios.

4.4.1 Supervised Fine-Tuning (SFT)

In the SFT approach, the model is trained directly to produce the desired structured output. We train two primary configurations: AUDSEMT THINKER, trained on the full AUDSEM dataset, and AUDSEMT THINKER-QA, trained exclusively on the subset containing multiple-choice audio question answering examples. For both configurations, the training data is formatted into a conversational structure, including a system prompt, a user prompt containing the audio input and a question (or instruction), and an assistant’s response, which is the ground truth text containing the `<think>`, optional `<semantic_elements>`, and `<answer>` sections. The loss is computed only on the model completion part, which is the assistant’s response.

Parameter-efficient fine-tuning is used with LoRA [179], targeting projection layers. Both models are trained using the AdamW optimizer [268] with a learning rate of $2e-04$ for one epoch, using a linear decay scheduler and bf16 precision. Training uses a batch size of four on a single H100 GPU, which takes approximately 12 hours for the full dataset (AUDSEMT THINKER) and six hours for the QA subset (AUDSEMT THINKER-QA). A detailed instructional prompt is prepended to the user’s question (see Appendix 4.8.4), where the model is guided on the expected output format and reasoning depth (see Appendix 4.5.3 for a detailed ablation study).

4.4.2 Reinforcement Learning with Group Relative Policy Optimization (GRPO)

Beyond supervised fine-tuning, the applicability of reinforcement learning to the task and dataset is investigated using Group Relative Policy Optimization (GRPO) [409]. This technique has proven effective in finetuning models and has been used to train state-of-the-art models such as DeepSeek R1 [409] and Qwen2.5 [371]. The approach allows exploration of whether the AUDSEM dataset is suitable for RL-based finetuning and enables specific testing of the impact of controlling the model’s “thinking budget” during generation. The method combines GRPO with Verifiable Rewards (RLVR) [75, 430], based on a QWEN2.5-OMNI base model.

The following reward functions are defined to guide the model’s generation policy toward accuracy, structural adherence, and controlled reasoning length:

- **Accuracy Reward:** This function evaluates the accuracy of the content within the `<answer>` tags. For the multiple-choice QA tasks used in GRPO training, it uses string matching to compare the model’s selected answer directly against the ground truth choice.
- **Format Adherence Reward:** This reward encourages the model to strictly follow the prescribed XML-tag structure (`<think>`, optional `<semantic_elements>`, `<answer>`). The reward checks for the presence, correct order, and proper encapsulation of content within these tags.

- **Length Constraint Reward:** Specifically targeting the <think> phase, this reward function implements a “thinking budget”. It penalizes deviations from a target length for reasoning content, using parameters to control the strength and tolerance of the penalty. The approach is adapted from [6] (with $\alpha = 0.1$, $\delta = 0.5$) by adding a reward component for being close to the target length (see Appendix 4.8.2 for details). This encourages the model to generate thinking content near a specified target length while penalizing excessive length.

These reward functions produce scalars for each generation, which are weighted and summed. For the group of generations corresponding to each prompt, the mean and standard deviation of these summed rewards are calculated. An advantage (summed reward – group mean) is computed for each generation, and these advantages are optionally normalized using the group’s standard deviation to produce the final scalar reward signal used for policy updates.

For GRPO training, we focus exclusively on the multiple-choice question answering subset of AUDSEM (approx. 140k examples). This decision is driven by the nature of Verifiable Rewards (RLVR), which are most effective when the correctness of an answer can be easily and objectively verified. While used for Boolean (true/false) verification, RLVR adapts well to multiple-choice questions by treating the single correct option as “true” and the remaining options as “false”. This contrasts with open-ended tasks like audio captioning or free-form QA, where verifying the quality or correctness of longer, more subjective generated text is more challenging for an automated reward model.

The resulting model, AUDSEMTINKER-QA GRPO, is trained with a target thinking budget of 25 words. Similarly to SFT, LoRA [179] is used for parameter-efficient tuning. The input prompts for GRPO also include instructions that refer to the target length of thinking. For training, we use the default GRPO loss type with $\beta = 0.01$, generate six completions per prompt ($k = 6$), and employ a batch size of two per device with bf16 precision. The model is trained on four H100 GPUs for approximately 10 hours, utilizing DeepSpeed ZeRO-3 [379] and vLLM [226] for efficient training and inference.

4.5 Experiments

This section details the evaluation of the performance of AUDSEMTINKER and the effectiveness of the AUDSEM dataset. The experimental setup is described first, followed by the main results on established benchmarks, and finalized with an ablation study.

Name	Sound		Music		Speech		Avg	
	Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
Baselines								
Random Guess	26.72	25.73	24.55	26.53	26.72	25.50	26.00	25.92
Most Frequent Choice	27.02	25.73	20.35	23.73	29.12	30.33	25.50	26.50
Human (Test-Mini)	86.31	-	78.22	-	82.17	-	82.23	-
Pretrained + Supervised Finetuned Models								
GAMA 7B [131]	41.44	45.40	32.33	30.83	18.91	19.21	30.90	31.81
Qwen Audio [73]	55.25	56.73	44.00	40.90	30.03	27.95	43.10	41.86
Qwen2 Audio [72]	54.95	45.90	50.98	53.26	42.04	45.90	49.20	52.50
Mellow [87]	61.26	64.90	54.19	52.67	29.73	38.77	48.40	52.11
Gemini Pro v1.5 [125]	56.75	54.46	49.40	48.56	58.55	55.90	54.90	52.97
Audio Flamingo 2 [128]	61.56	65.10	73.95	72.90	30.93	40.26	55.48	59.42
Gemini 2.0 Flash [363]	56.46	61.73	58.68	56.53	51.65	61.53	55.60	59.93
Kimi-Audio [205]	61.68	-	73.27	-	60.66	-	65.00	-
AUDEMTINKER (<i>ours</i>)	63.06	66.10	71.56	67.47	54.04	59.67	62.90	64.41
AUDEMTINKER-QA (<i>ours</i>)	<u>61.86</u>	66.60	76.65	70.07	52.25	60.43	63.60	65.70
Finetuned with Reinforcement Learning								
Audio-Reasoner [541]	60.06	-	64.30	-	60.70	-	61.71	-
Audio-CoT [279]	61.86	-	56.29	-	55.26	-	57.80	-
R1-AQA [240]	68.77	69.76	64.37	61.40	63.66	62.70	65.60	64.36
Gemini 2.5 Flash [77]	73.87	-	65.57	-	62.16	-	67.18	-
Qwen2.5-Omni-7B [544]	67.87	-	69.16	-	59.76	-	65.60	-
AUDEMTINKER-QA GRPO (<i>ours</i>)	<u>69.67</u>	<u>69.20</u>	69.16	63.13	61.26	65.77	66.70	66.03

Table 4.3: **Performance Comparison on MMAU**. Models presented in this chapter are marked with (*ours*). Models marked with (QA) are models trained on the subset of the dataset with only audio question answering examples. All our models are only finetuned on the dataset presented in Section 4.3. All results of others models are directly derived from their respective reports. The best-performing models in each category are highlighted in **bold**, and the second-best scores are underlined.

4.5.1 Experimental Setup

Evaluation employs the two main training paradigms detailed in Section 4.4: SFT and GRPO. All models undergo fine-tuning on the AUDSEM dataset exclusively (Section 4.3). Performance assessment uses three established benchmarks:

- **MMAU** (Massive Multi-Task Audio Understanding) [393] (Table 4.3): Evaluates expert-level knowledge and complex reasoning in speech, environmental sounds, and music in 27 diverse tasks. Scores are reported in both Test-Mini (1k clips) and Test (10k clips) sets; some authors only report Test-Mini. MMAU is a multiple choice question benchmark, where the model has to select the answer out of four possible answers. MMAU uses the accuracy metric, as scores are calculated by the match between the predicted answer and the correct answer. The answer part is extracted from the model’s full response before it is rated on MMAU.
- **OmniBench** [247] (Table 4.4): A multimodal benchmark evaluating audio comprehension paired with image captions. Models receive both audio and corresponding visual descriptions, answering questions that require integrating information from both modalities. The questions include multiple-choice and open-ended formats in the speech, sound events, and music domains.
- **AudioBench** [475] (Table 4.5): A universal benchmark that covers various audio tasks, focusing on the following capabilities of the instruction. For our evaluation, we used a relevant subset that focuses on Audio Question Answering (AQA), including AudioCaps QA [199], Clotho AQA [254] and WavCaps AQA [296]. Automated Audio Captioning (AAC), including AudioCaps [199] and WavCaps [296], and Music Question Answering Music (QA), including MuchoMusic [496]. Qwen2.5-Omni is included in the benchmarks for comparison. The AQA and Music QA benchmarks are open question answering benchmarks, in contrast to MMAU. For AQA, Music QA and AAC we evaluate the full output of the model against a model-as-judge. A Llama 3 model [142] is tasked with evaluating the response against the correct answer.

4.5.2 Main Results

GRPO shows promise for specific reasoning enhancements but does not uniformly outperform SFT. Fine-tuning with GRPO can guide the model toward desired reasoning patterns or thinking lengths. AUDSEMTHINKER-QA GRPO demonstrates a performance increase over its base model, Qwen2.5-Omni-7B, and exhibits less performance divergence across different tasks compared to supervised fine-tuning. However, GRPO may not always surpass SFT, as seen in the music understanding performance of AUDSEMTHINKER-QA GRPO compared to AUDSEMTHINKER-QA. While GRPO can achieve consistent improvements without catastrophic forgetting, SFT appears more

Model	Speech	Sound Event	Music	Overall
Qwen-Audio-Chat (7B) [73]	18.39	14.66	22.64	16.64
Reka-core [453]	26.27	28.53	29.43	23.12
Audio-Flamingo (1.3B) [213]	24.78	26.98	21.51	23.82
LTU (7B) [139]	23.12	25.42	20.00	23.91
Gemini-1.5-Pro [125]	21.02	39.82	33.96	28.02
UnifiedIO2 [270]	27.15	31.13	38.87	32.49
Qwen2-Audio-7B [†] [72]	30.48	44.82	33.02	34.23
Qwen2.5-Omni [†] [544]	33.47	37.93	40.57	<u>35.27</u>
AUDSEMTHINKER (<i>ours</i>)	32.19	34.86	51.89	34.80
AUDSEMTHINKER-QA (<i>ours</i>)	<u>34.33</u>	35.63	40.57	<u>35.27</u>
AUDSEMTHINKER-QA GRPO (<i>ours</i>)	35.29	<u>42.00</u>	<u>44.34</u>	37.51

Table 4.4: **Performance comparison on OmniBench.** Models presented in this chapter are marked with (*ours*). The best-performing models are highlighted in **bold**, second best are underlined. [†] Self-evaluated score, no equivalent score by authors of benchmark exists.

Model	AQA			Music QA	AAC	
	AudioCaps QA	Clotho AQA	WavCaps QA	MuchoMusic	AudioCaps	WavCaps
Qwen-Audio [73]	50.22	61.93	42.70	59.06	47.04	32.94
MERaLiON [162]	49.78	63.15	46.32	57.79	38.00	33.98
Qwen2-Audio [72]	45.75	50.92	44.47	71.61	40.78	33.78
Phi4 [2]	38.47	47.87	35.13	54.42	26.39	21.88
SeaLLMs [329]	53.74	53.04	42.11	63.18	53.21	-
WavLLM [180]	29.84	43.01	26.25	44.31	5.50	6.90
SALMONN [444]	50.29	57.75	47.30	50.88	37.45	23.77
Qwen2.5-Omni [544]	<u>59.62</u>	59.91	<u>53.29</u>	70.09	<u>51.97</u>	41.48
AUDSEMTHINKER-QA GRPO (<i>ours</i>)	54.19	50.85	47.70	62.00	44.06	31.92
AUDSEMTHINKER-QA (<i>ours</i>)	58.40	68.74	31.83	<u>74.47</u>	39.83	31.83
AUDSEMTHINKER (<i>ours</i>)	62.81	<u>63.79</u>	55.20	76.66	44.82	<u>35.36</u>

Table 4.5: **Performance comparison on AudioBench.** Models presented in this chapter are marked with (*ours*). The benchmarks on AQA, Music QA and AAC in AudioBench were used. The best-performing models are highlighted in **bold**, and the second-best scores are underlined.

effective for substantial gains in understanding novel data, as shown by the larger performance gap between SFT-tuned models (AUDSEMTHINKER-QA, AUDSEMTHINKER) and their base model. RLVR performs well on closed caption benchmarks, but fails to generalize to AudioBench’s open audio question benchmark.

Training dataset composition directly impacts model performance. AUDSEMTHINKER excels in Music and Sound Understanding tasks, achieving exceptional results in lyrical reasoning (100%), texture interpretation (94.12%), instrumentation (91.43%), and melodic structure (90.91%). This strength stems from the rich music data in AUDSEM and the integration of LP-MusicCaps during dataset creation. AUDSEMTHINKER-QA shows modest improvements on MMAU but not AudioBench, while outperforming competitors on MuchoMusic. The model’s weaker speech understanding performance reflects our intended focus on environmental sound and music reasoning, without incorporating speech-to-text models in the data pipeline.

Length constraint reward impacts performance. Experiments investigating thinking budget constraints, detailed in Appendix 4.8.2, demonstrate the effectiveness of the GRPO length constraint reward function (Equation 4.1, using default parameters $\alpha = 0.1$ and $\delta = 0.5$). The best results on the MMAU Test-Mini benchmark were observed with thinking budgets (25 words, and 100-150 words, see Table 4.7) that allowed the model to operate within a range where the reward signal was active. Excessively long target reasoning phases, if they fall outside this effective reward range, do not necessarily translate to better performance under the current setup.

4.5.3 Ablation Study

Ablation Type			Qwen2-Audio				Qwen2.5-Omni			
Full/QA	Semantic/Simple	LoRA	Sound	Music	Speech	Avg	Sound	Music	Speech	Avg
Supervised Fine-tuning										
QA	Simple	✓	65.17	73.95	39.34	59.50	65.17	73.65	46.85	61.90
QA	Semantic	✓	67.57	73.95	42.94	61.50	64.86	74.85	51.95	63.90
QA	Semantic	✗	26.43	25.15	19.22	23.60	14.41	17.07	15.02	15.50
Full	Semantic	✗	2.7	3.29	0.3	2.1	22.22	33.83	16.52	24.20
Full	Semantic	✓	60.06	72.16	45.95	59.4	63.06	71.56	54.05	62.90
Full	Simple	✓	18.32	22.46	30.63	23.8	63.66	70.36	47.45	60.50
Group Relative Policy Optimization										
QA	Simple	✓	59.76	61.98	48.65	56.80	65.47	66.17	54.35	62.00
QA	Simple	✗	54.05	56.29	42.34	50.90	52.25	52.69	44.14	49.70
QA	Semantic	✓	60.96	61.08	49.25	57.10	63.36	62.87	56.76	61.00

Table 4.6: **Ablation study results on MMAU Test-Mini.** The table compares performance across Qwen2-Audio and Qwen2.5-Omni architectures with 3 ablation configurations.

An ablation study was conducted to investigate the impact of three training strategies. The study used the MMAU Test-Mini dataset and focused on two primary model architectures: Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B. SFT ablations involved training for one epoch on the AUDSEM dataset, taking approximately six hours for QA

and 12 hours for the full training, while GRPO ablations were trained for 10 hours (using the checkpoint at step 8400). Ablations were performed over three configurations: Full/QA indicates using all four subsets, the AAC, multiple choice AQA, open AQA, and creative writing data. QA indicates only the multiple choice AQA. Semantic/Simple describes which data configuration is used: with or without semantic descriptors. The pretrained base model is included for comparison.

The key findings of these investigations (summarized in Table 4.6) indicate that incorporating semantic descriptors generally improves performance in SFT settings, particularly when fine-tuning the QA subset. However, this benefit was not consistently observed in the GRPO setting, possibly because the GRPO reward mechanism primarily targets the correctness of the final answer and may not fully use the intermediate reasoning steps. The use of LoRA for parameter-efficient fine-tuning improved results across both SFT and GRPO, mitigating catastrophic forgetting and enhancing performance. In contrast, training models from scratch results in poorer performance.

4.6 Conclusion

In this chapter, AUDSEMT THINKER is introduced, a model that advances audio understanding through structured reasoning over semantic descriptors, and AUDSEM, a newly curated dataset designed to minimize overlap with existing resources. A new pipeline is proposed that includes filtering audio captions using the original closed caption to make sure the description still corresponds to the audio. Experimental results show that our model outperforms state-of-the-art approaches on multiple benchmarks, demonstrating particularly strong performance in music-related tasks. Ablation studies indicate that decomposing audio into semantic descriptors boosts performance when effectively integrated with the base models. Additionally, our thinking budget experiments reveal that both short (25 words) and moderately long (100-150 words) reasoning lengths that allow the model to operate within a range within the reward signal yield optimal results. These findings contribute to the field of audio-language modeling by demonstrating that explicit semantic reasoning improves model performance and enables the construction of cleaner, more diverse datasets, helping to address data contamination in zero-shot evaluations.

4.7 Acknowledgement

This work was supported by the Dutch Research Council (NWO 406.20.GO.030 to Prof. Elia Formisano), the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-12157, Data Science Research Infrastructure (DSRI; Maastricht University) and the Dutch Province of Limburg. We would like to thank Jopik from Filmot.com for making a dataset of subtitles available for us. We would like

to thank Jenia Jitsev and the LAION team for funding this work by providing compute time on both the JUWELS Booster at Jülich Supercomputer Centre (JSC) and Leonardo at CINECA Datacenter.

We gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this part of work by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS Booster at Jülich Supercomputing Centre (JSC), with access to compute provided via LAION cooperation on foundation models and datasets research.

We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium through an EuroHPC Extreme Access grant EHPC-EXT-2023E02-068, provided via LAION cooperation on foundation models and datasets research.

4.8 Appendix

4.8.1 Ethics Statement

Responsible Data Science

Dataset creation via YouTube scraping introduces a risk of inadvertently including harmful content. Mitigation strategies involved systematically checking and found no terms related to child abuse, hate speech, sexual content, and harassment. YouTube’s community guidelines, which prohibit such material, offer an additional safeguard regarding the source data.

Experimental validation involved extensive runs; ablation studies evaluated 62 models 192 times on the MMAU benchmark at various checkpoints and configurations. A performance discrepancy was observed when evaluating models in compute clusters equipped with A100 versus H100 GPUs. The AUDSEM dataset was constructed to minimize overlap with existing benchmarks, facilitating a more candid evaluation. To current knowledge, training data for AUDSEMTINKER do not include samples from the MMAU benchmark, though certainty is limited due to the obfuscation of MMAU audio filenames. It is acknowledged that the underlying pretrained model Qwen2.5-Omni may have encountered data present in test sets during its initial pretraining.

Broader Impact

The scope of this chapter is primarily fundamental and does not propose any real-world applications that could directly benefit society. Nevertheless, this research offers potential positive social impacts by advancing audio-language understanding. Applications could include enhanced audio transcription and captioning for people

who are deaf or hard of hearing, sophisticated monitoring systems for environmental sounds such as avian populations, and automated closed-caption generation for multimedia.

4.8.2 Thinking Budget

To investigate the relationship between thinking length and model performance, the experiments systematically tested different thinking budget configurations. The study examined whether allowing the model to perform more extensive reasoning (using more tokens) before generating an answer improves prediction accuracy across different audio understanding tasks. For this experiment, we implemented a controlled approach using GRPO with a maximum length constraint reward function. This reward function, shown in Equation 4.1, balances correctness with adherence to specified thinking lengths:

$$r(y, y_{gold}, n_{gold}) = \begin{cases} \text{clip}(1 - \alpha(n_{gold} - n_y) + \delta, 0, 1), & \text{if } n_y \leq n_{gold} \\ \text{clip}(\alpha(n_{gold} - n_y) + \delta, 0, 1), & \text{otherwise} \end{cases} \quad (4.1)$$

Where y represents the model's output, y_{gold} is the target output, n_y is the word length of the thinking section, and n_{gold} is the target thinking length. The parameters α and δ control the penalty strength and tolerance margin, respectively. In our experiments, we set $\alpha = 0.1$ and $\delta = 0.5$.

For the case where the generated thinking length is less than or equal to the target length ($n_y \leq n_{gold}$), the reward function becomes:

$$r(y, y_{gold}, n_{gold}) = \text{clip}(1 - \alpha(n_{gold} - n_y) + \delta, 0, 1) \quad (4.2)$$

This component becomes zero when:

$$n_{gold} - n_y \geq \frac{1 + \delta}{\alpha} \quad (4.3)$$

For example, with our parameter values, if $n_{gold} - n_y = \frac{1+\delta}{\alpha} = \frac{1+0.5}{0.1} = 15$, we get:

$$1 - \alpha(n_{gold} - n_y) + \delta = 1 - 0.1 \cdot 15 + 0.5 = 1 - 1.5 + 0.5 = 0 \quad (4.4)$$

For the case where the generated thinking length exceeds the target ($n_y > n_{gold}$), which corresponds to the original LPCO Max Length Constraint Mode as defined in [6], the reward function becomes:

$$r(y, y_{gold}, n_{gold}) = \text{clip}(\alpha(n_{gold} - n_y) + \delta, 0, 1) \quad (4.5)$$

This component clips to zero when:

$$n_y - n_{gold} \geq \frac{\delta}{\alpha} \quad (4.6)$$

For example, with our values, if $n_y - n_{gold} = \frac{\delta}{\alpha} = \frac{0.5}{0.1} = 5$, this term evaluates to:

$$\alpha(n_{gold} - n_y) + \delta = 0.1 \cdot (-5) + 0.5 = -0.5 + 0.5 = 0 \quad (4.7)$$

This clipping behavior has implications for training when α is 0.1: if the model’s average generated thinking length is shorter (e.g., more than 15 words below n_{gold}) than high target lengths like 125 or 150 words, it may consistently receive a zero reward component from this length constraint. This could hinder convergence toward such longer budgets, as the model does not receive a continuous gradient “nudge” from the length reward to increase its thinking length into the desired range. Although adjusting α to a much lower value could widen the range before clipping, an overly small α might diminish the length reward’s impact, potentially allowing it to be overshadowed by other rewards in the GRPO objective. In experiments, decreasing α was found to be non-effective. The reward function must have a slope (before clipping) to ensure effective gradient propagation in GRPO. The referenced Max Length Constraint Mode [6] is designed to apply a soft constraint, gradually penalizing outputs exceeding the target length instead of imposing a hard cutoff, and incentivizes token efficiency without sacrificing correctness. The term $\delta = 0.5$ helps ensure that correct answers with minor budget violations are still preferred over incorrect ones.

In the experiments, the target length parameter n_{gold} was varied in intervals of 25 words, allowing one to precisely control the model’s thinking budget and measure the impact of reasoning length on performance across various audio understanding tasks.

Table 4.7 presents the detailed results of our thinking budget experiments in different audio domains. The results show that constraining the model’s thinking to specific token lengths has varying effects on performance. Notably, the shortest (25 words) and longer (100-150 words) thinking budgets tend to yield better overall performance compared to mid-range budgets.

Since the model starts with an average output length in characters of about 95 characters, we see that for models with 125 and 150 words as a constraint, the target is too high, and the model is not able to convert to the correct target length. Targets with 25 and 100 words differentiate just enough from the original length for the length

Thinking Budget (words)	Sound	Music	Speech	Total
25	69.07	70.66	59.16	66.30
50	64.86	67.07	58.26	63.40
75	65.47	66.47	55.26	62.40
100	67.27	67.07	60.96	65.10
125	67.87	67.96	58.56	64.80
150	66.67	67.96	60.66	65.10

Table 4.7: **Performance across different thinking budget constraints.** Results show accuracy (%) on MMAU Test-Mini dataset for different audio domains when the model is constrained to use specific thinking lengths.

constraint to be effective. As a result, we see that the performance for any model above 100 words is not better than the model with 100 words as a constraint.

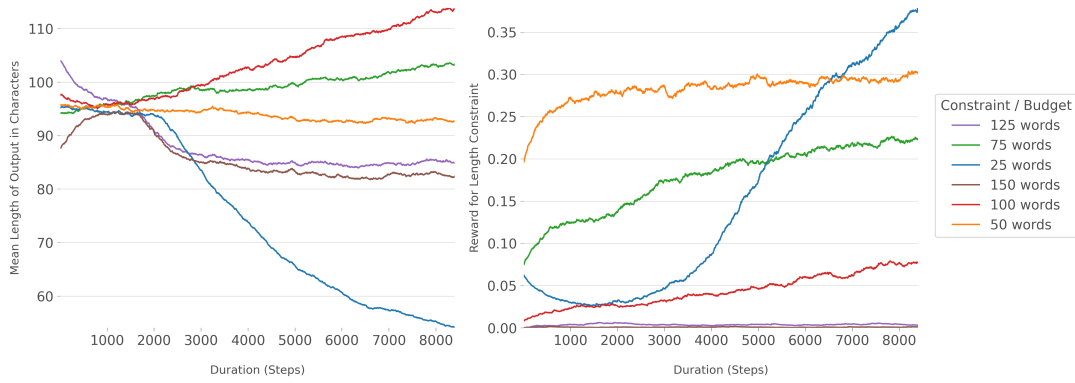


Figure 4.4: Analysis of thinking budget impact on model performance. Left: How applying a length constraint to the model’s thinking nudges the mean length of the output. Right: The relationship between various length constraints and its length constraint reward.

4.8.3 Example Model Outputs

`<thinking>`

Let's think about this. The audio seems to be set in a home environment with a dog and some background noise. Initially, there's some music playing, but it's not the main focus. The dog is definitely vocalizing, growling, and barking. There's also a man speaking, but it's not clear what he's saying. The dog's sounds are the most prominent, indicating a high level of activity from the dog. The environment feels domestic and warm, which aligns with the sounds of a pet and a person interacting. The dog's vocalizations suggest it might be responding to something, possibly the person or another stimulus. The overall atmosphere is one of interaction and engagement between the person and the dog, with the dog being quite vocal throughout the segment. The music and speech are more in the background, adding to the domestic setting but not dominating the scene.

`</thinking>`

`<answer>`

A dog is barking and growling in a home environment, with a man speaking in the background.

`</answer>`

Figure 4.5: Example of the audio-language model's output (two-phase, without semantic descriptors). The output contains a `<thinking>` section with detailed reasoning about the audio content, followed by an `<answer>` section with the concise caption.

```

<thinking>
I start by considering the primary sound, which is a seal roaring. This is a distinct, powerful
sound that stands out in the audio. The roar is followed by a series of grunts, which add to the
intensity of the scene. In the background, I hear intermittent wind noise, which suggests an
open, coastal environment. There are also some bird sounds, which are not as prominent but
add to the natural setting. The overall sound is consistent with what you might expect in a beach
environment with a large seal population. I also note some background sounds that could be
vehicles or trains, but these are less noticeable and do not seem to be the main focus of the audio.
The primary sound is the seal roaring, which is the most significant and noticeable element in
the audio.
</thinking>
<semantic_elements>
1. Sound-generating animated beings with descriptive adjectives:
- Seals roaring and grunting (powerful, intense)
2. Physical objects/substances generating sound:
- Birds (singing)
3. Actions/mechanisms of sound generation:
- Roaring and grunting (seals)
4. Temporal context if present:
- Continuous roaring followed by grunts
5. Spatial context and environment:
- Open, coastal environment with seals on a beach
6. Acoustic surfaces and materials contributing to the sound:
- Sandy beach, open air
7. Signal-level sound descriptors:
- Powerful, deep roaring
- Intermittent, soft grunting
8. Auditory sensation attributes:
- Deep, resonant roaring
- Intermittent, low-pitched grunting
9. Subjective/emotional descriptors:
- Intense, powerful, natural
</semantic_elements>
<answer>
A seal roars powerfully, followed by deep grunts, in a coastal environment with wind and birds
in the background.
</answer>

```

Figure 4.6: Example of the audio-language model's output (three-phase). The output contains an additional `<semantic_elements>` section.

4.8.4 Prompts

Filtering Prompt

Prompt for Caption Filtering

You are a friendly chatbot whose task it is to filter out bad data. You will get a closed caption corresponding to a video clip. Your task is to state whether the caption is a correct subtitle for deaf or hard-of-hearing people. Correct captions in this task are those that correspond to words that could represent an actual sound being made. This could either include a verb that states an impact or sound types or properties like "sound", "noise" or "music". Incorrect closed captions include sentences that someone is saying in the video clip, or sentences that are not related to the video clip at all. All captions are in English. All captions are within curly brackets or square brackets []. Examples of correct captions include: - "(laughs)" or "(laughter)" - "[XBOX SOUND]" - "[chicken bocking imitation]" - "(cereal grains smacking onto wood)" - "(collision)"

Examples of incorrect captions include: - "[transport]" - "(Wishes are left to wither by time.)" - "(look, I like my nightmareless sleep; I'll play some scary games when I feel too peaceful)" - "[A calm navy color] [TinyTAN character detail]" - "[Haotian Sword Tower]"

Is the following caption correct? Please only answer "yes" or "no"

Caption Generation Prompt

Prompt for Judging Captions

You are a strict judge for an audio caption generator. Your task is to verify whether the generated output adheres to all the rules from the original prompt. In particular, check the following:

1. The 'thinking' process should contain a Chain-of-Thought (CoT) reasoning process.
2. The 'thinking' process must not mention "predictions per second" or any similar phrasing.
3. The 'thinking' process must not include any of the original data fields directly.
4. The 'answer' should be a valid audio caption containing no visual elements or contexts.

Examine the generated output below carefully and respond with a JSON object that includes: - "valid": set to true if all rules are followed, or false if any rule is broken. - "reason": if false, a brief explanation of which rule(s) were violated.

— Generated Output Start — generated_output — Generated Output End —

Prompt for Caption Generation

You are an expert audio caption generator to create training data for an audio model. Your task is to create a detailed caption that describes what happens in an audio segment, including a Chain-of-Thought (CoT) reasoning process. You will be provided with various types of information extracted from audio processing models and supporting visual context. Your goal is to write a thinking process and answer as if you would only have the audio itself, without any of the following information.

Given Information: 1. Basic Information: - Video ID: video_id - Time Segment: start to end seconds - Original Closed Caption: text (This is the most important information to keep in mind) 2. Model-Generated Audio Information: - Audio Caption: audio_caption - Audio Tags (each with a confidence score): audio_tags - Short Audio Caption: conette_candidates - Predictions Per Second (key is the second, value is dict of sound and confidence score): sat_predictions music_caption_section

3. Supporting Visual Context: Scene Description: caption Detected Objects (COCO labels): objects Scene Classification (Places365): places

Context Evaluation Guidelines: 1. Use visual information ONLY if it: a) Strongly aligns with AND confirms audio evidence b) Provides essential acoustic environment context unavailable from audio 2. Ignore visual information if: a) Contradicts audio evidence b) Talks about text/graphics/static images c) Describes visual-only elements NEVER mention the visual context or visual elements in the thinking step, ONLY use it to infer the audio context.

In your output in the thinking step, analyze the audio scene in detail, reason about the primary and background sounds, and describe what happens in the audio. Include key events and activities, and the environment and context.

Guidelines: - Use natural, descriptive language - The thinking should be at least 50 words - Keep the final caption under 50 words - Do not include timestamps - Do not mention specific speech content unless crucial to understanding the audio scene - Use the prediction per second to determine the order of the sounds in the caption - The reasoning process MUST include thought expressions in natural language. This includes discourse markers, hesitation phrases, cognitive markers and casual suggestions. - NEVER describe or mention the original data fields directly in your reasoning process. You are generating training data for an audio model, and the model should learn to reason from the audio itself and NOT from the extracted data given including any of the visual context.

DO NOT mention any of the outputs of the models in the thinking step.

Please provide your response as a JSON object with the following keys: - thinking: The thinking process - answer: The caption

Prepended Prompt on Model Generation

Prompt for Generation with Semantic Descriptors

You are given a question and an audio clip. Your task is to answer the question based on the audio clip. First, think about the question and the audio clip and put your thoughts in <think> and </think> tags. Then reason about the semantic elements involved in the audio clip and put your reasoning in <semantic_elements> and </semantic_elements> tags. Then answer the question based on the audio clip, put your answer in <answer> and </answer> tags.

Prompt for Generation without Semantic Descriptors

You are given a question and an audio clip. Your task is to answer the question based on the audio clip. First, think about the question and the audio clip and put your thoughts in `<think>` and `</think>` tags. Then answer the question based on the audio clip, put your answer in `<answer>` and `</answer>` tags.

ACES: Evaluating Automated Audio Captioning Models on the Semantics of Sounds



ADAPTED FROM

Gijs Wijngaard, Elia Formisano, Bruno L. Giordano, and Michel Dumontier. “ACES: Evaluating Automated Audio Captioning Models on the Semantics of Sounds”. In: *2023 31st European Signal Processing Conference (EUSIPCO)*. Helsinki, Finland: IEEE, Sept. 2023, pp. 770–774. doi: 10.23919/EUSIPCO58844.2023.10289793

Abstract

Automated Audio Captioning is a multimodal task that aims to convert audio content into natural language. The assessment of audio captioning systems is based on quantitative metrics applied to text data. Previous studies have used metrics derived from machine translation and image captioning to evaluate the quality of generated audio captions. Drawing inspiration from auditory cognitive neuroscience research, we introduce a metric approach: Audio Captioning Evaluation on Semantics of Sound (ACES). ACES takes into account how human listeners parse semantic information from sounds, providing an alternative evaluation perspective for automated audio captioning systems. ACES combines semantic similarities and semantic entity labeling. On the Clotho-Eval FENSE benchmark, ACES outperforms comparable automated audio captioning metrics in two evaluation categories.

Preface

This chapter answers Research Question 4 (Semantics-Aware Evaluation Metric for Audio Captioning) by introducing ACES, a metric that evaluates captions through extracted semantic sound descriptors (who, what, how, where, and when) and their embedding similarity. It contributes to the human-inspired structure theme by aligning evaluation with audible semantics rather than n-gram overlap and by providing diagnostic signals about which descriptor types are missing from a caption.

This chapter was accepted at the European Signal Processing Conference (EUSIPCO) 2023 [512]. The EUSIPCO version had the 4-page limit, and presented a simpler metric formulation based on averaged cosine similarity and label overlap, and reported that ACES did not yet match the performance of other metrics on the FENSE benchmark. The current version in this chapter redesigns the metric around max-aligned semantic-token similarity with recall-weighted F-score, integrates a fluency penalty, and adds hyperparameter tuning. This redesign enables ACES to outperform comparable metrics in two FENSE evaluation categories on Clotho-Eval.

5.1 Introduction

Automated audio captioning (AAC) is an emerging field of audio processing. As introduced in 2017 by Drossos et al. [101], the goal of AAC is to describe the content of an audio clip using natural language, that is, using structured text that describes sound. The performance of AAC models is estimated using metrics that compare model-predicted captions with human annotations.

Standard AAC models are based on the encoder-decoder architecture [293], where the encoder converts the input audio to an embedding, which is then fed to the decoder. In this architecture, the decoder learns to generate the audio caption by minimizing the cross-entropy loss on the probabilities of the decoder output. During inference, the decoder calculates the most probable sentence given the embedded audio.

AAC models are benchmarked using metrics tailored to various criteria (see Section 5.2.1, below). Additionally, it is possible to benchmark the metrics themselves and gauge their alignment with human judgment. In this context, Zhou et al. [600] introduced the FENSE benchmark, which evaluates how AAC metrics perform compared to human evaluation.

Research in the field of auditory cognition and neuroscience has shown that humans listen to sounds to derive information about sources, events, and changes in the environment and that this information is reflected in listeners' verbal descriptions of everyday sounds [133]. When asked to describe sounds, listeners refer to the presence of animate (*who*) or inanimate (*what*) sources, identify sound generation mechanisms or actions (*how*), and, eventually, to a spatial (*where*) or temporal (*when*) context (see Table 5.2, and [133], for the derivation of these classes of descriptors from previously published taxonomies of natural sounds).

This chapter proposes a novel metric to evaluate audio captioning algorithms on annotated datasets: **Audio Captioning Evaluation on Semantics of Sounds (ACES)**¹. ACES draws inspiration from auditory cognition research on the extraction of sound descriptors: words or phrases that describe who or what makes the sound, where it occurs, and how the sound is produced (see Table 5.2 for a list of all categories). This is related to semantic role labeling, which is the process of assigning labels to words to indicate their semantic role [38]. ACES uses the cosine similarity of vector representations of these sound descriptors to measure the correspondence between generated and “ground truth” captions. ACES returns a score that indicates the quality of the generated caption based on these ground truths.

In short, the ACES score consists of a cosine similarity of the pairwise token combination. ACES uses precision and recall from this cosine similarity, and takes the token with the highest score as overlap. ACES also uses a METEOR [27] based approach in weighing precision and recall, a small penalty for longer captions, and a fluency error detection model. This work is an extension of an earlier version of ACES, presented

¹Code, data and models available at: <https://github.com/GIJS/ACES>.

at the European Signal Processing Conference 2023 (EUSIPCO). The current version of ACES supersedes the initial release, rendering the previous version obsolete and deprecated.

5.2 Related work

The evaluation metrics from the Microsoft COCO Caption Dataset represent a set of audio captioning evaluation algorithms [64]. This set of evaluation metrics was originally intended to measure performance in image captioning tasks but has since been adopted in other domains, including audio captioning. The toolset includes the metrics BLEU [355], ROUGE [251], METEOR [27], and CIDEr [471], and, in later COCO versions, SPICE [13]. The current standard in AAC metrics are SPIDEr and its derivative SPIDEr-FL. SPIDEr is a combination of CIDEr and SPICE, which outperforms both metrics based on human evaluation of a randomly sampled COCO test set (10.56% increase compared to a baseline MLE model) [259]. SPIDEr-FL is SPIDEr enhanced with an additional pretrained fluency detection model, specifically designed to identify incomplete sentences and sentences with repeated or missing words [600]. For an overview of recent metrics for evaluating Automated Audio Captioning models, see Table 5.1.

5.2.1 Drawbacks of current metrics

Performance metrics currently adopted in the field of AAC are characterized by several drawbacks. For example, BLEU, ROUGE, CIDEr, and METEOR metrics are sensitive to n-gram overlap, which is neither necessary nor sufficient to establish meaning similarity between sentences [13]. For example, the captions “Rain coming from a big cloud”, and “Music coming from a big band”, are dissimilar in meaning but have a 4-gram in common, resulting in a high n-gram score. Although SPICE was proposed to mitigate this problem, it makes the overly stringent assumption that the compared captions have exactly the same wording for its entities, attributes, and relations. For example, the candidate sentence “Young woman talking with crunching noise” and the reference sentence “Paper crackling with female speaking lightly in the background.” result in a SPICE score of 0, despite their clear affinity in meaning [600]. In other cases, the caption may not include text descriptors of the sound-generating agents or objects (e.g. “Metallic scraping that stops and then starts again”) or descriptors of actions (e.g. “Very loud static sound without any other noise”). In such cases, a scene graph (a way of encoding objects, attributes, and relations) cannot be composed or calculated, and the SPICE score returns a low value regardless of semantic relatedness.

Here, a novel metric is proposed based on semantic role structure (e.g. who did what to whom). The Semantic-role was previously considered by Lo et al. [267], who proposed the MEANT metric in the context of Machine Translation. A drawback of this metric is the use of syntactic parsing of the sentence, which results in a low score

Table 5.1: Overview of metrics used for evaluating Automated Audio Captioning models

Metric	Description	Primary use
BLEU _{1..4}	Uses the co-occurrences of n-grams. Adds a brevity penalty.	Machine translation
METEOR	Harmonic mean of precision and recall of caption chunks. Recall weighted 9 times more than precision. Matches also stems, synonyms and paraphrases	Machine translation
ROUGE	Calculates F-measure using, longest common subsequence. Oriented to recall with $\beta = 1.2$.	Text summarization
SPICE	Matches captions on object classes, relation types and attributes.	Image captioning
CIDER	Calculates a weighted sum of the cosine similarity for n-grams with $n \in [1, 4]$. Uses TF-IDF.	Image captioning
BertScore	Calculate F1, precision, and recall of matched words from the pairwise cosine similarity of the sentence embeddings.	Text generation
BLEURT	BERT model trained on human annotated dataset of sentence pairs,	Text generation
FENSE	Cosine similarity of sentence embeddings combined with a fluency detection model	Audio captioning
SPICE+	Combines SPICE with matching using sentence embeddings	Audio captioning

when synonyms are used. Pretrained models such as BLEURT [406] and BERTScore [579] have also been used as a metric for various machine learning tasks. One of the benefits of these models is that you can extract a contextualized embedding that captures the meaning of the term, and that similar meanings are correlated with the computational similarity derived from a cosine of the two embeddings.

The comparison of sentences based on lexical structure has been investigated in recent studies on Automated Audio Captioning. The FENSE metric [600] uses a language model to calculate similarity and includes a fluency penalty model to capture coherent structures in audio captions. Specifically, FENSE penalizes captions in five categories: incomplete sentences, repeated events, repeated adverbs, missing conjunctions, and missing verbs. This fluency penalty was also added to the SPIDEr metric in the DCASE Challenge 2023 Task 6a [427]. The SPICE+ metric [141] finally tries to solve the issues that arise with its counterpart, SPICE, by performing the evaluation within a language model framework.

5.2.2 Research in other fields

Analogies can be drawn between our research and active areas within natural language processing (NLP), specifically: (i) semantic role labeling (SRL), where predicate-argument structures are modeled from sentences, and (ii) part-of-speech tagging (POS), where words are tagged based on their grammatical functions (e.g. subject or verb). Although our dataset is directly labeled by us, there is a clear relation between our labels and the classifications that an automated SRL or POS model would assign. For example, the *ARG-0* and *V* labels in a SRL model correspond, respectively, to the WHO and HOW property in our model.

Our model performs entity recognition by labeling terms, similar to the MEANT metric, and incorporates pretrained masked language models (BERT and its derivative models) to capture the sound descriptors of captions. Furthermore, similar to the SPIDEr-FL score, our metric integrates the fluency penalty from the FENSE metric. In summary, our approach involves computing the cosine similarity of the embeddings, focusing on the correspondence of sound descriptor categories. We also incorporate a fluency penalty into our approach and, in the absence of overlap, default to sentence-BERT embeddings.

5.3 Methods

The backbone of the proposed metric is a NLP model that is capable of classifying words from human (or model) generated captions into a set of sound descriptor categories. These categories reflect semantic attributes of the sound-generating objects and events that listeners derive from sounds, and have been derived from a recent survey of taxonomies of everyday sounds [133] (see Table 5.2). This word

Table 5.2: Sound descriptor categories used in calculating correspondence between captions. Items marked with * exist only in the dataset of 13 labels.

Label	Description
WHO	sound-generating agent
WHO/WHAT PROPERTY*	describes object or person
WHAT	vibrating object or substance
HOW	sound-generating actions/mechanisms
HOW PROPERTY*	specifies action
WHEN	temporal context
WHERE	spatial context
WHAT/WHERE*	objects that contribute to acoustics
SOUND TYPE	sound-signal categories
SOUND PROPERTY	acoustic/auditory sound properties
NON-AUDITORY SENSATION	non-auditory attributes of sound
OTHER	labels that do not describe sound
O	omitted labels

model is then applied to both the candidate and the reference captions, and a score is calculated that reflects the similarity of their sound descriptor categories.

5.3.1 Model training

To obtain this word-classification model, a random subset of captions was annotated from the Clotho dataset. The Clotho dataset is widely used in AAC and consists of 29,645 captions for 5,929 audio excerpts, each 15 to 30 seconds long [102]. Various pre-trained NLP models were fine-tuned to this dataset, and their performance compared to determine the most effective classifier (see below). Specifically, two word-labeled caption datasets were generated using the Prodigy web annotation tool [370]. In the initial training iteration, every word in each caption was labeled with one category from a set of 10 (dataset 1). For the final model, a set of 13 labels was used (dataset 2) (see Table 5.2).

Dataset 1 consists of 2300 captions labeled by 3 annotators, with 1120 captions being duplicates across annotators for the estimation of annotator agreement statistics. The annotators labeled 694, 1387 and 219 labels captions, respectively. The average inter-annotator agreements (Cohen’s kappa coefficient), averaged across annotator pairs, and annotations, were 0.808, 0.836 and 0.839. To train our models, duplicates were filtered out to ensure that they did not lead to biased results from training items multiple times per epoch or data leakage in the test set, leaving 1158 captions in the final dataset. Of these 1158 captions, the annotators labeled 285, 727 and 146, respectively.

Table 5.3: Comparison of BERT-based models on token classification of our dataset. The highest scores are in bold. The groups of models are 10 labels (top) and 13 labels (bottom).

Name	F1	F1 How	F1 What	F1 Where	F1 Who
10 Labels					
BERT	81.6	88.3	80.6	78.5	85.3
RoBERTa	81.2	89.5	79.3	79.2	80.0
XLNet	79.7	85.8	80.5	78.0	81.0
ALBERT	80.6	87.0	81.2	76.8	83.7
DeBERTa	79.7	85.6	80.1	75.6	83.9
13 Labels					
BERT	80.3	90.8	80.6	86.0	91.2
RoBERTa	84.2	91.8	84.7	92.3	90.9
XLNet	84.2	93.7	85.6	83.9	92.9
ALBERT	83.0	91.4	87.5	84.5	95.5
DeBERTa	80.9	88.7	87.1	84.7	87.9

We applied a similar procedure to dataset 2, which was labeled by two annotators, with a Cohen’s kappa coefficient of 0.794. Dataset 2 is a subset of dataset 1, but contains more descriptor categories. Dataset 2 initially contained 989 captions, where the annotators covered 494 and 495 captions, respectively. After removing duplicates, the final dataset contained 500 unique captions. Although in dataset 2 there were more sound descriptors and less training data than dataset 1, this resulted in an increase in classification F1 scores.

Using the 2 labeled datasets, various pretrained Transformer encoder models were fine-tuned with a classification head (see Table 5.3), as implemented in the HuggingFace Transformers library [516]. During tokenization, each caption was split into tokens, each corresponding to a sound descriptor. Tokens from words that did not correspond to a sound descriptor in our dataset were assigned the label O and omitted. For training, we used an AdamW optimizer [268], with a learning rate of $2e-5$ and weight decay. The labeled dataset was divided into 80% train and 20% test sets. The model was trained for 5 epochs, since initial tests showed that after 5 epochs overfitting might occur, as identified by observing a growing difference between the training and validation losses after the fifth epoch. For each trained NLP model, the averaged and category-specific (Who-What-How-Where) F1-scores on the test set for the two distinct variants of the models were obtained using 10 (top) or 13 (bottom) labels (Table 5.3). This F1-score gives an indication how well the model is capable of classifying the sound descriptors on unseen sentences.

5.3.2 Metric definition

The fine-tuned model is then applied to both the candidate and the reference captions. For example, for the sentence “a person is walking on a hard surface”, groups of tokens

person, walking on and *hard surface* are categorized into *WHO*, *HOW* and *WHAT/WHERE*, respectively. In the case of 5 reference captions instead of 1 reference caption, the words per category in each of the reference captions are combined (e.g. *WHO* in “bird caws” “bird croaks” becomes “bird bird”).

For each of these group of words, the corresponding sentence embed is calculated with SBERT [381]. The cosine similarity is computed on overlapping sound descriptors (e.g. *WHO* in candidate and reference). This ensures a focus on entities present in the candidate text and optimizes for better wording per descriptor category. However, a drawback is that it may neglect entities found in reference texts that do not exist in the candidate (see Appendix 5.8.2 for examples).

The cosine similarity of the overlap is calculated as follows. For each candidate and reference caption C and R , for each candidate and reference token $c \in C$ and $r \in R$, given the sound descriptor of c, r : $\mathbb{c}, \mathbb{r} \in \mathbb{C}_l \cap \mathbb{R}_l$, the cosine similarity is calculated as:

$$\text{CosSim}(c, r) = \frac{E(c)^\top \cdot E(r)}{\|E(c)\| \|E(r)\|}$$

with $E(\cdot)$ representing the embedding of the token. As we use pre-normalized vectors, cosine similarity is reduced to $E(c)^\top \cdot E(r)$. This cosine similarity score indicates the relation of each pair of tokens so that similar tokens in sound descriptors also get a high score, contrary to using the n-gram.

From this cosine similarity, the maximum score of precision overlap (Pr) and recall overlap (Re) of the candidate and reference tokens is taken, similar to BertScore [579].

$$Pr(\mathbb{c}, \mathbb{r}) = \frac{1}{|\mathbb{r}|} \sum_{r \in \mathbb{r}} \max_{c \in \mathbb{c}} \text{CosSim}(c, r)$$

$$Re(\mathbb{c}, \mathbb{r}) = \frac{1}{|\mathbb{c}|} \sum_{c \in \mathbb{c}} \max_{r \in \mathbb{r}} \text{CosSim}(c, r)$$

The F-score is calculated by weighting the recall of the token 9 times more than the precision, as done similarly for the METEOR metric, and has been shown to improve correlation with human judgment [27]. Consistently, we also observed a positive correlation between the emphasis on recall and a better outcome on the FENSE benchmark.

The ACES metric puts focus on the token with the highest similarity score between the candidate and the references due to the maximization in the above formula. This approach has limitations. For instance, it may overlook irrelevant tokens even if they are categorized similarly. Additionally, the metric does not consider captions composed of compound sentences. In such cases, the ACES score emphasizes only the category in one part of the compound sentence with the greatest similarity, disregarding the rest.

$$F_{\text{score}}(\mathbb{C}, \mathbb{R}) = \frac{10PrRe}{Re + 9Pr}$$

All embeddings are then averaged:

$$F_{\text{single}}(C, R) = \frac{1}{|C \cap R|} \sum_{\mathbb{C}, \mathbb{R} \in C \cap R} F_{\text{score}}(\mathbb{C}, \mathbb{R})$$

If there is no overlap at all, we set the score to 0. As a last step, a small penalty is added to our scoring method. This penalty is designed to slightly decrease the scores of shorter sentences that include fewer entities. The penalty formula is as follows:

$$\text{penalty} = (|\mathbb{L}| - |\mathbb{C}_l \cap \mathbb{R}_l|) \frac{1}{|\mathbb{L}|} \frac{1}{1850}$$

$\mathbb{L}$ is defined by the total number of possible sound descriptors, in this case 13. $|\mathbb{C}_l \cap \mathbb{R}_l|$ is the number of sound descriptors that are the same in both the candidate and the reference captions. This penalty is subtracted from the F-score for each pair of candidate and reference sentences, resulting in the following formula:

$$\text{ACES}_1(C, R) = F_{\text{single}}(C, R) - \text{penalty}$$

The normalization constant 1850 was determined empirically to gently penalize shorter sentences without allowing the penalty to dominate the semantic similarity score. However, tests showed that higher penalties did not lead to better performance on the FENSE benchmark. By applying only a slight penalty, a preference for detailed sentences is still included, and the metric still correlates well with human evaluation.

For each candidate and reference caption C and R a single value is returned. An extra penalty for fluency error detection is added, similar to the one in the FENSE and SPIDER-FL metrics [600]. The best results were found by putting a weight of 0.5 instead of 0.9, the default value in the FENSE metric (see the appendix 5.8.1). For every single candidate and reference caption, ACES is computed as:

$$\text{ACES}(C, R) = \text{ACES}_1(C, R) - 0.5 \times \text{fluency}(C) \times \text{ACES}_1(C, R)$$

A combined ACES score for a whole dataset is calculated by taking the average across the pairs of candidate and reference captions.

5.3.3 Hyperparameter tuning

To find the optimal parameters for our model, a hyperparameter tuning phase is performed. During hyperparameter tuning, the library WandB and its sweep function with Bayesian optimization [36] carry out a search through 18 different parameters

(see Appendix 5.8.1, for an overview). Bayesian optimization narrows the search scope by focusing on the parameters that show the most promise. When the parameters are adjusted, the total Clotho-Eval score is maximized, with the aim of outperforming the other models on the FENSE benchmark (see Table 5.5). The emphasis for maximization is put on the Clotho-Eval benchmark, since its captions have longer sentences than Audiocaps-Eval (11.334 vs 8.796 words on average) and contain on average more entities (5.433 vs. 4.742). However, the initial results showed that optimizing for one benchmark also resulted in better results for the other benchmark.

The final parameters are from a run of a model that specifically outperforms some values on the FENSE benchmark and has a good overall result. Multiple configurations of the ACES model outperformed other metrics in the benchmark, but the final configuration also included values that resembled previous metric research in the METEOR and Fluency detection methods [27, 600].

5.3.4 Optimization

Since the use case for this benchmark is during model validation, optimizing the model for implementation speed is a priority. The final model requires about 6.4 GB of RAM for the token classification part, and the embedding model uses about 1.2 GB of RAM. This is optimized by quantizing the model weights to float16 precision, resulting in a model size reduction of about 50%. In addition, the model can easily perform distributed calculations by automatically scaling up to multi-GPU inference with the Accelerate Python package [146].

Additionally, relying on a smaller version of the RoBERTa model (RoBERTa-base instead of RoBERTa-large), we are able to achieve a 25% speedup in the computation of ACES at the cost of a negligible drop in performance.

5.4 Experiments

We evaluated our models by comparing them to similar metrics in AAC. In parallel, our models are tested against the FENSE benchmark to allow the estimation of its predictions of human evaluations. The FENSE benchmark consists of two datasets, Clotho-Eval and AudioCaps-Eval. These are derived from the Clotho and AudioCaps datasets, respectively, and contain human evaluations for 4 categories of pairs (see section 5.4.2).

5.4.1 Model evaluation

We evaluated several AAC models using the BERTScore, FENSE, SPIDER, and the proposed ACES metrics to compare the metrics against each other. The DCASE baselines for this year and last year were used as benchmarks for the metrics. The 2021 baseline

Table 5.4: Models were evaluated considering several metrics, including our proposed metric ACES.

	BERTScore	FENSE	ROUGE _L	SPIDER	ACES
Baseline 2021	82.3	1.9	27.3	6.4	15.7
AudioCaption	89.1	37.7	28.9	14.2	42.2
Baseline 2022	90.5	45.7	36.5	23.1	46.1
PANNs + BART	90.7	46.8	37.6	25.2	47.9
PASST + BART	90.8	48.3	37.8	26.0	48.8
Baseline 2023	90.7	48.3	38.7	27.0	49.3

model was an encoder-decoder architecture based on GRU’s [69], whereas the 2022 model had a VGGish encoder and a BART-based decoder. In addition to the 2022 model baseline, two other versions were added. In these two versions, the VGGish encoder network was replaced by a PANNs [212] and PaSST [217] encoder, respectively. The 2023 model baseline is also included, which uses a special pretrained CNN encoder based on PANNs.

Metrics	AudioCaps-Eval					Clotho-Eval				
	HC	HI	HM	MM	Total	HC	HI	HM	MM	Total
SPIDER	53.2	89.9	84.1	55.2	65.4	47.9	88.1	67.9	52.5	59.8
BERTScore	60.6	97.6	92.9	65	74.3	57.1	95.5	70.3	61.3	67.5
BLEURT	77.3	93.9	88.7	72.4	79.3	59	93.9	75.4	67.4	71.6
Sentence-BERT	64	99.2	92.5	73.6	79.6	60	95.5	75.9	66.9	71.8
FENSE	64.5	98.4	91.6	84.6	85.3	60.5	94.7	80.2	72.8	75.7
SPICE+	59.1	85.4	83.7	49	62	46.7	88.1	70.3	48.7	57.8
SPICE+emb	63.5	96.4	91.6	70	77	61	94.7	76.3	61.6	68.9
ACES	64.5	95.1	89.5	82.0	83.0	56.7	95.5	82.8	69.9	74.0

Table 5.5: Performance of various metrics on Audiocaps and Clotho (bold = top values).

Table 5.4 displays the evaluation results of the AAC models. ACES has more variation than BERTScore, which is helpful in discriminating results that are close to each other. The ACES scores are consistent with other metrics. In particular, BERTScore and FENSE rate PASST + BART similarly to the Baseline 2023 model. Other metrics favor the baseline 2023 model, including ACES. This could indicate that ACES is a more reliable metric than BERTScore and FENSE, although further investigation is required.

5.4.2 Human evaluation

We used the FENSE benchmark [600] to evaluate our score against human evaluations and to compare ACES to several other metrics (see Table 5.5). The FENSE benchmark comprises four components, each designed to calculate the quality of a candidate metric compared to its reference: (i) human-correct (HC), (ii) human-incorrect (HI), (iii) human-machine (HM) and (iv) machine-machine (MM). For each of these components, 4 human evaluators compare two candidate captions for each sound, and indicate which target sound is better (final outcome based on majority vote with target sound not considered in the absence of a majority). For the HC component, both candidate captions are generated by human annotators describing the same sound (human captioning data from Clotho and AudioCaps datasets) For HI, a caption generated by a human annotator for a target sound is compared with a human-generated caption for a different sound. The HM component instead compares a human caption with a caption for the same target sound generated with a captioning model. The MM component finally compares machine-generated captions from two captioning models (see [600], for details). The audio captioning metrics are finally bench marked against the majority-based human evaluation of the candidate captions, with the final score measuring the proportion of target sounds for which the captioning metric and human evaluators agreed on their choice of the best captions among the two candidates.

On this FENSE benchmark, approaches that use computational similarity of contextualized embeddings (ACES, FENSE, BERTScore, BLEURT, Sentence-BERT, SPICE+emb), are favored over approaches that use entity labeling (SPIDEr, SPICE+). This could indicate that the participants’ judgments of descriptions from the FENSE benchmark are correlated with distances between descriptions in the BERT embedding space.

5.5 Discussion

The ACES metric demonstrates comparable performance to other metrics in the FENSE benchmark and notably outperforms all other metrics in the human-incorrect (HI) and human-machine (HM) categories on the Clotho-Eval dataset. ACES provides a versatile backbone that can be used to recognize a sentence’s entities, which can be helpful for explainability purposes.

ACES calculates a score based on semantic descriptors and shows promising results in comparison to other metrics on the FENSE benchmark. A comparison between metrics with several examples is shown in Appendix 5.8.2. This appendix also illustrates some drawbacks of our approach, including cases where there is no overlap of semantic descriptors. In case of no overlap, ACES assigns the score 0. It also only takes the most similar entity into account: in case of smaller sentences with linking words, it focuses on the most similar word per semantic descriptor.

Results from candidate captions that exhibit a higher ACES score contain several sound descriptors that are high in cosine similarity of the vector representation to the ground truth vector representation. These sound descriptors are important to include in a caption, as they represent listeners' verbal descriptions of everyday sounds.

5.6 Conclusion

This chapter introduces the Audio Captioning Evaluation on Semantics of Sound (ACES), a novel AAC evaluation metric based on audio semantics. ACES combines both semantic similarities and semantic entity labeling, resulting in an approach for evaluating audio captions that departs from earlier work relying exclusively on embeddings, or n-grams, or entities. ACES combines elements from earlier research on AAC evaluation and outperforms other metrics on the FENSE benchmark.

Future research can explore several promising directions, such as the involvement of large language models such as GPT-4 to evaluate sentences for fluency issues and to grade semantic richness and factual correctness in generated captions. A strategy for this could be to train a large language model on how a human would evaluate a dataset of generated audio captions and create an automatic pipeline for the caption evaluation.

5.7 Acknowledgment

This work was supported by the Dutch Research Council (NWO 406.20.GO.030 to Prof. Elia Formisano), the French National Research Agency (ANR-21-CE37-0027-01 to Bruno L. Giordano; ANR-16-CONV-0002 – ILCB; ANR11-LABX-0036 – BLRI), Data Science Research Infrastructure (DSRI; Maastricht University) and the Dutch Province of Limburg.

5.8 Appendix

5.8.1 Overview of hyperparameters

In bold the parameter that performed the best in the hyperparameter tuning. Percentages are correlation scores with Clotho-Eval Total in total for all runs. See also the WandB Report.

- *model*: ("gijs/aces-roberta-10", "**gijs/aces-roberta-13**") Selects which pretrained ACES model to use for named entity recognition and embedding extraction. The options are the 10 label or 13 label versions, which detect different numbers of entity types.

-
- *fl_weighting*: (**True**, False) (34.9%). Determines whether a fluency model should be added to the model. This model detects sentences with errors in them and penalizes them with *F1_weight* (float between 0.0 and 1.0) (1.4%): which controls the amount the score is reduced if an error is detected.
 - *average_strategy*: ("simple", "first", "**max**", "average") Controls how the per-token predictions from the ACES model are aggregated into entities.
 - *use_sbert*: (**True**, False) (22.7%) Determines if SentenceBERT is used to calculate entity embedding similarity instead of the native ACES embeddings.
 - *division*: (float between 0.0 and 1.0. **0.998**) (-2.6%) Controls weighting between the cosine similarity score and the entity overlap score.
 - *F1_calc*: ("mean", "mean-max", "**max-mean**", "max-max"). Sets how precision and recall are combined to calculate F1 between the candidate and reference entities.
 - *F1_beta*: (float between 0.1 and 10.0. **9.0**) (1.9%) Controls beta parameter that balances precision and recall for F-beta score for the cosine similarity
 - *F1*: (float between 0.1 and 10.0. **3.798**) (9.3%) Controls the beta parameter that balances precision and recall for F-beta score for the overlap.
 - *apply_penalty*: (**True**, False) (-8.6%). Boolean for whether to apply the entity overlap penalty. This is to penalize sentences that have only a few classes, as we prefer longer more descriptive sentences over shorter ones ("bird caws noisily with an alarming tone" over just "bird caws"). *penalty_score* (int between 1 and 2000. **1850**) (-5.5%): Penalty applied when there is no entity overlap. Higher values decrease this penalty. If there are a lot of labels in the sentences, it still corresponds to a high score.
 - *overlap_type*: ("cand", "ref", "**both**", "F1"). Sets the formulation of entity overlap used. Based on the candidate, reference, both, or F1-weighted overlap.
 - *distance_technique* ("**cosine**", "Euclidean"): Distance metric to calculate entity embeddings.
 - *use_score*: ("pairwise", "mean", "**no**"). With the pairwise score, both the combination of the similarity score and the overlap of the entities per combination of sentences are calculated. "mean" corresponds to calculating the overlap after taking the mean. *score_weighting* (float between 0.0 and 1.0. **0.5**) (18.0%): Controls balance between embedding similarity and overlap when combining values.
 - *overall_sbert*: (True, **False**). Whether to incorporate an additional overall SentenceBERT similarity between the candidate and reference. *overall_sbert_weight* (float between 0.0 and 1.0. **0.5**) (-8.2%): Weighting on overall SBERT similarity

when used. *sbert_based_on_scores* (**True**, False) (4.3%): Only uses SBERT similarity if the ACES score is 0, which is the case when there are no labels at all present.

5.8.2 In depth comparison of ACES with previous metrics

This Appendix presents selected examples of the FENSE benchmark, featuring both successful and unsuccessful cases of scoring from the ACES metric. The examples are displayed in Table 5.6. The first three examples demonstrate situations where ACES outperforms FENSE and other measures. The final two instances reveal ACES's limitations.

In the initial example, “a door is followed by a”, both SPICE and ACES score 0.0 due to their reliance on overlapping semantic pairs. SPICE uses non-existing graph overlap based on its semantic graph, and ACES relies on its semantic sound categories in the captions. For ACES, it returns 0 due to “door” being the sole recognized WHAT entity and in the reference captions no WHAT entities. Other metrics, however, return similar, albeit higher, scores.

The second example highlights a more pronounced gap between the ACES and FENSE scores. The lower score for ACES arises because of the difference in length between the references and candidate captions, affecting the overlap in categories considered. The candidate caption is classified into three categories. WHO (“birds”, “young peoples”), HOW: (“cackling”). WHAT: (“voices”). For WHAT (“voices”), ACES matches it to the only WHAT in the references: “wings”, which drives the score down. However, less descriptive sentences may be the origin of the lower score.

For the third scenario, ACES again assigns a lower score compared to other metrics. This is due to the reference caption containing only WHO and HOW categories, creating a mismatch on HOW between the candidate and references. The candidate’s “rolling” and “strikes” differ from the reference’s “roars”, “falls” and “falling”. Despite similar overall overlap, they are missing some key components, i.e. no mention of lightning in any of the references.

The fourth example exhibits considerable overlaps resulting in a higher ACES score. However, certain elements, such as “in a quiet environment”, are absent. In the candidate caption, there are no similar WHERE matches, which results in ACES wrongly attributing a high score to a caption missing contextual details.

Finally, the fifth example highlights another instance where ACES scores higher on average compared to other metrics. This results from the similarity in WHO and HOW categories between the candidate and reference captions. Yet, ACES does not fully discern the difference between “speaking” and “shouting”, focusing instead on the entity with the greatest overlap, in this case, the HOWs “squeaking” and “bouncing”, which are present in both the candidate and reference captions.

Table 5.6: Candidate and reference examples from the FENSE benchmark with corresponding AAC metrics

C1	
Candidate A door is followed by a	AAC metrics METEOR 0.0265 ROUGE _L 0.1393 SPICE 0.0 BERTScore 0.9814 FENSE 0.0311 ACES 0.0
References • Some scratching and rustling with small clicks • Rustling and some knocks • Some rustling and scratching with a short click • Continuous clanking and rustling	
C2	
Candidate Birds cackling and young peoples voices	AAC metrics METEOR 0.0440 ROUGE _L 0.1680 SPICE 0.0541 BERTScore 0.9646 FENSE 0.6621 ACES 0.3230
References • A man speaking followed by pigeons cooing and flapping wings then a kid speaking as someone claps loudly • A man talking then a young boy talking followed by a loud pop as pigeons coo and bird wings flap • A man talking followed by pigeons cooing and bird wings flapping then a young man talking • A male voice speaks and a bird coos and flaps its wings	
C3	
Candidate Rolling thunder with lightning strikes	AAC metrics METEOR 0.1741 ROUGE _L 0.2618 SPICE 0.2000 BERTScore 0.9858 FENSE 0.6899 ACES 0.4755
References • Thunder and a gentle rain • Thunder roars in the distance as rain falls • Rain falling with thunder in the distance • Rain and thunder	

C4	
Candidate Typing on a computer keyboard	AAC metrics METEOR 0.2049 ROUGE _L 0.5031 SPICE 0.2500 BERTScore 0.9882 FENSE 0.7101 ACES 0.9677
References • Typing on a keyboard is occurring in a quiet environment • Typing on a keyboard is ongoing in a quiet environment	

C5	
Candidate Squeaking and bouncing followed by a man speaking	AAC metrics METEOR 0.1373 ROUGE _L 0.2254 SPICE 0.0 BERTScore 0.9752 FENSE 0.5331 ACES 0.8676
References • Several basketballs bounce and shoes squeak on a hardwood surface as a man yells in the distance • A man yelling in the background as several basketballs bounce and shoes squeak on a hardwood surface • A man yelling in the distance as several basketballs bounce and shoes squeak on hardwood floors • Multiple basketballs bounce on a hard surface and shoes squeak as a man shouts in the distance	

AudioToolAgent: An Agentic Framework for Audio-Language Models



ADAPTED FROM

Gijs Wijngaard, Elia Formisano, Michel Dumontier, and Jenia Jitsev. *AudioToolAgent: An Agentic Framework for Audio-Language Models*. Feb. 2026. doi: 10.48550/arXiv.2510.02995

Abstract

Large Audio-Language Models (LALMs) perform well on audio understanding tasks but lack the multistep reasoning and tool-calling capabilities found in recent Large Language Models (LLMs). This chapter presents AudioToolAgent, a framework that coordinates audio-language models as tools via a central LLM agent that accesses tool adapters for audio question answering and speech-to-text. The agent reasons about which tools to invoke, how to formulate follow-up queries, and how to arbitrate conflicting tool outputs, without accessing the audio. Experiments with MMAU, MMAR, and MMAU-Pro show state-of-the-art accuracy: up to 77.50% in MMAU, 77.00% in MMAR, and 61.90% in MMAU-Pro. Monte Carlo sampling for Shapley values across 490 configurations identifies effective agent-tool combinations. The modular design allows the integration of new tools and removes the need for additional training data and training runs. The code and reproduction materials are available at <https://github.com/GLJS/AudioToolAgent>.

Preface

This chapter answers Research Question 5 (Agentic Framework for Coordinating Audio Tools) by evaluating whether a text-only agent can orchestrate specialized audio tools to solve audio reasoning tasks without additional training data. It contributes to the human-inspired structure theme by using inference-time verification and tool coordination to improve performance: the agent selects tools, cross-checks evidence, and synthesizes answers in a controlled format. This chapter frames audio-language modeling as a system where outcomes depend on both tool quality and the agent's decision process.

6.1 Introduction

Understanding and reasoning about audio is central to human cognition. Recent progress in transferring this capability to machines spans two areas: the advancement of Large Language Models (LLMs) with reasoning and tool-calling capabilities [85, 449, 559, 348, 77, 124], and the development of Large Audio-Language Models (LALMs) for tasks such as audio captioning, audio question answering, and speech recognition [544, 72, 205, 269, 519, 134, 255, 374].

Recent LALMs perform well on audio benchmarks [393, 224] but most lack tool use [519]. Conversely, general LLMs excel at reasoning and using external tools but lack direct audio processing. The present work combines these strengths by enabling an LLM agent to use audio models as tools. The approach achieves more adaptive audio understanding than LALMs while maintaining the deeper reasoning capabilities of general LLMs.

The chapter introduces **AudioToolAgent**, which treats audio-language models as tools and uses a central agent to coordinate them. The agent, a text-only LLM, cannot process the audio directly; it receives only the audio file path, a question or prompt, and possible answers. The agent then reasons about which tools to invoke and delegates instructions to LALMs (tools) that process the audio. Performance gains arise from two mechanisms: (i) selective use of complementary tools, combining specialized ASR with general audio understanding models, and (ii) multi-hop reasoning across multiple tools to reduce errors from individual models. Because the framework reuses pretrained models, it requires no new datasets or training.

The agent receives an audio input, a question, and answer choices. The agent uses this information to reason about the task, upon which it calls tools to be able to answer the question or prompt. For speech, the agent prioritizes speech-to-text tools to transcribe the audio. For environmental sounds or music, the agent uses general audio models to gather information. AudioToolAgent asks follow-up questions, iteratively invokes tools, compares outputs, and verifies disagreements by continuing to call tools with different inputs to increase reliability. Figure 6.1 shows an example of the framework in a chatbot.

The contributions of this work are twofold:

1. A reproducible, agent-based coordination framework with modular tool adapters. Using pretrained foundational models without additional data or fine-tuning, this provides a cost-effective approach to high performance on MMAU [393], MMAR [280], and MMAU-Pro [224]. Both closed-source and open-source versions of AudioToolAgent reach or exceed previously reported results for several benchmark subsets.
2. A controlled empirical study across three benchmarks with a large ablation space (agents and tools) and Shapley-based contribution analysis, establishing

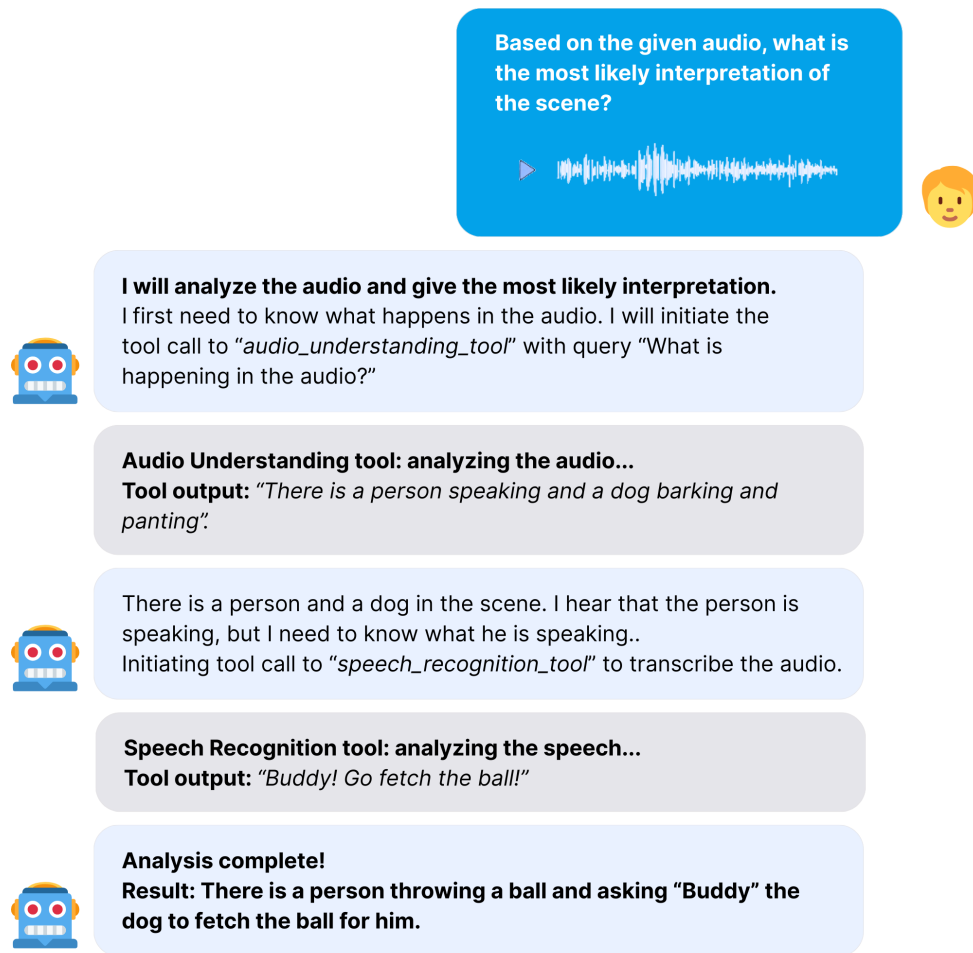


Figure 6.1: Example of a chatbot using the AudioToolAgent framework. The agent is a large language model that coordinates the tools; the tools are audio-language models.

a measurement methodology for tool-using audio systems where tools can be swapped as new models appear without retraining.

6.2 Related Work

Recent advances in large language models (LLMs) have resulted in agents who call tools to solve tasks [359]. This started with the function calling of GPT-3.5 [346] and includes the Model Context Protocol (MCP) [15], which standardizes interactions with external tools. This work uses the ReAct framework [566]. In ReAct, the agent first reasons about the task and then performs actions to solve it. The agent selects the appropriate tools to answer the question. After receiving tool responses, the

agent decides whether to make additional tool calls or answer the question with the information already gathered.

Related work in vision includes Image-of-Thought (IoT) prompting to external tools for a multimodal model to gain information [596]. The ScienceQA model [271] uses an image captioning model to provide context for a reasoning model. Similarly, SRICE [592] guides inference by allowing the model to select parts of the image through interactions with external tools.

Recent developments include large multimodal models that integrate audio processing and agentic capabilities within a single architecture. Models such as Gemini 3 [124] and GPT-4o [347] handle both audio comprehension and speech recognition while performing tool calls. Training these models costs substantial resources, and they remain closed-source, accessible only through API endpoints.

Another relevant work, StepAudio 2 [519], capable of audio understanding, received explicit training in tool call and benchmarking on four specific tools: audio search with multimodal RAG, date and time retrieval, weather search, and web search. This model processes audio input, performs tool calls, understands audio content, and generates speech output. Training this integrated model consumed 1.356 trillion tokens over 21 days [519]. In contrast, the AudioToolAgent framework eliminates this training cost by coordinating existing pretrained models.

Similarly, other works with the same name include AudioAgent [14], which uses audio attributes to optimize prompts using a fine-tuned LLM for audio tools, and the AudioAgent framework [495], which uses an LLM to orchestrate audio generation and editing. The current work differs by using a text-only agent that delegates audio understanding to specialized tools without fine-tuning. Instead of using fixed classifiers, AudioToolAgent queries multiple interchangeable tools and cross-checks their outputs for verification.

6.3 Methodology and Experimental Setup

6.3.1 Framework Overview

The agent, a reasoning model, receives an audio file path and task description and selects tools to produce the output. The agent accesses audio signals only through the tools. The agent identifies suitable tool calls for the task, then invokes them through structured tags: `<tool_call>` to initiate a request and `</tool_call>` to conclude it. Within these tags, the agent specifies the target tool, audio file path, and prompt. Each tool's output enters the agent's context, enabling it to reason and invoke additional tools as needed. The set of tools includes multiple speech recognition and audio understanding tools. Figure 6.2 shows a schematic visualization. To prevent runaway loops, each agent can invoke a maximum of 20 tool calls. In AudioToolAgent, the agent makes 5-10 calls, depending on the configuration.

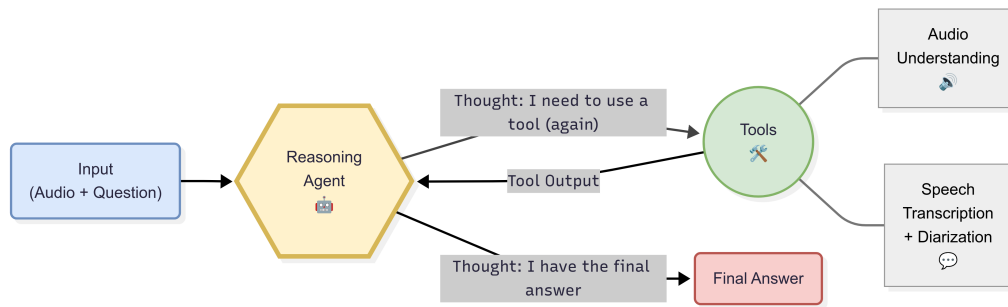


Figure 6.2: Schematic overview of the AudioToolAgent framework. The framework has two components: a central agent and a set of tool models. The agent is a large language model that coordinates the tools; the tools are audio-language models.

Instructions in the system prompt help the agent to issue follow-up tool calls when outputs conflict or remain ambiguous, and to gather targeted evidence. The prompt also directs the agent to resolve conflicting or ambiguous tool output through targeted follow-up queries and to format answers within `<answer>` tags for parsing. The system prompt starts as follows:

System Prompt

You are an expert audio analyst with access to specialized tools. Answer the question given. Put the answer between `<answer>` and `</answer>` tags. If the question is multiple choice, there is always just one choice correct. If the tool says it can't listen to audio, try invoking the tool again. Use as many tools as needed to answer the question, even using the same tool multiple times if needed. If initial tool outputs are conflicting or ambiguous, do not guess; instead, you must generate specific, follow-up tool calls to isolate the point of disagreement and gather more detailed evidence. The following tools are available ...

In this framework, all tools connect via HTTP API tool adapters for modularity. This includes public endpoints for proprietary models and self-hosted endpoints for open-source models, running on vLLM [226] or Transformers [516]. To demonstrate the versatility of the framework, the implementation offers two configurations, each exposing four tools¹:

- **AudioToolAgent:** This configuration uses a proprietary agent and closed-source tools, accessed through public API endpoints to maximize performance. The agent is Gemini 3 Pro [124], and the tool suite includes Gemini 3 Pro [124], Qwen2.5 Omni [544], Audio Flamingo 3 [134], and Whisper [374].

¹The exact tool registry and configuration files are available at <https://github.com/GLJS/AudioToolAgent/>.

-
- **AudioToolAgent-Open:** This configuration uses an open-source agent with high-performing open-source audio tools to balance performance with self-hosting capabilities. The agent is Qwen 3 235B [559], and the tool suite includes Qwen3 Omni [559], Qwen2.5 Omni [544], Audio Flamingo 3 [134], and Whisper [374].

The agent uses a multi-hop approach to reduce errors from individual tools. When the agent queries multiple tools and receives conflicting outputs, it triggers targeted re-queries: the agent formulates follow-up questions that constrain the scope to the disputed span or category (e.g., asking specifically about a contested timestamp or sound event). This iterative process continues until the evidence converges across tools or uncertainty remains explicit. By comparing answers across tools, the agent reduces reliance on any single model and mitigates biases inherent in individual tools.

6.3.2 Evaluation Setup

The study evaluated AudioToolAgent and AudioToolAgent-Open on three benchmarks: Massive Multi-Task Audio Understanding (MMAU) [393], MMAR [280] and MMAU-Pro [224]. The MMAU benchmark includes 10,000 audio clips for multi-task audio understanding and reasoning, with 1,000 in the test-mini split and 9,000 in the test split. The experiments used only the test-mini split to reduce costs. The MMAR benchmark tests deep reasoning capabilities with 1,000 audio-question-answer triplets requiring multi-step reasoning across modalities. The MMAU-Pro benchmark measures audio intelligence using 5,305 instances containing question-answer pairs generated by human experts in speech, sound, music, and combinations.

6.3.3 Ablation Study Design

To evaluate the effectiveness of different components in the AudioToolAgent framework, an ablation study examines two aspects. First, an agent comparison benchmarks the tool-use performance of state-of-the-art reasoning LLMs while holding the tool set constant. Each agent is evaluated across five independent runs with different random seeds to account for non-deterministic inference behavior. Second, a tool contribution analysis quantifies how much each audio-language model improves performance when added to different tool combinations. The study operates on a 10% subset (100 examples) of the MMAU test-mini split to balance computational cost with statistical reliability.

For tool analysis, this work uses Shapley values [410] to quantify whether adding a tool to an existing set of tools is beneficial. Shapley values account for interactions where a tool’s value depends on which other tools are present, to be able to calculate marginal contributions in multi-tool systems. To estimate Shapley values, this work uses Monte Carlo sampling on a representative small fraction of all $2^{20} = 1,048,576$

possible configurations. Without sampling, calculating all Shapley values would require evaluating every possible configuration.

The calculation consists of two stages. In the first stage, the method generates many random permutations of the available tools. For each permutation, the system calculates the accuracy on the benchmark subset of all ordered subsets (coalitions) from the empty set to the full set of tools. In the second stage, the final Shapley value for each tool is calculated across all sampled permutations. This Shapley value is the average of the tool’s marginal contribution when added to a pre-existing, non-empty set of tools. Standard errors are computed across permutations to quantify estimation uncertainty.

6.4 Results

Table 6.1 summarizes results across three benchmarks (MMAU, MMAR and MMAU-Pro). AudioToolAgent-Open achieves state-of-the-art open-source performance: 74.30% on MMAU, 65.90% on MMAR and 57.50% on MMAU-Pro, surpassing Audio Flamingo 3, Qwen2.5-Omni, and Kimi-Audio. AudioToolAgent reaches 77.50% on MMAU, 77.00% on MMAR and 61.90% on MMAU-Pro, outperforming GPT-4o Audio and Gemini 2.5 Pro.

Speech tasks show the strongest gains. AudioToolAgent achieves 78.38% on MMAU Speech and 82.99% on MMAR Speech. On MMAR speech categories, AudioToolAgent reaches 83.03% on Sound-Speech and 79.17% on Sound-Music-Speech, surpassing all baselines. This advantage comes from a specialized ASR tool (Whisper [374]), which provides value even when general omni models already support speech transcription.

ASR models trained with alignment constraints such as CTC, RNN-T, or transducer architectures return tokens only when the acoustics provides sufficient evidence, making them less prone to hallucination than general audio-language models. The ablation results in Section 6.4.1 confirm that adding dedicated ASR tools improves accuracy on speech-heavy subsets even when omni tools are present.

Beyond speech, both models perform well on cross-domain reasoning and music. On MMAR, AudioToolAgent leads all models on Music-Speech (81.71%) and Sound-Speech (83.03%); AudioToolAgent-Open leads all open-source models on every MMAR category. Music understanding is a consistent strength: AudioToolAgent reaches 72.57% on MMAU-Pro and 72.46% on MMAU, while AudioToolAgent-Open scores 65.70% and 69.46% respectively, the highest open-source results on both benchmarks.

6.4.1 Ablation Study

To identify the highest-accuracy configuration for AudioToolAgent, an ablation study examines 10% of the MMAU test-mini split (100 examples), analyzing agents and tools separately. The methodology for this analysis is also described in Section 6.3.3.

Agents

The experiments used a fixed set of five open-source tools to evaluate the LLMs capable of tool calling: Qwen2.5 Omni [544], AudioFlamingo 3 [134], DeSTA 2.5 [269], Whisper [374], and Voxtral [255]. Each evaluation ran five independent tests per agent with different random seeds and reported the mean. This approach accounts for accuracy variations from non-deterministic inference even with fixed seeds, partly due to vendor-recommended decoding defaults such as nonzero temperature. Figure 6.3 shows the results. Inspired by Omni-R1 [389], which showed that text-only models perform well on audio reasoning tasks, the black vertical tick shows each LLM’s no-tools baseline. The dots represent individual evaluations that show discrepancy between results; to combat this, the horizontal colored bar shows the average across 5 runs.

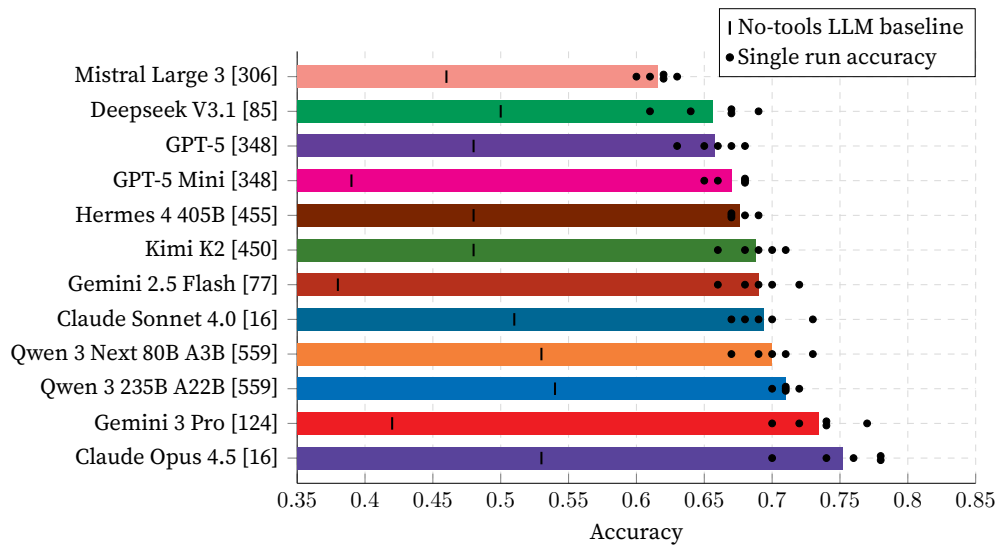


Figure 6.3: Accuracy comparison across 12 LLMs on a subset of the MMAU test-mini split. Bar is an average of 5 runs.

Claude Opus 4.5 achieves the highest mean accuracy of 0.752, followed by Gemini 3 Pro (0.734) and Qwen 3 235B (0.710). Closed-source models (Opus 4.5, Gemini 3 Pro) demonstrate the strongest orchestration capabilities, while open-source models like Qwen 3 235B and Qwen 3 Next 80B (0.700) achieve competitive results. All models benefit from tool access, with gains ranging from 15.6 percentage points (Mistral

Large 3) to 31.0 percentage points (Gemini 2.5 Flash). Consistency varies: Qwen 3 235B exhibits the narrowest accuracy range (0.02), while DeepSeek V3.1 shows higher variability (range of 0.08). The gap between the weakest agent (Mistral Large 3: 0.616) and the strongest (Opus 4.5: 0.752) spans 13.6 percentage points, demonstrating that orchestration quality is a key differentiator.

Based on these results, AudioToolAgent uses Gemini 3 Pro as the agent, since Claude Opus 4.5 was too expensive. Gemini 3 Pro also serves as an audio tool. AudioToolAgent-Open uses Qwen 3 235B as the highest-performing open-source agent.

Tools

This work applies the Shapley values methodology (see Section 6.3.3) to quantify each tool's contribution to system performance. Figure 6.4 displays the estimated Shapley values with standard error bars computed on 490 sampled configurations across 20 tools. The analysis groups the tools into 4 categories, shown on the y-axis of Figure 6.4.

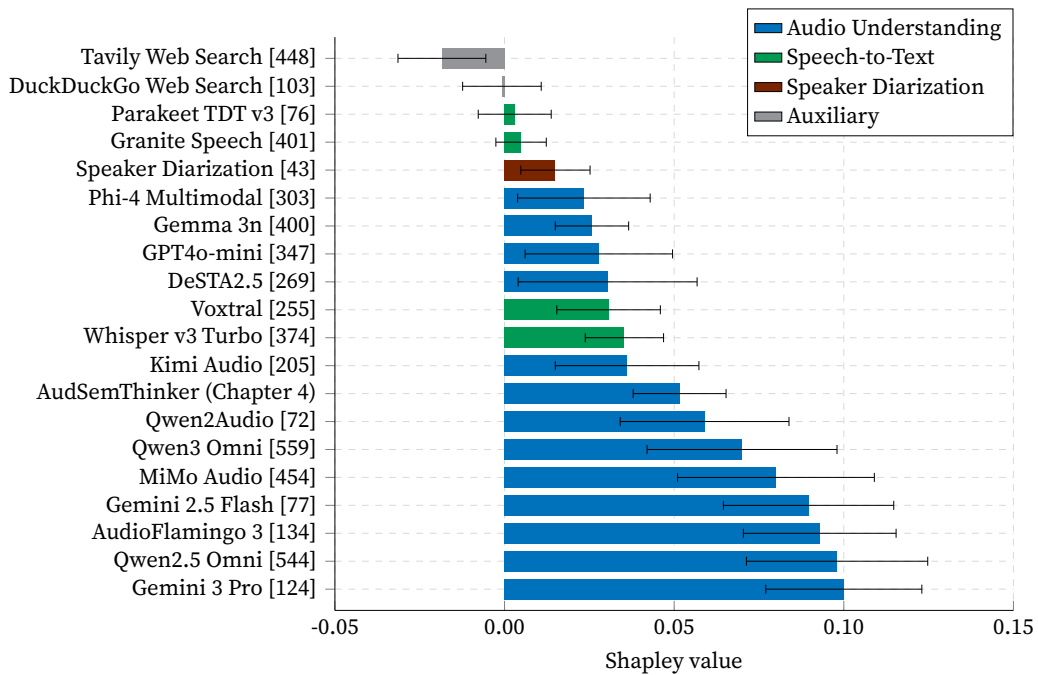


Figure 6.4: Shapley values per feature with standard error bars computed over 490 sampled configurations. Bars indicate the contribution of each audio tool to the overall system performance. Error bars are black.

The best performing tools were incorporated into the configurations of this work (see Section 6.3.1), with two exceptions: Qwen2Audio [72] processes only 30 seconds of

audio and the audios in the benchmarks are often longer, and AudSemThinker (see Chapter 4) shares its architecture with Qwen2.5 Omni.

The Shapley value analysis reveals five key findings. First, general-purpose audio models (Qwen2.5 Omni: 0.098, AudioFlamingo 3: 0.093, Gemini 2.5 Flash: 0.090) contribute 2.5–3 $\times$ more than speech-to-text models (Whisper v3 Turbo: 0.035), which shows that broader audio understanding models outweigh niche single-task models. Second, task-specific tools such as speaker diarization (0.015) provide modest value. Third, standalone benchmark performance does not predict tool effectiveness: AudioFlamingo 3 outperforms Qwen2.5 Omni on benchmarks [134] but achieves lower Shapley values (0.093 vs 0.098). Similarly, newer speech-to-text models (Parakeet TDT v3: 0.003, Granite Speech: 0.005) contribute less than older Whisper v3 Turbo (0.035) despite better transcription accuracy [76, 401]. Fourth, open-source tools match or exceed closed-source alternatives (AudioFlamingo 3: 0.093, Qwen2.5 Omni: 0.098 vs GPT4o: 0.028). Fifth, Gemini 3 Pro (0.100) is the highest contributor, while MiMo Audio (0.080) and Qwen3 Omni (0.070) are strong new additions that rank among the top tools.

6.4.2 Discussion and Future Work

The multi-hop consensus mechanism mitigates tool errors by querying multiple models and issuing targeted follow-up queries when outputs conflict, producing higher agreement rates than any single model. This approach implements a form of consensus aggregation where majority agreement across independent models provides a more reliable signal than any single model. Individual models may exhibit biases or hallucinations, but combining evidence from multiple sources reduces the impact of such errors.

The agent ablation reveals that tool-calling proficiency varies across models: the best closed-source agent (Opus 4.5: 0.752) outperforms the best open-source agent (Qwen 3 235B: 0.710) by 4.2 percentage points, while the gap between the weakest and strongest agent spans 13.6 percentage points. Orchestration quality depends on training methodology rather than model scale alone.

Shapley analysis shows that standalone benchmark performance does not predict tool effectiveness: general-purpose audio models contribute 2.5–3 $\times$ more than specialized speech-to-text tools, and AudioFlamingo 3 outperforms Qwen2.5 Omni on benchmarks yet achieves lower Shapley values (0.093 vs 0.098). Future work should identify properties beyond accuracy that make models effective tools, including output format, response consistency, and instruction following.

A practical limitation is latency. Individual audio-language models process audio in 1–5 seconds per example, depending on model size. In contrast, AudioToolAgent requires 30 seconds to one minute per example because the framework coordinates multiple models sequentially, with the agent waiting for each tool’s response before

proceeding. Four strategies mitigate this overhead. First, the orchestrator can invoke multiple tool calls simultaneously when the tools are independent, reducing the number of sequential round-trips. Second, concurrent batching across examples allows the orchestrator to process multiple audio questions in parallel: while waiting for tool responses on one example, the orchestrator can initiate tool calls for other examples. Third, deploying multiple replicas of tools across GPUs enables load balancing and reduces queuing delays. Fourth, using API endpoints from external providers eliminates the need to run local GPU infrastructure while enabling parallel API calls at low cost per query. Because no training is involved, the total compute cost remains modest compared to fine-tuning or training new models. Future research should focus on developing learned tool selection policies and lower-latency orchestration strategies to reduce sequential dependencies.

The experiments tested the framework with web search integration using both DuckDuckGo API and Tavily’s proprietary search API (see Figure 6.4). Ablation studies did not reveal consistent improvements, likely because the benchmarks focus on audio content requiring historical information and common knowledge facts. Web search integration may benefit real-world tasks where external knowledge retrieval is required beyond the audio content.

A next step is combining native multimodal processing with tool-calling capabilities, merging the efficiency of direct audio understanding with the extensibility of frameworks like AudioToolAgent.

Dataset	Models	Results
MMAU test-mini [393] Sound Music Speech Average	<i>Closed Source</i>	
	GPT-4o Audio [347]	64.56 56.29 66.67 62.50
	Gemini 2.5 Flash [77]	73.27 65.57 76.58 71.80
	Gemini 2.5 Pro [77]	75.08 68.26 71.47 71.60
	Gemini 3 Pro [124]	71.17 65.57 75.08 70.60
	Omni-R1 [†] [389]	81.70 73.40 76.00 77.00
	Step-Audio 2 [†] [519]	83.48 73.65 76.88 78.00
	AudioToolAgent	81.68 72.46 78.38 77.50
	<i>Open Source</i>	
	Phi-4 multimodal [303]	65.47 64.37 67.27 65.70
MMAR [280] Sound Music Speech Sound-Music Sound-Speech Music-Speech Sound-Music-Speech Average	DeSTA2.5-Audio [269]	70.27 56.29 71.47 66.00
	Audio Reasoner [541]	67.87 69.16 66.07 67.70
	Kimi-Audio [205]	75.68 66.77 62.16 68.20
	Qwen2.5-Omni [544]	78.10 65.90 70.60 71.50
	Step-Audio 2 mini [519]	76.28 71.56 71.47 73.20
	Audio Flamingo 3 [134]	79.58 66.77 66.37 73.30
	AudioToolAgent-Open	78.38 69.46 75.08 74.30
	<i>Closed Source</i>	
	GPT-4o Audio [347]	53.94 50.97 70.41 63.64 72.48 62.20 75.00 63.50
	Gemini 2.0 Flash [363]	61.21 50.97 72.11 81.82 72.48 65.85 70.83 65.60
MMAU-Pro [224] Sound Music Speech Sound-Music Speech-Music Speech-Sound Sound-Music-Speech Voice Multi-Audio Open-ended Instruction-Following Average	Gemini 3 Pro [124]	72.73 61.65 84.35 72.73 81.19 78.05 83.33 76.40
	Omni-R1 [†] [389]	67.30 51.50 64.30 45.50 70.20 64.60 70.80 63.40
	AudioToolAgent	73.33 63.59 82.99 63.64 83.03 81.71 79.17 77.00
	<i>Open Source</i>	
	Audio Reasoner [541]	43.64 33.50 32.99 45.45 42.66 31.71 25.00 36.80
	Qwen2.5-Omni [544]	58.79 40.78 59.86 54.55 61.93 67.07 58.33 56.70
	AudioToolAgent-Open	61.21 57.28 69.05 54.55 69.27 78.05 66.67 65.90
	<i>Closed Source</i>	
	GPT4o Audio [347]	44.70 63.10 68.20 40.40 43.50 62.50 57.10 21.40 57.50 32.60 43.20 82.50 52.50
	Gemini-2.5 Flash [77]	51.90 64.90 73.40 42.80 58.70 61.30 42.80 36.30 71.70 21.20 67.50 95.10 59.20
MMAU-Pro [224] Sound Music Speech Sound-Music Speech-Music Speech-Sound Sound-Music-Speech Spatial Voice Multi-Audio Open-ended Instruction-Following Average	Gemini 3 Pro [124]	50.76 69.68 83.39 38.00 73.91 73.86 71.43 32.62 59.66 24.88 60.96 78.16 60.72
	AudioToolAgent	54.96 72.57 87.09 52.00 69.57 76.14 57.14 46.15 81.03 30.00 36.80 40.23 61.90
	<i>Open Source</i>	
	Phi4 Multimodal [303]	25.70 47.80 47.60 30.00 39.10 30.10 28.60 39.70 42.70 11.40 42.50 65.40 38.70
	Audio-Reasoner [541]	34.20 50.10 44.00 26.00 36.90 43.20 28.60 20.30 43.40 22.60 38.60 43.40 39.50
	DeSTA2.5-Audio [269]	35.70 48.20 49.90 22.00 36.90 35.20 28.60 28.00 51.00 19.80 36.40 46.50 40.60
	Kimi-Audio [205]	46.00 57.60 52.20 46.00 54.30 48.90 42.80 43.70 50.60 17.20 34.50 42.30 46.60
	Audio Flamingo 3 [134]	55.90 61.70 58.80 40.00 41.30 47.70 57.10 26.80 58.60 26.00 44.20 33.30 51.70
	Qwen2.5-Omni [544]	47.60 61.50 57.40 40.00 53.20 60.20 28.50 41.20 60.00 24.30 52.30 61.30 52.20
	AudioToolAgent-Open	46.50 65.70 72.70 36.00 56.50 55.70 57.10 42.80 66.20 27.70 62.20 58.60 57.50

Table 6.1: Comparison of AudioToolAgent with baseline models on MMAU, MMAR, and MMAU-Pro benchmarks. AudioToolAgent achieves state-of-the-art or near state-of-the-art performance across benchmarks. [†]Self-proposed, no code or API available to verify.

6.5 Conclusion

This chapter introduced AudioToolAgent, a framework for multimodal audio understanding and reasoning where a central agent coordinates audio-language models as tools. The combination of Gemini 3 Pro, Qwen2.5 Omni, Audio Flamingo 3, and Whisper orchestrated by Gemini 3 Pro outperforms prior models on the MMAU, MMAR and MMAU-Pro benchmarks.

This framework combines the strengths of audio-language models and large language models. Ablation studies identified Gemini 3 Pro, Qwen2.5 Omni, and AudioFlamingo 3 as the highest-contributing audio tools across 20 evaluated tools. Among 12 LLMs tested as orchestrating agents, Claude Opus 4.5 and Gemini 3 Pro demonstrated the strongest performance, while open-source models like Qwen 3 235B achieved competitive results. This hybrid approach combines the audio processing of ALMs with the reasoning strengths of LLMs, creating a modular system in which tools and agents can be swapped independently.

The modular design enables integration of new tools and agents as they become available, without retraining. Future work includes developing learned tool selection policies, improving orchestration speed, and expanding the tool ecosystem to include audio retrieval and generation tools.

6.6 Acknowledgment

This work was supported by the Dutch Research Council (NWO 406.20.GO.030 to Prof. Elia Formisano), the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-14224. GW and JJ acknowledge funding by the Federal Ministry of Education and Research of Germany (BMBF) under grant no. 01IS24085C (OPENHAFM), under the grant 16HPC117K (MINERVA) and under the grant no. 01IS22094B (WestAI - AI Service Center West), as well as funding by EU Horizon under grant no. 101214398 (ELLIOT) and co-funding by EU from EuroHPC Joint Undertaking programm under grant no. 101182737 (MINERVA) and from Digital Europe Programme under grant no. 101195233 (openEuroLLM).

Discussion, Conclusion, and Future Work



This dissertation examines audio-language modeling through two themes: (i) data quality and (ii) human-inspired structure and reasoning. This chapter synthesizes these results into contributions, limitations, and directions for future work. The first section reviews the contributions of each chapter. The second section answers the five research questions posed in the introduction and connects them to the central thesis. The third section discusses limitations and challenges across five areas: data quality and quantity trade-offs, evaluation and benchmarking issues, computational constraints, legal and ethical considerations, and environmental concerns. The final section proposes directions for future research, including temporally grounded audio captioning, latent-space reasoning, end-to-end tool training, neuro-symbolic integration, prompt optimization, and multisensory world models.

7.1 Review of Contributions

Chapter 1, the introduction, establishes the field of audio-language modeling and defines the terminology used throughout the dissertation. It introduces audio-language models (ALMs) as systems that connect audio encoders with language models to describe sound using natural language, and identifies automated audio captioning (AAC) and audio question answering (AQA) as core tasks. The introduction presents the two recurring themes that structure subsequent chapters: data quality (how dataset design, curation, and overlap affect model performance) and human-inspired reasoning (how semantic sound descriptors and explicit reasoning steps can improve model output). It summarizes scaling laws and their implications for audio-language research, and outlines the five research questions addressed in Chapters 2 to 6.

Chapter 2 presents the first systematic cross-dataset overlap analysis of audio-language datasets, using CLAP embeddings to measure content duplication across the full landscape of 69 datasets (as of September 2024). This analysis reveals that YouTube-based and Freesound-based datasets dominate the field (23 and 21 datasets, respectively) and that overlap between datasets is extensive: 17.6% of AudioCaps [199] overlaps with WavCaps [296], and Clotho [102] shares up to 89% of its content with derivative datasets. These overlap findings expose data-quality problems that compromise model evaluation across the field: models trained and tested on overlapping sources report inflated performance. Additional issues include limited public availability (38 of the 69 datasets are not publicly accessible), severe underrepresentation of non-English data (only four datasets include non-English languages), bias in frequency, visual grounding and category diversity. These findings motivate the curation strategies developed in subsequent chapters and establish the data foundations used throughout the dissertation.

Chapter 3 addresses data quality through a balanced curriculum learning pipeline for Audio Question Answering. This chapter is the first to combine zero-overlap enforcement, statistical category balancing, and curriculum-stabilized GRPO [409] in a single pipeline for AQA. The pipeline tackles three challenges: training-evaluation

contamination, category imbalance (especially overrepresentation of speech and music), and instability in reinforcement learning. It enforces zero overlap between training and evaluation sets and explores two data preparation strategies: filtering overrepresented categories and introducing curriculum learning to stabilize GRPO. This results in a 3.3% improvement over the SFT+GRPO baseline.

Chapter 4 introduces AUDSEMTHINKER, the first audio-language model to structure reasoning around cognitive-science-inspired semantic descriptors, bridging auditory cognition research with language model reasoning. The model reasons explicitly about “who, what, how, where, when” prior to answering, and the human can inspect this reasoning trace before the answer is produced. This transparency makes visible how the model contextualizes sound through each semantic dimension. The chapter also introduces AUDSEM, a dataset built from YouTube closed captions using a multi-stage pipeline involving nine models. AUDSEM was filtered to minimize benchmark leakage (12 examples overlap with AudioSet, one also in AudioCaps, and 0 in VGGSound; see Chapter 2). AUDSEMTHINKER outperforms state-of-the-art models on multiple benchmarks (MMAU [393], AudioBench [475], and OmniBench [247]), achieving top-3 performance on music-related tasks.

Chapter 5 proposes ACES, the first caption-evaluation metric to apply semantic role labeling grounded in auditory cognitive neuroscience. This approach bridges how humans describe sounds with how metrics assess caption quality, and provides a more reliable evaluation signal than token-matching approaches. Unlike traditional metrics that rely on n-gram overlap, ACES uses fine-tuned BERT models to identify semantic descriptors in captions (who, what, how, when, sensation, etc.) and computes similarity scores using contextualized embeddings. The metric combines precision and recall weighted toward recall (1:9 ratio), following METEOR [27], incorporates fluency detection, and applies a small penalty for shorter captions. On the FENSE benchmark, ACES outperforms comparable metrics in the human-incorrect and human-machine conditions on Clotho-Eval.

Chapter 6 introduces AudioToolAgent, a multi-tool agentic framework that achieves state-of-the-art performance without any additional training data, relying on coordination and multi-hop reasoning alone. The framework orchestrates specialized audio-language models through a central text-only orchestrator (Gemini 3 Pro [124] or Qwen 3 235B [559]). The agent calls and selects audio understanding models iteratively (Gemini 3 Pro [124], Qwen3 Omni [559], Qwen2.5 Omni [544], Audio Flamingo 3 [134]) and speech recognition systems (Whisper [374]), resolves inconsistencies, and synthesizes final answers. On MMAU (77.50%) [393], MMAR (77.00%) [280], and MMAU-Pro (61.90%) [224], this approach sets new state-of-the-art results. The Shapley-based ablation across 490 configurations is the first systematic measurement of individual tool contributions in an audio understanding system, revealing that standalone tool performance does not predict contribution within the agentic framework: Gemini 3 Pro [124], Qwen2.5 Omni [544], and Audio Flamingo 3 [134] provide the highest

contributions as tools, while Claude Opus 4.5 [16] and Gemini 3 Pro perform best as orchestrating agents, with Qwen 3 235B [559] as the strongest open-source agent.

7.2 Answering the Research Questions

This section synthesizes the findings from Chapters 2 to 6 to answer the five research questions posed in Chapter 1. The answers collectively support the central thesis: that balanced datasets with low-overlap reduce contamination and improve training stability, and that semantic descriptors, semantics-aware metrics, and tool-based systems support interpretable and verifiable outputs and improve performance on audio-language tasks.

RQ1: What is the current landscape of audio-language datasets, and how do their origins, overlaps, and augmentation strategies shape the development and evaluation of audio-language models?

Chapter 2 establishes that the audio-language dataset landscape concentrates around two dominant sources: YouTube (23 datasets) and Freesound (21 datasets). This concentration creates systematic overlap: AudioCaps shares 17.6% of its content with WavCaps, and Clotho derivatives inherit up to 89% of the original dataset. The CLAP-based analysis reveals that a small set of categories (indoor environments, music, and male speech) account for nearly half of all samples, indicating severe distributional imbalance. Augmentation strategies that expand the dataset size often reinforce existing biases when LLMs rewrite captions for the same underlying audio. The analysis in Chapter 2 demonstrates that dataset quality, accessibility, and diversity matter more than scale alone. These findings directly support the thesis that balanced, low-overlap datasets improve training conditions by reducing contamination between training and evaluation sets.

RQ2: How can curriculum learning and data balancing improve audio question answering performance?

Chapter 3 demonstrates that careful data curation produces gains comparable to architectural improvements. The chapter addresses three interrelated problems: training-evaluation contamination from shared audio sources, category imbalance dominated by speech and music, and instability in reinforcement learning. The proposed pipeline enforces zero overlap between training and evaluation sets and applies statistical thresholds to limit overrepresented categories. Combined with curriculum-stabilized GRPO that presents easy examples before hard ones, the approach achieves a 3.3% improvement over the baseline. This result validates the first part of the central thesis: balanced datasets with low-overlap reduce contamination and improve training stability. The curriculum learning component further shows that training dynamics benefit from structured data presentation rather than random sampling.

RQ3: How can incorporating human-inspired reasoning through semantic descriptors enhance audio-language model performance?

Chapter 4 introduces AUDSEMT THINKER, a model that structures its reasoning around semantic descriptors derived from auditory cognitive neuroscience. The model explicitly analyzes who or what produces a sound, how the sound is generated, and where and when it occurs before producing an answer. This design mirrors how human listeners describe everyday sounds. The supporting dataset, AUDSEM, was constructed with minimal benchmark leakage (12 examples overlap with AudioSet, one with AudioCaps, zero with VGGSound). AUDSEMT THINKER achieves state-of-the-art performance on MMAU, AudioBench, and OmniBench, with particularly strong results on music-related tasks. These findings support the thesis that semantic descriptors enhance audio-language performance by providing interpretable intermediate representations that guide reasoning toward more accurate outputs.

RQ4: How can an audio captioning metric better align with human perception by using semantic descriptors?

Chapter 5 proposes ACES, a metric that evaluates captions based on semantic descriptors rather than n-gram overlap. ACES uses fine-tuned RoBERTa models to identify entities corresponding to who, what, how, where, when, and perceptual attributes, then computes similarity using contextualized embeddings. The metric combines precision and recall with a 1:9 ratio favoring recall (following METEOR), incorporates fluency detection, and applies length penalties. On the FENSE benchmark, ACES outperforms comparable metrics in the human-incorrect and human-machine conditions of Clotho-Eval. These results demonstrate that semantics-aware evaluation aligns more closely with human judgment than token-matching approaches. This finding supports the thesis by showing that semantic descriptors improve the metrics used to evaluate them, which enables a more accurate assessment of audio captioning quality.

RQ5: How can agentic frameworks that coordinate specialized audio tools achieve state-of-the-art performance without additional training data?

Chapter 6 presents AudioToolAgent, a framework that orchestrates specialized audio-language models and speech recognizers through a central text-only agent. The agent determines which tools to call, in what order, and with what prompts, issuing follow-up queries when responses are incomplete or inconsistent. On MMAU (77.50%), MMAR (77.00%), and MMAU-Pro (61.90%), AudioToolAgent achieves state-of-the-art results without requiring additional training data. Ablation studies show that general-purpose audio models contribute more to the agentic system than narrow speech recognizers, and that tool performance as standalone models does not predict their contribution within the framework. These results support the thesis by demonstrating that tool-based coordination and multi-hop reasoning can boost performance beyond what any individual model achieves alone.

Synthesis: Supporting the Central Thesis

The central thesis of this dissertation is that balanced datasets with low-overlap reduce contamination and improve training stability; semantic descriptors, semantics-aware metrics, and tool-based systems support interpretable and verifiable outputs and improve performance on audio-language tasks. The five chapters investigate specific aspects of this thesis:

- Dataset quality and overlap (Chapter 2).
- Balanced data curation and curriculum learning (Chapter 3).
- Semantic descriptor reasoning for audio understanding (Chapter 4).
- Semantics-aware caption evaluation (Chapter 5).
- Tool-based coordination without additional training (Chapter 6).

The five chapters collectively validate both components of the central thesis. Chapters 2 and 3 address the data quality theme: the survey reveals the overlap and imbalance problems, while the curriculum learning pipeline demonstrates concrete improvements from addressing them. Chapters 4 to 6 address the human-inspired reasoning theme: semantic descriptors improve model reasoning (AudSemThinker), enable better evaluation (ACES), and boost performance through multi-hop tool coordination (AudioToolAgent). In terms of data quality, AudSem required overlap filtering to enable semantic descriptor training. The interpretability aspect emerges from the semantic descriptor framework in AudSemThinker, where the explicit reasoning about who, what, how, where, and when allows humans to inspect the model’s understanding of the audio before it produces an answer. AudioToolAgent contributes differently: its central agent cross-checks tool outputs and resolves inconsistencies through iterative queries, producing performance gains through coordination rather than interpretable intermediate representations. These contributions demonstrate that data curation and structured reasoning frameworks advance audio-language modeling beyond what architectural scaling alone achieves.

7.3 Discussion, Limitations and Challenges

This section distills the challenges that arose across the dissertation’s chapters. Each subsection groups related limitations and points to directions for improvement.

7.3.1 Data-Related Challenges

A central tension throughout this dissertation is the trade-off between data quality and quantity. Although Chapter 2 shows that millions of audio-text pairs are available, many datasets suffer from visual bias [199, 435, 331, 435, 573, 154], overlapping content (e.g. [199] and [57]), and imbalanced data distributions [126, 57]. Chapter 3

demonstrates empirically that data curation techniques to circumvent imbalanced data distributions can produce stronger models (see, e.g., [296, 23]).

Scaling laws indicate that the size of the dataset should grow with model parameters [194], yet creating high-quality datasets (e.g., through pipelines like AudSem; see Chapter 4) is computationally expensive. A solution is data augmentation techniques, such as the ones depicted in Chapter 2, Section 2.3. Techniques such as SpecAugment [356], MixUp [577], and PairMix [203] show that the augmentation of audio data can improve performance without the need for additional data [225, 203].

Using LLMs for data augmentation, as described in Chapter 2, is now commonplace and is a practical way to improve dataset quality [604, 554, 23]. Often, this is done with the guidance of other modalities [435, 573]. LLMs are used to generate captions for audio-language datasets (e.g. [296, 203, 435, 129], see also Chapter 4) and multiple audio clips to create natural combinations [523, 377]. With the rise of chat models, datasets that include instruction tuning prompts are also becoming more common, and mostly include LLM-generated captions (e.g. [134, 544, 269, 205]).

With the rise of LLM-generated datasets, issues such as bias amplification [320, 391] or model collapse are becoming more prevalent [320, 97, 418] (see Chapter 2, Section 2.5). Consequently, future opportunities include the development of automated dataset quality assessment tools that measure acoustic diversity, caption quality, and (temporal) alignment between audio and text. Techniques such as data overlap matching with CLAP described in Chapter 2 can be used for this. In other works, contrastive techniques such as CLAP [525] are also used to assess audio-text alignment to filter out mismatched examples [23, 61].

Future research should develop metrics to quantify the dataset “synthetic-ness” to combat model collapse and establish thresholds for safe synthetic data ratios. For example, research on adversarial techniques that detect synthetic patterns could help identify when models overfit on synthetic data [317, 538]. Human annotations are often better quality than synthetic data generated by LLMs [311, 418, 326], but unfortunately humans do not scale or are very costly to do so [323]. However, it might be important to keep hybrid approaches with human and synthetic data combined. Humans could provide sparse high-quality annotations, which synthetic methods could then augment. This would achieve some level of human grounding while still providing scale, as demonstrated by works such as [130, 554, 73].

Further opportunities for optimizing this quality-quantity trade-off reside in the adoption of adaptive sampling strategies that dynamically adjust data selection and filtering during training based on model performance [504] or more efficient curriculum learning [580, 422] (see also Chapter 3). Temporal-aware pretraining, such as T-CLAP [572] and composition-aware variants of CLAP [132] expose how richer supervision can increase robustness without collecting entirely new corpora.

Standardized protocols are needed to assess and report dataset contamination [233, 392]. Future research should develop contamination detection frameworks similar to the ones in Chapter 2 that operate at multiple levels: (i) exact matching using identifiers or metadata and (ii) similarity based on metrics (see also [542]). These frameworks must be integrated into dataset creation pipelines and benchmark evaluation protocols to ensure true zero-shot learning conditions. Chapter 2 provides online guidelines in its GitHub repository¹ to help avoid dataset contamination for various datasets.

The overlap becomes more subtle when the audio clips differ but share underlying characteristics. For example, audio models trained on songs by specific artists often perform better on benchmarks containing other songs by the same artists, even when the exact tracks differ [429, 113]. This artist-level overlap can inflate performance metrics without explicit test-set contamination, which can be seen as a gray area in benchmark integrity. Similarly, in the domain of sound effects, models trained on specific sound effects often perform better on benchmarks containing other sound effects from the same source, even when the exact sound effects differ [114, 328].

7.3.2 Evaluation and Benchmarking Challenges

Chapter 5 highlights the limitations of the existing evaluation metrics. N-gram metrics poorly capture meaning; semantic metrics like SPICE [13] fail when phrasing varies. ACES [512] (Chapter 4) improves alignment with human judgments but still faces edge cases: zero scores when no semantic categories overlap and limitations when multiple entities share the same category. Similar concerns exist in music-focused evaluations such as MuChoMusic [496], which emphasize culturally grounded questions and human validation to reduce overfitting to metrics.

An issue with metrics including references is that one compares the generated caption or output to the reference. However, reference captions themselves may be incorrect, especially if they are LLM-generated or originate from noisy data without manual validation. Even with human labeling, the same audio can be described in multiple ways: Clotho [102], for example, provides five reference captions per audio clip, each fairly different from the others. In practice, models trained only on Clotho perform poorly on AudioCaps [199] and vice versa [228]. Another issue is that captions often follow specific style constraints: Clotho and AudioCaps contain captions with approximately 5-15 words, while datasets like Clotho-Detail [587] or AudioLog [24] contain paragraph-length descriptions. Metrics should account for this, something that ACES (see Chapter 5) tries to tackle.

Reference-free metrics avoid these issues by using models pretrained on many examples to assess caption quality directly, without relying on potentially flawed or

¹<https://github.com/GLJS/audio-datasets>

stylistically constrained references. These methods include metrics based on pre-trained critics, such as PAM [86], CLAPScore [528, 282], or benchmarks such as BRACE [149], using models to evaluate the quality of the captions.

Future work should explore reference-free caption evaluation while promoting semantic richness and variety in captions. Learnable evaluation metrics should also be adapted to specific domains and contexts, for example, biodiversity or the medical domain. Good quality metrics should be used as a policy for reinforcement learning algorithms (i.e. GRPO, see Chapter 4 and Chapter 3). Metrics used as policy gradients require fast performance: speed is key to be able to train a network reasonably fast with GRPO as rewards need to be feed back to update the models, although alternatives such as off-policy training exist and may even be beneficial [341, 314, 558].

A major concern is self-evaluation bias: when researchers assess their models on benchmarks they design, subtle biases can inflate performance. Researchers often create benchmarks to accompany their models, using these benchmarks to evaluate and compare their models against others in the field. However, these benchmarks frequently contain examples from test sets that, while held out during training, originate from the same data collection process as the training set (i.e. [131, 128, 519, 223]). Models trained on data that share the same source, distribution and collection methodology as the test set often perform better than related models that were trained on entirely different datasets [482]. This creates an unfair comparison as the performance advantage may stem from familiarity with the data distribution.

Another issue is that in public benchmarks (i.e., MMAU [393], MMAR [280], or MMAU-Pro [224]) the benchmark and its reference texts are publicly distributed. This publication can create intentional or unintentional overfitting by researchers. While directly training on the test set is clear cheating, more subtle forms of data leakage exist in a gray area. For instance, researchers could compare embeddings of reference data to candidate training examples (i.e. through embedding based methods such as in [154, 234]), and selecting samples that are similar to the test distribution. This practice, while not explicit test-set training, still biases models toward benchmark-specific patterns. To mitigate these risks, benchmarks should maintain both a public validation set and a truly held-out test set accessible only through secure evaluation servers (good examples include Kaggle² or the DCASE challenge³), to ensure that models cannot be optimized on data they will be evaluated on.

In practice, it is often not clear how the benchmark was used in terms of prompt or model output usage. It is, for example, possible to use exact matching of the model response to the answer options (as used in Chapter 3 and Chapter 6), as well as to compute a sentence embedding of all the multiple choice options and the generated output and take the options that have the highest similarity score (Chapter 4). Theoretically, the latter option will result in better scores as more possibilities can be matched. With

²<https://www.kaggle.com/>

³<https://dcase.community/challenge2025/task-audio-question-answering>

exact matching, the answer is wrong if one letter is misspelled. Future work should standardize the evaluation protocols for the benchmarks to ensure reproducibility and fairness.

During the research for this dissertation, this issue of self-evaluation bias was repeatedly encountered. When trying to replicate existing state-of-the-art models on multiple benchmarks, almost always a lower performance than reported was seen in the original articles (e.g., in Chapter 6). This reproducibility gap undermines the field's credibility and makes meaningful model comparison impossible. The problem exists in two ways. First, during inference when evaluating models on benchmarks, but also during training with unclear or incomplete descriptions, training procedures, and data preprocessing steps prevent replication of the original models. Moreover, datasets in the audio-language field are often proprietary and not publicly available, making it difficult to replicate results on the training side.

Future work on new benchmarks should require hold-out test sets accessible only through secure evaluation servers. More importantly, conferences and journals should prioritize papers with benchmarks where models are independently evaluated. Future work could also focus on testing models through human preference studies, where groups of users rate different models based on actual usage. More research on biases and replication issues should be done in this field.

7.3.3 Computational and Architectural Challenges

Chapters 4 and 6 show that longer reasoning and the use of more tools result in better performance, but also increase computational costs. The trade-off between performance and efficiency presents a challenge. The methods included in this dissertation achieve state-of-the-art performance, but their computational requirements limit their practical adoption. For the generation of AudSem and the training of AudSem-Thinker (Chapter 4), approximately 15k GPU hours on A100 and H100 GPUs were needed. To host AudioToolAgent-Open (Chapter 6) locally, 10 H100 GPUs are needed, particularly to run the Qwen 3 235B agent with its 235B parameters.

In general, the computational demands of audio-language models have increased drastically over the years. At the start of this PhD in 2021, audio-language models such as the Audio Spectrogram Transformer [137] and Audio Captioning Transformer [291] contained approximately 100M parameters. Today, audio models such as Qwen3 Omni Captioner [559] have reached 30B parameters. State-of-the-art audio-language models may soon no longer fit on a single GPU and require clusters of GPUs for training and fine-tuning. This computational barrier creates a bottleneck for new PhD students, who may lack access to large compute clusters. However, scaling laws demonstrate that model size remains a strong predictor of language model capabilities [194, 173], suggesting that models will continue to grow larger in the coming years.

However, larger models are not the only path forward. There are many opportunities for researchers to focus on smaller and more niche models that serve a specific purpose. Recent examples include Mellow [87] (167M parameters) and Pengi [88] (323M parameters). Models like this are particularly useful for on-device tasks, for example, for offline recognition of sounds in environmental sound classification [310, 469], for urban noise anomalies [156] but also for biodiversity [184]. For example, DCASE has the challenge in each year 2022-2025 to create low-complexity models for purposes like this⁴. These challenges impose strict computational constraints: in 2025, models must fit within 128 kB of memory and perform inference within 30 MMACs (million multiply-accumulate operations) per second of audio. A variant of DCASE, named BioDCASE⁵ also hosted a “Bioacoustics for Tiny Hardware” challenge in 2025 to detect birdsong on a microcontroller.

Future research should investigate smaller and faster models that still preserve reasoning and tool-use capabilities. As model sizes increase, it is also important to introduce benchmarks in audio-language learning for smaller and faster models, and even models trained on fewer data. To benchmark datasets, the amount of data needed to train a model to a specific accuracy should be plotted (similar to the scaling laws).

In this category also falls a phenomenon known as ‘speedrunning’, to train a model to a specific performance target as fast as possible. A popular speedrun⁶ is done with NanoGPT⁷, where researchers find the fastest algorithm to train a GPT2 language model to 3.28 cross-entropy loss on the FineWeb validation set using 8 NVIDIA H100 GPUs. Since its inception, speedrun has achieved 24.725x faster training than the baseline model (which took days to weeks on 256 TPU v3 cores [375]), whereas normally with a Python framework one would train a similar model in 0.86 days on 512 GPUs [415].

Solutions to the computation problem of larger models include developing adaptive computation strategies that selectively invoke expensive tools only when necessary, reducing costs while maintaining performance (see Chapter 6, Section 6.4.2).

Another recently proposed approach is using mixture-of-experts layers in the attention mechanism to route requests to particular expert layers, to reduce parameter usage in inference [315, 599]. Qwen3 Omni Captioner [559] uses this approach to reduce the number of parameters in the model, but is the first to do so for audio-language models. In the future, more research should be done to explore the use of mixture-of-experts layers in audio-language models and what kind of sparsity would be beneficial [324].

⁴<https://dcase.community/challenge2025/task-low-complexity-acoustic-scene-classification-with-device-information>

⁵<https://biодcase.github.io/>

⁶<https://github.com/KellerJordan/modded-nanogpt>

⁷<https://github.com/karpathy/nanoGPT>

Constrained decoding offers a method to guide model outputs by restricting the vocabulary at each generation step (see also Chapter 3). Libraries such as Outlines [514] and XGrammar [99] implement efficient algorithms to enforce context-free grammars or regular expressions during generation. In principle, this approach could benefit audio question answering by ensuring that a model always produces at least one valid answer, for example, by constraining the first token to one of the answer letters (A, B, C, or D). However, experiments in Chapter 3 found that guided decoding reduced performance by 11.2% compared to unconstrained generation. The findings show that strong format constraints can possibly hinder reasoning ability [443], forcing one to respond early with an answer letter prevents the model from using reasoning tokens that would otherwise improve accuracy. This trade-off between guaranteed output structure and reasoning flexibility remains an open challenge for audio-language systems that require both reliable parsing and accurate responses.

7.3.4 Legal, Ethical, and Societal Challenges

Chapter 2 identifies the dominance of English language content, with only 4 out of 69 datasets including non-English-languages: AudioCaption [522, 549] and FAVDBench [412] with original Mandarin captions, CAPTDURE [340] with Japanese captions, and mClothoAQA [34] with 8 languages. Most multilingual datasets are automated translations from English, resulting in inheritance of cultural and linguistic biases and translation errors. This monolingual bias limits the applicability of the model in contexts where sounds carry different cultural significance.

The challenge extends beyond language to the cultural interpretation of sounds. What constitutes a “loud” sound varies between cultures, as does the significance of specific acoustic events. The semantic descriptor framework in Chapter 4 assumes Western interpretations of who, what, where, when, and how, which may not align with other cultural frameworks for understanding sound [230, 94]. Future work should prioritize building multilingual datasets from native sources instead of translations. Research into culture-aware semantic frameworks that adapt captions or descriptions of audio based on linguistic and cultural context could improve the deployment of cross-cultural models.

A critical challenge facing audio-language research is the legal status of datasets. Chapter 2 shows that 23 of the 69 datasets derive from YouTube, despite YouTube’s Terms of Service explicitly prohibiting content download without authorization. Researchers circumvent these restrictions by sharing URLs instead of audio files, requiring users to scrape content themselves using tools like *yt-dlp*. The legality of scraping remains unclear. Many datasets based on YouTube exist (see Chapter 2, e.g. AudioSet [126, 199, 57]) and for almost all of them it is necessary to download the audio from YouTube.

Beyond dataset scraping, the use of copyrighted content for model training raises is a big issue. More than 50 copyright infringement cases have been filed in the United

States against technology companies for training models without permission [281]. Tech companies use the argument of fair use in court, which permits reproduction under specific circumstances. Recent rulings in *Bartz v. Anthropic* [10] and *Kadrey v. Meta* [66] found large language model training to be “transformative” because models analyze patterns to produce statistical weights rather than retain or reproduce expressive content [136, 29]. The distinction between transformative and substitutive use will determine in the future whether copyright law can address AI training at scale.

Another concern is the legal status of model usage. The EU AI Act imposes explicit prohibitions and high-risk classifications that directly affect audio-language research [109]. Article 5(1)(f) [109] prohibits AI systems that infer emotions in the workplace and educational settings, except for medical or safety purposes. The models discussed in this dissertation, including AudSemThinker (Chapter 4) and AudioToolAgent (Chapter 6) could analyze vocal prosody, tone, or stress patterns. Voice constitutes biometric data under Article 9 [108] of the GDPR and Annex III of the EU AI Act [109], which classifies remote biometric identification systems as high-risk (Article 6(2)) and prohibits real-time biometric identification in publicly accessible spaces under Article 5(1)(h). Speaker diarization, as used in section 6, therefore, requires explicit consent and risk assessment when deployed outside controlled research settings.

The semantic descriptor framework in Chapter 4, which reasons about “who” and “how” a sound is made, could inadvertently profile individuals based on protected characteristics if deployed on human speech without safeguards. In principle, the AudSemThinker model trained in Chapter 4 can be used for surveillance purposes, such as recognizing specific events in public areas. Without anonymization steps, this can be illegal.

It is also important to consider that audio datasets often capture bystanders who never consented to recording; GDPR [108] requires a lawful basis for processing such data, and the “legitimate interest” justification (Article 6(1)(f), GDPR) requires a balancing test that research convenience alone may not satisfy.

To address some legal concerns, all code and dataset releases associated with the chapters in this dissertation have been distributed under non-commercial share-alike Creative Commons licenses (CC BY-NC-SA 4.0)⁸. This license option ensures that the work remains accessible for academic and research purposes. This license also prevents commercial exploitation that could increase legal and ethical concerns. It also potentially mitigates the harms of deployment in surveillance or profiling contexts.

Future research should establish ethical frameworks that balance research needs with privacy rights. Specifically, privacy-preserving techniques for audio, such as voice anonymization, could enable dataset creation without compromising individual

⁸<https://creativecommons.org/licenses/by-nc-sa/4.0/>

privacy. Negotiating explicit data-sharing agreements with platforms or focusing on open-licensed content would provide more sustainable legal foundations for the field.

7.3.5 Environmental and Sustainability Concerns

The scaling laws emphasized throughout this dissertation also raise significant environmental and sustainability concerns. Training AudSemThinker required approximately 15k GPU hours on A100 and H100 GPUs (Chapter 4), and AudioToolAgent-Open requires 10 H100 GPUs for local deployment (Chapter 6). As model sizes grow from hundreds of millions to hundreds of billions of parameters, the carbon footprint of audio-language research increases proportionally [428, 360].

Future research should report energy consumption and carbon emissions alongside accuracy metrics, following emerging best practices in responsible AI development [168, 272]. Whenever training of a new model takes place, one should argue or reason about what the goal or purpose of such training run is. If the result would not necessarily be better than previous models, then one should question the feasibility of the project or at least take the environmental losses into account. Otherwise, models without any training, such as AudioToolAgent depicted in Chapter 6, are a viable alternative, as they pose no training costs at all.

Other strategies to reduce environmental impact include reusing pretrained models and training small layers on top of them, as well as employing parameter-efficient methods such as LoRA [179] (see Chapter 4 and Chapter 3) or prefix tuning [204]. These approaches minimize computational overhead by freezing the majority of model parameters and updating only a small subset, reducing training time and energy consumption. For example, LoRA adapters typically require less than 1% of the parameters of full fine-tuning while achieving comparable performance.

In the audio-language modeling domain specifically, a particularly effective strategy is to reuse pretrained language model decoders and freeze them entirely, training only a new audio encoder on top (e.g. [260, 444]). This approach uses the fact that encoders are typically much smaller than decoders: while state-of-the-art language model decoders can contain multiple billions of parameters, audio encoders often only contain hundreds of millions of parameters [72, 131].

7.4 Future Research Directions

This final section proposes new directions for future audio-understanding research and situates them within existing work. Each subsection represents a research area that, given more time during this PhD, would have been explored as a standalone research paper. These themes emerged from the work presented in earlier chapters,

but remained beyond the scope of this dissertation. They are presented here as potential starting points for new students entering the field.

7.4.1 Temporally Grounded Sound Events with Captions

The chapters in this dissertation focussed partly on audio captioning [101], with one caption summarizing an entire clip from start to end. This task compresses multiple acoustic events into a single sentence and ignores when precisely in the audio each event occurs. The result has low temporal granularity [71], as audio descriptions consist of multiple events together.

The next step is an event-centric formulation: each clip contains multiple events, and each event has a start time, an end time, and a free-form description [319, 318]. In essence, this approach extends beyond sound-event detection [300], which often allows only a fixed set of existing labels [526]. The name for this task could be “temporal audio captioning”.

To develop well-performing models in this field, multiple components are needed: datasets, models, and evaluation methods. In the following, each component is described in detail.

Datasets for temporal audio captioning

Datasets are needed that contain time-stamped, free-form captions at multiple granularities. Current efforts have introduced simulated datasets like (1) Clotho-Moment [319], with 15-30 second sound events from Clotho [102] overlaid on background soundscapes of Walking Tours [472]. (2) AudioTime [540], which provides annotations with ordering, duration, frequency, and timestamps. (3) ASFX-SED, sound effect mixtures of background sounds and events. These datasets use synthetic generation: scripts combine individual sounds at specified times, producing automatic timestamps. This approach guarantees accurate temporal labels because the placement of each sound is known.

However, synthetic mixing introduces the risk of implausible combinations. Mechanically overlaying unrelated sounds can produce audio that does not occur naturally, which results in training models on unrealistic acoustic scenes. One solution is to use a language model to select audios that coexist in the same environment before combining them. The language model identifies sounds that plausibly occur together (such as traffic noise with pedestrian chatter or birdsong with wind) and overlays these sounds (e.g. [531]). This method reduces implausible combinations while maintaining the scalability of synthetic generation.

Human annotation of events and timestamps in naturally occurring audio, such as in TACOS [368], produces higher-quality data but does not scale for large pretraining

datasets. During pretraining, noisy data can be used, and mechanically overlaying data may help scale a large amount of examples, but human-annotated datasets like TACOS should be prioritized for evaluation and fine-tuning.

Models for temporal audio captioning

These models should be capable of jointly learning start and end times and the generation of captions. Recent frameworks such as DASM [50] and FLAM [526] have explored dual-stream and frame-wise architectures, respectively, to perform event recognition together with temporal localization. Both models formulate sound event detection as a retrieval task: the model determines when a specific query at inference time is active in an audio signal. These models are limited by the text input provided from a database: they compute similarity scores only for the text embeddings that are measured, meaning that the possible output descriptions are restricted to whatever the user supplies to the model.

An alternative approach is that of encoder-decoder models such as Temp4Cap [71] and TimeAudio [480], using LLMs to generate structured “temporal captions” to train models directly on sequential relationships. Encoder-decoder models can also generate all possible captions for a clip, unlike retrieval-based models.

This encoder-decoder approach is to use pretrained models to preserve general audio understanding and fine-tune for specific event detection by adding special tokens to the tokenizer. The model decoder outputs multiple captions in a paragraph-like format, with each event marked by temporal tokens. For example, the output might be: 0.5s 2.3s dog barking 3.1s 5.8s car passing by”. This format allows the model to generate structured temporal information without requiring separate regression heads. The temporal tokens are learned during fine-tuning, while the pretrained weights retain general audio understanding capabilities.

A promising extension of the encoder-decoder approach is the introduction of controllability through input-level tokens that steer the behavior of the model [228, 451, 493]. These control tokens (also known as style tokens) would specify the desired output format and granularity before the model processes the audio.

For example, temporal precision tokens could indicate whether timestamps should be reported in seconds, or even more detailed with floating point precision. Granularity tokens could instruct the model to produce fine-grained event-level descriptions or higher-level summaries that group related sounds. Fine-grained descriptions would separate individual occurrences (e.g., each horse neigh as a distinct event), while coarse descriptions would group repeated occurrences (e.g., multiple horse neighs within a short interval as a single event). Event-type tokens direct the model to focus on specific acoustic categories such as music, speech, or environmental noise, enabling selective attention to relevant sound classes.

This approach is similar to Magenta RealTime [451], which allows users to control specific music parameters such as key, tempo, and brightness through UI controls that are converted to specific tokens fed to the model.

An additional extension is to integrate source separation into the pipeline, a field that has been explored extensively [550, 551, 495, 547, 263, 278] Blind source separation (BSS) tackles the ‘cocktail party problem’ to isolate individual source signals from a mixture of multiple sounds without knowing how they were mixed (without labels) [51, 70]).

In language-queried audio source separation (LASS) (see also Section 2.2.2), the model extracts isolated audio streams for target sounds given a text query [264, 98, 574, 477]. Most models isolate the target audio stream, while some also remove the target sound from the source to preserve the remaining audio without the specified sound (e.g. PromptSep [505], CLAPSep [98], and HSM-TSS [571]).

Source separation cannot simply mask portions of the input audio, as this would create temporal gaps and remove sounds that overlap in time. Due to this overlap, the model must reconstruct the acoustic scene by targeting specific sounds within overlapping regions. For the reconstruction of audio streams, both in the target sound separation and in the source without the target sound separation, models use generative frameworks that synthesize the target audio. These frameworks include techniques such as diffusion [505, 495] and flow matching [574] to generate audio streams that are similar to the target audio stream.

Evaluation methods for temporal audio captioning

Evaluation methods must assess both temporal accuracy and caption quality. Temporal metrics should reward predictions that closely match ground-truth intervals with some tolerance. For example, a predicted event spanning 2.0 to 3.5 seconds should receive partial credit if the true interval is 2.1 to 3.4 seconds, with, for example, scores decreasing smoothly as overlap diminishes. Caption quality metrics also present a challenge. Traditional reference-based metrics such as BLEU [355], METEOR [27], and CIDEr [471] compare the generated captions with human references. However, these metrics penalize valid descriptions and alternative descriptions that differ lexically from the reference but remain semantically correct (see 7.3.2).

Evaluation protocols should also account for multiple levels of granularity. A model that produces fine-grained event-level descriptions (e.g., separating each dog bark) should be evaluated differently from one that generates coarse summaries (e.g., grouping multiple barks as a single event). The metrics must distinguish between these levels of granularity and assess whether the chosen level matches the prompt tokens for all the prompt tokens available.

Earlier sound separation evaluation methods included signal-level metrics to measure signal-to-distortion, interference, and artifact ratios (SDR, SIR, SAR, respectively) to compare an estimated source with ground truth sources [474]. However, the original SDR implementation allowed artificial score inflation through signal distortion [390], and it is still possible to achieve high SDR scores while perceptually failing to achieve true semantic separation [583]. (Scale-Invariant SDR (SI-SDR) [390] was suggested as an alternative. Other techniques include Log Spectral Distance (LSD) and Peak Signal-to-Noise Ratio (PSNR) to evaluate separation quality by comparing the spectrograms of predicted and target samples [574]. However, these metrics are not suitable for generative separation models as they do not take the text query into account [528].

A solution for this is the use of reference-free metrics, such as Fréchet Audio Distance (FAD), which is used for generative models to assess the overall distribution quality of the generated audio relative to a set of real recordings [384]. A bonus of reference-free metrics is that they do not require a reference signal, which is often difficult to obtain in real-world scenarios. The correlation between FAD and human perception is strongly dependent on the choice of embedding used to compute the distance [439]. However, matching distributions such as in FAD do not guarantee a perceptually closer result [12] and were found to capture the perceptual quality of synthesized sounds insufficiently [473]. A better alternative would be to use CLAPScore [528], which measures the semantic consistency between the separated audio and the text query by calculating their similarity in a shared embedding space [525].

In terms of temporal accuracy, sound event detection metrics (SED) can be used to evaluate the temporal accuracy of the generated captions. Most simple metrics in SED fall into segment-based or event-based metrics categories [297]. Segment-based evaluation divides audio into fixed-length segments (e.g., 1 second) and classifies each segment as active or inactive for each sound class, computing true positives, false positives, and false negatives per segment. This approach forgives this issue with timing errors (off by hundreds of milliseconds) within segments but does not explicitly score onset and offset accuracy. Event-based evaluation, on the contrary, matches predicted events with ground-truth events instance by instance. This requires temporal agreement within a tolerance limit (e.g., ± 250 ms or something with the duration of the event taken into account). A predicted event counts as a true positive only if it matches a reference event of the same label within these tolerances, making this approach more sensitive to precise temporal boundaries. Since SED is a classification task, metrics such as F-score and error rate are used for a limited number of classes, not for longer free-form descriptions.

For temporal audio captioning, the evaluation should combine temporal and semantic metrics. The temporal component can use simple metrics similar to the F-score from sound event detection to assess the onset and offset accuracy. For caption quality, embedding-based techniques, such as BERTScore [579], MACE [95], or ACES (see Chapter 5), are preferable to traditional reference-based metrics, as they better capture semantic similarity and reward valid paraphrases. When evaluating language-queried

audio source separation (LASS) in addition to temporal captioning, the evaluation protocol requires both separation quality and semantic alignment metrics. Simple metrics similar to F-score or error rate can assess whether correct sound sources were isolated. For semantic consistency between separated audio and text query, CLAPScore [528] provides a reference-free measure of alignment. FAD (Fréchet Audio Distance) [384] can serve as a backup metric to evaluate the overall quality of the separated audio distribution.

7.4.2 Reasoning in Latent Space

Current models reason through token sequences [499]. They output text during a “thinking-phase” before producing an answer [148]. However, this type of reasoning model must express every intermediate thought as a discrete word or subword from a fixed vocabulary. GPT-2, for example, uses a vocabulary of 50,257 tokens [375]. Every reasoning step must map onto one of these tokens, regardless of how nuanced or continuous the underlying concept is. This loss of information limits the model’s ability to reason even better.

This constraint is also found in linguistics: the Sapir-Whorf hypothesis (also known as linguistic relativism) states that the structure and vocabulary of a language shape the thought patterns of its speakers [402, 507]. An individual who wants to communicate a concept must select words from an available lexicon. If the lexicon lacks a precise term, the speaker must approximate, potentially losing information. Languages with richer vocabularies for specific domains enable finer distinctions. For example, studies of color perception show that speakers of languages with more color terms discriminate hues more rapidly than speakers of languages with fewer terms [515, 195].

Similarly, a model constrained to a finite token vocabulary may lose precision when compressing continuous internal states into discrete outputs. Latent representations offer a higher-bandwidth alternative. In the same GPT-2 XL model, each hidden state consists of a 1,600-dimensional vector of floating-point values [375]. This representation is continuous and much more expressive than a single token index. The hidden state preserves the full distribution of activations, instead of discarding all information except the winning token. Reasoning in this latent space avoids the information bottleneck imposed by tokenization.

Recent work explores this direction. COCONUT (Chain of Continuous Thought) [159] trains language models to use the last hidden state as a “continuous thought” that feeds back into the model as the next input embedding. This approach bypasses token generation entirely during intermediate reasoning steps, resulting in multiple reasoning paths. This helps in reasoning tasks, particularly problems that require backtracking. Another method is Continuous Chain-of-Thought via Self-Distillation (CODI). CODI extends this the continuous thought paradigm through self-distillation,

which compresses explicit chain-of-thought traces into continuous representations. CODI and COCONUT demonstrate that reasoning does not have to occur in language, continuous latent updates can be a substitute for token-level thinking.

A promising research path is to allow the model to run a fixed or adaptive number of latent steps, with a controller that stops the computation based on an uncertainty or consistency criterion [144, 28, 597]. Other objectives may be needed during training to learn the model to halt, such as latent-to-audio/text alignment to align spaces [525]. A potential hybrid variant could use latent steps for exploration and emit short textual plans only when needed. This approach could pair well with the auditory semantic descriptors in Chapter 4.

The evaluation of this idea should compare token-only, latent-only, and hybrid reasoning under a certain amount of computation. Evaluation should also report accuracy and speed with multiple prompt perturbations. Given the limited work on latent reasoning in audio, this direction could test whether high-dimensional internal computation might in principle result in better audio understanding than text-level reasoning.

7.4.3 End-to-End Training for Tool Use

A possible next direction for an agentic audio system as AudioToolAgent (Chapter 6) would be to have multiple models in an end-to-end training fashion (also known as multi-hop reasoning). In this setting, a single training loop optimizes the orchestrator’s tool-selection policy and its prompt to the tools, as well as to the tools themselves, using reward tasks from answering audio questions and related objectives. The goal is to learn when to call which tool and how to structure the request and responses for all parts of the pipeline.

This end-to-end training approach is particularly useful for enriching the content available to the orchestrator, which directly supports test-time scaling. Test-time scaling, similar to the scaling laws discussed throughout this dissertation, operates on the principle that the more information a model has during inference, the better. This information could also be an additional context in the form of tool outputs. The AudSemThinker model (Chapter 4) demonstrates this phenomenon: longer reasoning traces with semantic descriptor reasoning and extended thinking resulted in improved performance up to a certain point. Latent reasoning approaches (Section 7.4.2) could further enhance this by allowing richer intermediate representations. By training the orchestrator and tools jointly, the system learns to generate more informative tool responses. This increases the quality and quantity of context available at test time without requiring architectural changes or larger models.

Several works have explored how multiple LLMs or tools should be orchestrated when training them. Various reinforcement learning techniques exist [189, 606, 524, 557] that include multi-module policy-gradient updates for its tools for multi-turn

reasoning. A next step would be similar to mmGRPO [606], which enables online weight optimization for multiple LM programs by propagating final rewards backward across the modules.

This direction would benefit from a combined approach that optimizes both weight updates and prompt tuning. For example, the BetterTogether pipeline [423] demonstrates that alternating between weight optimization and prompt optimization outperforms either approach in isolation. This pipeline achieves up to 78% gains on multi-hop QA and reasoning tasks by first optimizing prompts, then fine-tuning weights on bootstrapped traces, and finally re-optimizing prompts on the fine-tuned model. A trained variant of AudioToolAgent (Chapter 6) could adapt this paradigm. In the first stage, reinforcement learning methods such as GRPO [409] or its multi-module extension mmGRPO [606] propagate reward gradients to the orchestrator and the tool models that were invoked during forward passes. In principle, whenever a tool call occurs during rollout, the tool’s parameters receive gradient updates proportional to its contribution to the final reward. This joint optimization would align tool outputs with orchestrator expectations and learn tool-selection policies.

In the second stage, after freezing all model weights, prompt-optimization frameworks (see section 7.4.5) refine the system prompt, tool-query templates, and demonstration examples through reflective search or Bayesian instruction tuning, extracting additional accuracy gains without further gradient computation. This pipeline would be evaluated on MMAU [393], MMAR [280] and MMAU-Pro [224] (similar to Chapter 6 and Chapter 4) to determine whether the combined approach outperforms individual methods in isolation.

7.4.4 Neuro-symbolic AI and Knowledge Graph Integration

Several experiments not included in this dissertation explored the integration of knowledge graphs and symbolic reasoning with audio models. The underlying hypothesis was that structured knowledge could improve audio understanding. Implementations included retrieval-augmented generation pipelines, in which models used knowledge graphs to enrich their context, and agentic layers that dynamically generated SPARQL queries to retrieve relevant information. Despite multiple architectural variations, these neuro-symbolic approaches did not improve performance, often degrading accuracy compared to purely neural baselines.

This outcome suggests a potential domain mismatch between audio understanding and traditional knowledge representation. The hypothesis is that audio captioning mainly involves identifying sound sources such as animals, vehicles, or environmental noise, and describing actions like “dog barking” or “car passing by” [293, 133]. These descriptions require minimal world knowledge or complex reasoning. Knowledge graphs, on the other hand, are structured knowledge models that explicitly store rich factual knowledge [353] and excel in knowledge-intensive tasks such as answering

questions about historical events, scientific facts, or relational reasoning [239]. This simplicity in audio descriptions means that audio models can learn the necessary associations directly from pretrained knowledge or training data without requiring external knowledge bases.

However, recent work in adjacent domains demonstrates how knowledge graphs and retrieval-augmented generation can enhance multimodal reasoning and shows it is possible to combine symbolic and sub-symbolic approaches in multimodal reasoning. Multimodal Knowledge Graphs (MMKGs) organize information from multiple modalities into a single graph structure, to allow cross-modal reasoning and retrieval [232, 357, 177]. Afterward, the information from the graph can be queried and retrieved to pass to an LLM for reasoning [177, 434]. Often, this information query occurs through retrieval-augmented generation (RAG) [239, 129].

In the audio domain specifically, examples of adjacent work include iKnow-audio [342] and WavRAG [65]. iKnow-audio [342] addresses limitations in Contrastive Language-Audio Pretraining (CLAP) by refining CLAP predictions through an Audio-centric Knowledge graph iKnow-audio enhances disambiguation of acoustically similar sounds. Models by integrating knowledge graphs to provide semantic grounding. The framework constructs an Audio-centric Knowledge Graph (AKG) that encodes ontological relations, including semantic, causal, and taxonomic connections found in everyday sound scenes. The evaluations show consistent improvements over the baseline CLAP models. WavRAG [65] introduces the first retrieval-augmented generation framework with native end-to-end audio support. This framework bypasses automatic speech recognition to process raw audio directly, in a way that integrates audio and text into a unified knowledge representation.

These advances suggest several directions for future research on the understanding of neuro-symbolic audio. First, the construction of audio-specific knowledge graphs that encode acoustic information with contextual knowledge of sounds could enable structured reasoning over sound events. This could benefit tasks that require complex audio understanding or audio question answering. Second, benchmarks that explicitly require combining audio perception with world knowledge could reveal whether the limitations observed in the earlier research of this dissertation lie in current task definitions, or whether it is fundamentally unfeasible to have a state-of-the-art neuro-symbolic audio understanding model.

7.4.5 Prompt Sensitivity and Automated Prompt Optimization

Audio-language models exhibit high sensitivity to prompt formatting. For multiple-choice questions, constraining the outputs to begin with one of the answer letters (A, B, C, D) so that the model always returns at least one answer. For captioning tasks, role instructions and output constraints, e.g. “You are an AI assistant for audio captioning” or “Respond with a one-sentence caption”, can meaningfully alter behavior. These

effects introduce variance between benchmarks that does not reflect the underlying capability, highlighting the need for a systematic study of prompt sensitivity.

One direction is to decompose prompts into components such as role framing, problem templates, output schemas, reasoning styles (direct answer vs. brief plan), and tool-use hints when applicable. These components can then be treated as prompt hyperparameters and searched on a fixed compute budget. Such a search would report both accuracy and efficiency metrics (tokens, latency, number of calls). A qualitative analysis of high-performance prompts could further illuminate which prompt structures matter and why. Recent work on ReAct-style prompting [566] suggests that interleaving reasoning and action tokens can also serve as a systematic prompt dimension for audio agents, and systematized prompt optimization frameworks such as Self-Instruct [489] provide blueprints for reproducible prompt release practices.

Prompt components can also be optimized automatically with frameworks such as GEPA [7], PromptBreeder [111] and OPRO [560]. These methods treat prompts as variables and apply search or meta-prompting to improve task performance without modifying model weights. The DSPy framework [197, 196] extends this paradigm by compiling declarative language-model programs into optimized prompt chains: developers specify modules (e.g., retrieval, chain-of-thought, tool calls) and DSPy automatically tunes demonstrations, instructions, and module orderings through teleprompters that maximize a task metric. Applied to an audio tool pipeline (Chapter 6), DSPy could optimize the orchestrator’s system prompt and the decision-making process on which tools to call, and the queries sent to each tool, all from a small set of labeled examples. This approach does not require updating any weights in any model. Other prompt-optimization approaches include MIPROv2 [350] for Bayesian instruction tuning and COPRO [197] for reflective search over prompt templates.

To support reproducibility, future research should release all prompts used in benchmark evaluations (see issues described in section 7.3.2). In addition, evaluation protocols should standardize the prompts between models so that the comparisons reflect the differences between the models rather than the prompt engineering.

7.4.6 World Models and Multi-Sensory Perception

This dissertation focuses on machine listening, a field with clear implications for robotics. World models provide a framework for teaching robots how to perceive and act by learning predictive dynamics of the environment and using them for planning and control. A world model is an internal representation that an agent constructs to understand the dynamics of the environment [153]. The model learns to compute the next state given the current state and an action, and thus can reason about potential future outcomes.

Most of the world model research is based on video as the primary sensory modality [30, 337, 48]. However, sound is also a continuous stream that carries information

about object interactions. Humans rely on hearing as one of their primary senses: auditory cues signal events outside the visual field, provide feedback during manipulation tasks, and enable anticipation of future states. To mimic how humans perceive the world in artificial systems, robots should integrate auditory perception along with vision, touch, and other modalities. Integrating audio perception into vision-language-action models improves robotic manipulation [502].

A direct application of world models to audio involves learning to predict future audio observations. For example, training models for latent generative flow matching to predict the audio consequences of robot actions [575]. This audio world model approach mirrors the success of visual world models: just as NVIDIA Cosmos [337] learns to predict future visual states from multimodal inputs, audio world models learn to imagine future auditory states. The challenge lies in the temporal structure of environmental audio: unlike speech or music, environmental sounds lack the regularity that facilitates prediction.

Humans integrate information from all senses. A robot that perceives through multiple modalities could interact with its surroundings more reliably and in a manner closer to human perception. For example, a robot nowadays can combine visual, auditory, and tactile perception to solve complex tasks that outperform single-modality baselines [242]. A concrete research direction is developing multisensory world models that unify more than three modalities. Inputs could include audition, vision, somatosensation, olfaction, and gustation. A shared latent state would encode these signals, and a decoder would map the latent representation to actions.

Key challenges include learning cross-modal correspondences (e.g., the sound of breaking correlates with tactile discontinuities and visual fragmentation), but also handling missing modalities when sensors fail. Research at this intersection of robotics, perception, and control could build on the demonstration of tool coordination in Chapter 6 and extend world-model techniques from vision-language robotics into the audio domain.

A future where an AI system can hear, see, feel, smell, taste, reason, understand, interpret, act, predict, navigate, and more; all in the world as complex as ours, is rapidly approaching.

Bibliography

-
- [1] Jerome Abdelnour, Giampiero Salvi, and Jean Rouat. *CLEAR: A Dataset for Compositional Language and Elementary Acoustic Reasoning*. Feb. 2019. doi: 10.21227/7X26-A025.
- [2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, et al. *Phi-4 Technical Report*. Dec. 2024. doi: 10.48550/arXiv.2412.08905.
- [3] Shreya G. Abhyankar, Shashank S. Bharadwaj, G. Shobha Rani, Pruthvi G. Karigiri, Sahana Srikanth, et al. “A Survey on Music Genre Classification Using Multimodal Information Processing and Retrieval”. In: *2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC)*. Feb. 2023, pp. 1–6. doi: 10.1109/ICRTEC56977.2023.10111926.
- [4] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. *TAU Spatial Sound Events 2019 - Ambisonic and Microphone Array, Development Datasets*. Feb. 2019. doi: 10.5281/zenodo.2599196.
- [5] Adobe Creative Cloud. *Adobe Audition Sound Effects*. <https://www.adobe.com/products/audition/offers/adobeauditiondlcsfx.html>.
- [6] Pranjal Aggarwal and Sean Welleck. *L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning*. Oct. 2025. doi: 10.48550/arXiv.2503.04697.
- [7] Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziemis, Rishi Khare, et al. “GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning”. In: *First Workshop on Foundations of Reasoning in Language Models*. 2025.
- [8] Oliver Alonzo, Hijung Valentina Shin, and Dingzeyu Li. “Beyond Subtitles: Captioning and Visualizing Non-Speech Sounds to Improve Accessibility of User-Generated Videos”. In: *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 2022, pp. 1–12. doi: 10.1145/3517428.3544808.
- [9] Rosa Ma Alsina-Pagès, Joan Navarro, Francesc Alías, and Marcos Hervás. “homeSound: Real-time Audio Event Detection Based on High Performance Computing for Behaviour and Surveillance Remote Monitoring”. In: *Sensors* 17.854 (2017). doi: 10.3390/s17040854.
- [10] William Alsup. *Bartz et al. v. Anthropic PBC*. Aug. 2024.
- [11] Vanessa Theme Ament. *The Foley Grail: The Art of Performing Sound for Film, Games, and Animation*. Routledge, 2014.
- [12] Ishwarya Ananthabhotla, Sebastian Ewert, and Joseph A. Paradiso. “Towards a Perceptual Loss: Using a Neural Network Codec Approximation as a Loss for Generative Audio Models”. In: *Proceedings of the 27th ACM International Conference on Multimedia*. Mm ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1518–1525. doi: 10.1145/3343031.3351148.
- [13] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. “SPICE: Semantic Propositional Image Caption Evaluation”. In: *ECCV 2016*. Springer International Publishing, July 2016. doi: 10.48550/arXiv.1607.08822.
- [14] Anonymous. *AudioAgent: Enhancing Task Performance Through Modality-Driven Prompt Optimization*. 2024.

- [15] Anthropic. *Introducing the Model Context Protocol*. <https://www.anthropic.com/news/model-context-protocol>. Nov. 2024.
- [16] Anthropic Team. *Introducing Claude 4*. <https://www.anthropic.com/news/claude-4>. May 2025.
- [17] Anupriya. “Transfer Learning for Audio Processing: A Review of Recent Advances”. In: *International Journal of Mechanical Engineering* (2023). doi: 10.56452/6-3-665.
- [18] AudioStock | *High-Quality Royalty-Free Music and Sound Effects*. <https://audiostock.net/>.
- [19] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. *SoundNet: Learning Sound Representations from Unlabeled Video*. 2016. doi: 10.48550/arxiv.1610.09001.
- [20] Jaeyeon Bae, Seokhoon Jeong, Seokun Kang, Namgi Han, Jae-Yon Lee, et al. “Sound of Story: Multi-Modal Storytelling With Audio”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Association for Computational Linguistics, 2023, pp. 13467–13479. doi: 10.18653/V1/2023.FINDINGS-EMNLP.898.
- [21] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. “Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12449–12460. doi: 10.48550/arXiv.2006.11477.
- [22] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, et al. *Qwen Technical Report*. Sept. 2023. doi: 10.48550/arXiv.2309.16609.
- [23] Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, et al. “AudioSetCaps: Enriched Audio Captioning Dataset Generation Using Large Audio Language Models”. In: *Audio Imagination: NeurIPS 2024 Workshop AI-driven Speech, Music, and Sound Generation*. 2024.
- [24] Jisheng Bai, Han Yin, Mou Wang, Dongyuan Shi, Woon-Seng Gan, et al. “Audiolog: LLMs-Powered Long Audio Logging with Hybrid Token-Semantic Contrastive Learning”. In: *2024 IEEE International Conference on Multimedia and Expo (ICME)*. Niagara Falls, ON, Canada: IEEE, July 2024, pp. 1–6. doi: 10.1109/ICME57554.2024.10688214.
- [25] Yatong Bai, Trung Dang, Dung Tran, Kazuhito Koishida, and Somayeh Sojoudi. “ConsistencyTTA: Accelerating Diffusion-Based Text-to-Audio Generation with Consistency Distillation”. In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 3285–3289. doi: 10.21437/Interspeech.2024-1333.
- [26] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. “Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval”. In: *IEEE International Conference on Computer Vision*. IEEE, 2021, pp. 1708–1718. doi: 10.1109/ICCV48922.2021.00175.
- [27] Satanjeev Banerjee and Alon Lavie. “METEOR: An Automatic Metric for MT Evaluation With Improved Correlation With Human Judgments”. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72.
- [28] Andrea Banino, Jan Balaguer, and Charles Blundell. “PonderNet: Learning to Ponder”. In: *8th ICML Workshop on Automated Machine Learning (AutoML)*. 2021.
- [29] Megan K. Bannigan, Christopher S. Ford, Abigail Liles, and Joshua Plastrik. *Anthropic and Meta Decisions on Fair Use*. <https://www.debevoise.com/insights/publications/2025/06/anthropic-and-meta-decisions-on-fair-use>. June 2025.
- [30] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. “Navigation World Models”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 15791–15801.
- [31] *BBC Sound Effects*. <https://sound-effects.bbcrewind.co.uk/>.
- [32] Edward Beeching, Lewis Tunstall, and Sasha Rush. *Scaling Test-Time Compute - a Hugging Face Space by Huggingface4*. Dec. 2024.

-
- [33] Swarup Ranjan Behera, Krishna Mohan Injeti, Jaya Sai Kiran Patibandla, Praveen Kumar Pokala, and Balakrishna Reddy Pailla. *AQUALLM: Audio Question Answering Data Generation Using Large Language Models*. Dec. 2023. doi: 10.48550/arXiv.2312.17343.
- [34] Swarup Ranjan Behera, Pailla Balakrishna Reddy, Achyut Mani Tripathi, Megavath Bharadwaj Rathod, and Tejesh Karavadi. “Towards Multi-Lingual Audio Question Answering”. In: *INTER-SPEECH 2023*. ISCA, Aug. 2023, pp. 356–360. doi: 10.21437/Interspeech.2023-1816.
- [35] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. “Curriculum Learning”. In: *International Conference on Machine Learning*. Icml ’09. ACM, June 2009, pp. 41–48. doi: 10.1145/1553374.1553380.
- [36] Lukas Biewald. *Experiment Tracking With Weights and Biases*. 2020.
- [37] Jennifer K. Bizley and Yale E. Cohen. “The What, Where and How of Auditory-Object Perception”. In: *Nature Reviews. Neuroscience* 14.10 (Oct. 2013), pp. 693–707. doi: 10.1038/nrn3565.
- [38] Hans C Boas and Ryan Dux. “From the Past into the Present: From Case Frames to Semantic Frames”. In: *Linguistics Vanguard* 3.1 (Aug. 2017), p. 20160003. doi: 10.1515/lingvan-2016-0003.
- [39] Wim Boes and Hugo Van Hamme. *Impact of Visual Assistance for Automated Audio Captioning*. 2022. doi: 10.48550/arXiv.2211.10539.
- [40] Oliver Bones, Trevor J Cox, and William J Davies. “Sound Categories: Category Formation and Evidence-Based Taxonomies”. In: *Frontiers in Psychology* 9 (July 2018), p. 1277. doi: 10.3389/fpsyg.2018.01277.
- [41] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, et al. “AudioLM: A Language Modeling Approach to Audio Generation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2523–2533. doi: 10.1109/TASLP.2023.3288409.
- [42] Dimitrios Bralios, Gordon Wichern, François G. Germain, Zexu Pan, Sameer Khurana, et al. “Generation or Replication: Auscultating Audio Latent Diffusion Models”. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 1156–1160. doi: 10.1109/ICASSP48485.2024.10447705.
- [43] Hervé Bredin. “Pyannote.Audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe”. In: *Proc. INTERSPEECH 2023*. Aug. 2023, pp. 1983–1987. doi: 10.21437/interspeech.2023-105.
- [44] Julie A. Brefczynski-Lewis and James W. Lewis. “Auditory Object Perception: A Neurobiological Model and Prospective Review”. In: *Neuropsychologia* 105 (Oct. 2017), pp. 223–242. doi: 10.1016/j.neuropsychologia.2017.04.034.
- [45] *Bring Your Story to Life | Music & Sfx for Videos | Epidemic Sound*. <https://www.epidemicsound.com/>.
- [46] Andrei Z Broder. “On the Resemblance and Containment of Documents”. In: *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, 1998, pp. 21–29. doi: 10.1109/sequen.1997.666900.
- [47] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. “Language Models Are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901. doi: 10.5555/3495724.3495883.
- [48] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, et al. “Genie: Generative Interactive Environments”. In: *Forty-First International Conference on Machine Learning*. 2024.
- [49] Bruce D. Burns. “The Effects of Speed on Skilled Chess Performance”. In: *Psychological Science* 15.7 (July 2004), pp. 442–447. doi: 10.1111/j.0956-7976.2004.00699.x.

- [50] Pengfei Cai, Yan Song, Qing Gu, Nan Jiang, Haoyu Song, et al. *Detect Any Sound: Open-Vocabulary Sound Event Detection with Multi-Modal Queries*. Oct. 2025. doi: 10.48550/arXiv.2507.16343.
- [51] Xi-Ren Cao and Ruey-wen Liu. “General Approach to Blind Source Separation”. In: *IEEE Transactions on Signal Processing* 44.3 (1996), pp. 562–571. doi: 10.1109/78.489029.
- [52] Mark Cartwright, Ana Elisa Mendez Mendez, Aurora Cramer, Vincent Lostanlen, Graham Dove, et al. “SONYC Urban Sound Tagging (SONYC-UST): A Multilabel Dataset from an Urban Acoustic Sensor Network”. In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. 2019, pp. 35–39. doi: 10.33682/j5zw-2t88.
- [53] Ernie Chang, Pin-Jie Lin, Yang Li, Sidd Srinivasan, Gael Le Lan, et al. “In-Context Prompt Editing for Conditional Audio Generation”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. arxiv.org, Apr. 2024, pp. 5320–5324. doi: 10.1109/icassp48485.2024.10446431.
- [54] Moitrey Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. “Visual Scene Graphs for Audio Source Separation”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 1184–1193. doi: 10.1109/ICCV48922.2021.00124.
- [55] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, et al. “Soundspaces 2.0: A Simulation Platform for Visual-Acoustic Learning”. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. Ed. by Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, et al. arXiv, 2022.
- [56] David L. Chen and William B. Dolan. “Collecting Highly Parallel Data for Paraphrase Evaluation”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*. Portland, OR: Association for Computational Linguistics, June 2011, pp. 190–200.
- [57] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. “VGGSound: A Large-Scale Audio-Visual Dataset”. In: *ICASSP*. Barcelona, Spain: IEEE, May 2020, pp. 721–725. doi: 10.1109/ICASSP40776.2020.9053174.
- [58] Kun Chen, Yusong Wu, Ziyue Wang, Xuan Zhang, Fudong Nian, et al. “Audio Captioning Based on Transformer and Pre-Trained Cnn.” In: *DCASE*. 2020, pp. 21–25.
- [59] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, et al. “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”. In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (Oct. 2022), pp. 1505–1518. doi: 10.1109/JSTSP.2022.3188113.
- [60] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, et al. “BEATs: Audio Pre-Training with Acoustic Tokenizers”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2022, pp. 5178–5193. doi: 10.48550/ARXIV.2212.09058.
- [61] Shunian Chen, Xinyuan Xie, Zheshu Chen, Liyan Zhao, Owen Lee, et al. *FusionAudio-1.2 M: Towards Fine-Grained Audio Captioning with Multimodal Contextual Fusion*. 2025. doi: 10.48550/arXiv.2506.01111.
- [62] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, et al. *VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset*. Apr. 2023. doi: 10.48550/arXiv.2304.08345.
- [63] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, et al. “Vast: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset”. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, et al. arXiv, 2023.

-
- [64] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, et al. *Microsoft Coco Captions: Data Collection and Evaluation Server*. Apr. 2015. doi: 10.48550/arXiv.1504.00325.
- [65] Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, et al. *WavRAG: Audio-Integrated Retrieval Augmented Generation for Spoken Dialogue Models*. Feb. 2025. doi: 10.48550/arXiv.2502.14727.
- [66] Vince Chhabria and Sallie Kim. *Kadrey et al. v. Meta Platforms, Inc.* July 2023.
- [67] Jae-Heung Cho, Yoon-Ah Park, Jaewon Kim, and Joon-Hyuk Chang. “Hyu Submission for the Dcase 2023 Task 6a: Automated Audio Captioning Model Using Al-Mixgen and Synonyms Substitution”. In: *Proc. Detection and Classification of Acoustic Scenes and Events*. 2023.
- [68] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches”. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 2014, pp. 103–111. doi: 10.3115/v1/w14-4012.
- [69] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, et al. “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. doi: 10.48550/arXiv.1406.1078.
- [70] Seungjin Choi, Andrzej Cichocki, Hyung-Min Park, and Soo-Young Lee. “Blind Source Separation and Independent Component Analysis: A Review”. In: *Neural Information Processing-Letters and Reviews* 6.1 (2005), pp. 1–57.
- [71] Ho-Young Choi, Jae-Heung Cho, Pil Moo Byun, Won-Gook Choi, and Joon-Hyuk Chang. “Temp4Cap: Temporally-Aligned Automated Audio Captioning”. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. International Speech Communication Association, 2025, pp. 3135–3139.
- [72] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, et al. *Qwen2-Audio Technical Report*. July 2024. doi: 10.48550/arXiv.2407.10759.
- [73] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, et al. *Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models*. Dec. 2023. doi: 10.48550/arXiv.2311.07919.
- [74] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, et al. “Scaling Instruction-Finetuned Language Models”. In: *Journal of Machine Learning Research* 25 (2024), 70:1–70:53.
- [75] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, et al. *Training Verifiers to Solve Math Word Problems*. Nov. 2021. doi: 10.48550/arXiv.2110.14168.
- [76] Jonathan Cohen. *Now We’re Talking: Nvidia Releases Open Dataset, Models for Multilingual Speech Ai*. Aug. 2025.
- [77] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, et al. *Gemini 2.5: Pushing the Frontier With Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*. July 2025. doi: 10.48550/arXiv.2507.06261.
- [78] Matéo Cousin, Étienne Labbé, and Thomas Pellegrini. *Multilingual Audio Captioning Using Machine Translated Data*. Sept. 2023. doi: 10.48550/arXiv.2309.07615.
- [79] Muzhi Dai, Shixuan Liu, and Qingyi Si. *Stable Reinforcement Learning for Efficient Reasoning*. 2025. doi: 10.48550/ARXIV.2505.18086.

- [80] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, et al. “InstructBLIP: Towards General-Purpose Vision-Language Models With Instruction Tuning”. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, et al. arXiv, 2023.
- [81] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, et al. “Scaling Egocentric Vision: The Epic-Kitchens Dataset”. In: *European Conference on Computer Vision (ECCV)*. 2018.
- [82] Quy-Anh Dang and Chris Ngo. *Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't*. Mar. 2025. doi: 10.48550/arXiv.2503.16219.
- [83] Theodore Lamort de Gail, Bartłomiej Zgórzyński, Anna Płeś, and Kamil Górzynski. *Take It Seriously: Improving on Last Year With All Sorts of Tricks*. Tech. rep. DCASE2023 Challenge, June 2023.
- [84] Amalia De Götzen, Erik Sikström, Francesco Grani, and Stefania Serafin. “Real, Foley or Synthetic? An Evaluation of Everyday Walking Sounds”. In: *Proceedings of SMC* (2013).
- [85] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. *DeepSeek-V3 Technical Report*. Feb. 2025. doi: 10.48550/arXiv.2412.19437.
- [86] Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, et al. “PAM: Prompting Audio-Language Models for Audio Quality Assessment”. In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 3320–3324. doi: 10.21437/Interspeech.2024-325.
- [87] Soham Deshmukh, Satvik Dixit, Rita Singh, and Bhiksha Raj. *Mellow: A Small Audio Language Model for Reasoning*. Mar. 2025. doi: 10.48550/arXiv.2503.08540.
- [88] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. “Pengi: An Audio Language Model for Audio Tasks”. In: *Proc. NeurIPS*. Ed. by Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, et al. 2023.
- [89] Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. “Audio Retrieval with WavText5K and CLAP Training”. In: *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 2948–2952. doi: 10.21437/Interspeech.2023-1136.
- [90] Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, et al. *Audio Entailment: Assessing Deductive Reasoning for Audio Understanding*. 2024. doi: 10.48550/arXiv.2407.18062.
- [91] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
- [92] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. “Demographics and Dynamics of Mechanical Turk Workers”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. Wsdm '18. New York, NY, USA: Association for Computing Machinery, Feb. 2018, pp. 135–143. doi: 10.1145/3159652.3159661.
- [93] Cleyton Aparecido Dim, Rafael Martins Feitosa, Marcelle Pereira Mota, and Jefferson Magalhães de Morais. “Alert Systems to Hearing-Impaired People: A Systematic Review”. In: *Multimedia Tools and Applications* 81.22 (2022), pp. 32351–32370. doi: 10.1007/s11042-022-13045-1.
- [94] Mark Dingemanse. “Redrawing the Margins of Language: Lessons from Research on Ideophones”. In: *Glossa: a Journal of General Linguistics* 3.1 (2018). doi: 10.5334/gjgl.444.

-
- [95] Satvik Dixit, Soham Deshmukh, and Bhiksha Raj. “Mace: Leveraging Audio for Evaluating Audio Captioning Systems”. In: *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2025, pp. 1–5. doi: 10.1109/ICASSPW65056.2025.11011270.
- [96] Seungheon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. “LP-MusicCaps: LLM-Based Pseudo Music Captioning”. In: *24th International Society for Music Information Retrieval Conference*. International Society for Music Information Retrieval Conference, 2023, pp. 409–416. doi: 10.48550/arxiv.2307.16372.
- [97] Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. “Strong Model Collapse”. In: *The Thirteenth International Conference on Learning Representations*. 2025.
- [98] Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian J. McAuley, and Taylor Berg-Kirkpatrick. “CLIPSep: Learning Text-Queried Sound Separation With Noisy Unlabeled Videos”. In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [99] Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ziyi Xu, Yilong Zhao, et al. “XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models”. In: *Eighth Conference on Machine Learning and Systems*. 2024. doi: 10.48550/arXiv.2411.15100.
- [100] Download FREE Sound Effects at Zapsplat. <https://www.zapsplat.com/>.
- [101] Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen. “Automated Audio Captioning With Recurrent Neural Networks”. In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. arXiv:1706.10006. Oct. 2017, pp. 374–378. doi: 10.48550/arXiv.1706.10006.
- [102] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. “Clotho: An Audio Captioning Dataset”. In: *ICASSP*. May 2020, pp. 736–740. doi: 10.1109/ICASSP40776.2020.9052990.
- [103] DuckDuckGo - Protection. Privacy. Peace of Mind. <https://duckduckgo.com>.
- [104] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. “CLAP Learning Audio Concepts from Natural Language Supervision”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, June 2023, pp. 1–5. doi: 10.1109/icassp49357.2023.10095889.
- [105] Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. “Natural Language Supervision for General-Purpose Audio Representations”. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 336–340. doi: 10.1109/ICASSP48485.2024.10448504.
- [106] J. L. Elman. “Learning and Development in Neural Networks: The Importance of Starting Small”. In: *Cognition* 48.1 (July 1993), pp. 71–99. doi: 10.1016/0010-0277(93)90058-4.
- [107] Ayşegül Özkaya Eren and Mustafa Sert. “Audio Captioning Based on Combined Audio and Semantic Embeddings”. In: *2020 IEEE International Symposium on Multimedia (ISM)*. Dec. 2020, pp. 41–48. doi: 10.1109/ISM.2020.00014.
- [108] European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance)*. Apr. 2016.
- [109] European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (AI Act)*. Official Journal of the European Union. June 2024.
- [110] Haytham M Fayek and Justin Johnson. “Temporal Reasoning via Audio Question Answering”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2283–2294. doi: 10.1109/TASLP.2020.3010650.

- [111] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. *Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution*. Sept. 2023. doi: 10.48550/arXiv.2309.16797.
- [112] *FindSounds - Browse for Sounds*. <https://www.findsounds.com/>.
- [113] Arthur Flexer and Dominik Schnitzer. “Effects of Album and Artist Filters in Audio Similarity Computed for Very Large Music Databases”. In: *Computer Music Journal* 34.3 (Sept. 2010), pp. 20–28. doi: 10.1162/COMJ_a_00004.
- [114] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. “FSD50K: An Open Dataset of Human-Labeled Sound Events”. In: *IEEE ACM Trans. Audio Speech Lang. Process.* 30 (2022), pp. 829–852. doi: 10.1109/TASLP.2021.3133208.
- [115] Eduardo Fonseca, Manoj Plakal, Daniel PW Ellis, Frederic Font, Xavier Favory, et al. “Learning Sound Event Classifiers from Web Audio with Noisy Labels”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 21–25.
- [116] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, et al. *General-Purpose Tagging of Freesound Audio With Audioset Labels: Task Description, Dataset, and Baseline*. Oct. 2018. doi: 10.48550/arXiv.1807.09902.
- [117] Frederic Font Corbera, Gerard Roma Trepas, and Xavier Serra. *Freesound Technical Demo*. ACM Association for Computer Machinery, 2013. doi: 10.1145/2502081.2502245.
- [118] Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D Plumbley. “Chime-Home: A Dataset for Sound Source Recognition in a Domestic Environment”. In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, Oct. 2015, pp. 1–5. doi: 10.1109/waspaa.2015.7336899.
- [119] *Free Sound Clips | Soundbible.Com*. <https://soundbible.com/>.
- [120] *Free Sound Effects | Soundjay.Com*. <https://www.soundjay.com/>.
- [121] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, et al. “DataComp: In Search of the Next Generation of Multimodal Datasets”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 27092–27112. doi: 10.48550/arXiv.2304.14108.
- [122] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. “Tall: Temporal Activity Localization via Language Query”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 5277–5285. doi: 10.1109/ICCV.2017.563.
- [123] William W. Gaver. “What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception”. In: *Ecological Psychology* 5.1 (Mar. 1993), pp. 1–29. doi: 10.1207/s15326969eco0501_1.
- [124] Gemini Team. *A New Era of Intelligence with Gemini 3*. <https://blog.google/products/gemini/gemini-3/>. Nov. 2025.
- [125] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, et al. *Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context*. Dec. 2024. doi: 10.48550/arXiv.2403.05530.
- [126] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, et al. “Audio Set: An Ontology and Human-Labeled Dataset for Audio Events”. In: *Proc. IEEE ICASSP 2017*. New Orleans, LA: IEEE, 2017, pp. 776–780. doi: 10.1109/ICASSP.2017.7952261.
- [127] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. *Text-to-Audio Generation Using Instruction-Tuned Llm and Latent Diffusion Model*. May 2023. doi: 10.48550/arXiv.2304.13731.
- [128] Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S. Sakshi, Jaehyeon Kim, et al. *Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities*. Mar. 2025. doi: 10.48550/arXiv.2503.03983.

-
- [129] Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Ramani Duraiswami, and Dinesh Manocha. "RECAP: Retrieval-Augmented Audio Captioning". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2024, pp. 1161–1165. doi: 10.1109/icassp48485.2024.10448030.
- [130] Sreyan Ghosh, Sonal Kumar, Zhifeng Kong, Rafael Valle, Bryan Catanzaro, et al. "Synthio: Augmenting Small-Scale Audio Classification Datasets with Synthetic Data". In: *The Thirteenth International Conference on Learning Representations*. 2025.
- [131] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, et al. "GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities". In: *EMNLP*. 2024, pp. 6288–6313. doi: 10.18653/v1/2024.emnlp-main.361.
- [132] Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, et al. "CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models". In: *International Conference on Learning Representations*. 2023. doi: 10.48550/arXiv.2310.08753.
- [133] Bruno L. Giordano, Ricardo Miranda Azevedo, Yenisel Plasencia-Calaña, Elia Formisano, and Michel Dumontier. "What Do We Mean With Sound Semantics, Exactly? A Survey of Taxonomies and Ontologies of Everyday Sounds". In: *Frontiers in Psychology* 13 (2022).
- [134] Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, et al. "Audio Flamingo 3: Advancing Audio Intelligence With Fully Open Large Audio Language Models". In: *Neural Information Processing Systems (NeurIPS)*. July 2025. doi: 10.48550/arXiv.2507.08128.
- [135] Arushi Goel, Zhifeng Kong, Rafael Valle, and Bryan Catanzaro. *Audio Dialogues: Dialogues Dataset for Audio and Music Understanding*. Apr. 2024. doi: 10.48550/arXiv.2404.07616.
- [136] Artthur Gollwitzer. *Federal Courts Find Fair Use in AI Training: Key Takeaways from Kadrey v. Meta and Bartz v. Anthropic*. July 2025.
- [137] Yuan Gong, Yu-An Chung, and James Glass. "AST: Audio Spectrogram Transformer". In: *Proceedings of Interspeech 2021*. Aug. 2021, pp. 571–575. doi: 10.21437/interspeech.2021-698.
- [138] Yuan Gong, Yu-An Chung, and James Glass. "PSLA: Improving Audio Tagging With Pretraining, Sampling, Labeling, and Aggregation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3292–3306. doi: 10.1109/TASLP.2021.3120633.
- [139] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James R Glass. "Listen, Think, and Understand". In: *International Conference on Learning Representations*. 2023. doi: 10.48550/arXiv.2305.10790.
- [140] Félix Gontier, Romain Serizel, and Christophe Cerisara. "Automated Audio Captioning by Fine-Tuning Bart With Audioset Tags". In: *DCASE 2021 - 6th Workshop on Detection and Classification of Acoustic Scenes and Events*. Nov. 2021.
- [141] Félix Gontier, Romain Serizel, and Christophe Cerisara. "SPICE+: Evaluation of Automatic Audio Captioning Systems with Pre-Trained Language Models". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece, June 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10097021.
- [142] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. *The Llama 3 Herd of Models*. 2024. doi: 10.48550/ARXIV.2407.21783.
- [143] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, et al. "Ego4D: Around the World in 3,000 Hours of Egocentric Video". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2022, pp. 18973–18990. doi: 10.1109/CVPR52688.2022.01842.
- [144] Alex Graves. *Adaptive Computation Time for Recurrent Neural Networks*. Feb. 2017. doi: 10.48550/arXiv.1603.08983.

- [145] Fuyu Gu, Yang Gu, Yiyan Xu, Haoran Sun, Yushan Pan, et al. “Language-Based Audio Retrieval With GPT-Augmented Captions and Self-Attended Audio Clips”. In: *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. Tianjin, China: IEEE, May 2024, pp. 858–863. doi: 10.1109/CSCWD61410.2024.10580534.
- [146] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, et al. *Accelerate: Training and Inference at Scale Made Simple, Efficient and Adaptable*. 2022.
- [147] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, et al. *Textbooks Are All You Need*. Oct. 2023. doi: 10.48550/arXiv.2306.11644.
- [148] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. “DeepSeek-R1 Incentivizes Reasoning in LLMs Through Reinforcement Learning”. In: *Nature* 645.8081 (Sept. 2025), pp. 633–638. doi: 10.1038/s41586-025-09422-z.
- [149] Tianyu Guo, Hongyu Chen, Hao Liang, Meiyi Qiang, Bohan Zeng, et al. “BRACE: A Benchmark for Robust Audio Caption Quality Evaluation”. In: *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2025.
- [150] Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, et al. “Audio Generation with Multiple Conditional Diffusion Model”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.16 (Mar. 2024), pp. 18153–18161. doi: 10.1609/aaai.v38i16.29773.
- [151] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. *AudioCLIP: Extending CLIP to Image, Text and Audio*. June 2021. doi: 10.48550/arXiv.2106.13043.
- [152] Brian Gygi, Gary R Kidd, and Charles S Watson. “Similarity and Categorization of Environmental Sounds”. In: *Perception and Psychophysics* 69.6 (Aug. 2007), pp. 839–855. doi: 10.3758/bf03193921.
- [153] David Ha and Jürgen Schmidhuber. “Recurrent World Models Facilitate Policy Evolution”. In: *Advances in Neural Information Processing Systems*. Neural information Processing Systems Foundation, 2018, pp. 2450–2462.
- [154] Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, et al. “EzAudio: Enhancing Text-to-Audio Generation with Efficient Diffusion Transformer”. In: *CoRR* (2024). doi: 10.48550/ARXIV.2409.10819.
- [155] Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, Sergey Tulyakov, et al. *Taming Data and Transformers for Audio Generation*. Oct. 2024. doi: 10.48550/arXiv.2406.19388.
- [156] Sahibzada Saadoon Hammad, Ditsuhi Iskandaryan, and Sergio Trilles. “An Unsupervised TinyML Approach Applied to the Detection of Urban Noise Anomalies Under the Smart Cities Environment”. In: *Internet of Things* 23 (2023), p. 100848. doi: 10.1016/j.iot.2023.100848.
- [157] Qichen Han, Weiqiang Yuan, Dong Liu, Xiang Li, and Zhen Yang. “Automated Audio Captioning With Weakly Supervised Pre-Training and Word Selection Methods.” In: *DCASE*. 2021, pp. 6–10.
- [158] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, et al. *Deep Speech: Scaling up End-to-End Speech Recognition*. Dec. 2014. doi: 10.48550/arXiv.1412.5567.
- [159] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, et al. “Training Large Language Models to Reason in a Continuous Latent Space”. In: *Workshop on Reasoning and Planning for Large Language Models*. 2025.
- [160] Xiaoshuai Hao, Yi Zhu, Srikanth Appalaraju, Aston Zhang, Wanqian Zhang, et al. “MixGen: A New Multi-Modal Data Augmentation”. In: *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*. IEEE, 2023, pp. 379–389. doi: 10.1109/WACVW58289.2023.00042.
- [161] Haibo He and Edward A. Garcia. “Learning from Imbalanced Data”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (Sept. 2009), pp. 1263–1284. doi: 10.1109/TKDE.2008.239.

-
- [162] Yingxu He, Zhuohan Liu, Geyu Lin, Shuo Sun, Bin Wang, et al. “MERaLiON-AudioLLM: Advancing Speech and Language Understanding for Singapore”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Ed. by Pushkar Mishra, Smaranda Muresan, and Tao Yu. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 22–30. doi: 10.18653/v1/2025.acl-demo.3.
 - [163] Rajat Hebbar, Digbalay Bose, Krishna Somandepalli, Veena Vijai, and Shrikanth Narayanan. “A Dataset for Audio-Visual Sound Event Detection in Movies”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, June 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10094781.
 - [164] Konrad Heidler, Lichao Mou, Di Hu, Pu Jin, Guangyao Li, et al. “Self-Supervised Audiovisual Representation Learning for Remote Sensing Data”. In: *International Journal of Applied Earth Observation and Geoinformation* 116 (Feb. 2023), p. 103130. doi: 10.1016/j.jag.2022.103130.
 - [165] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. *TAU Urban Acoustic Scenes 2019, Development Dataset*. <https://zenodo.org/records/2589280>. Mar. 2019.
 - [166] Richard P. Heitz. “The Speed-Accuracy Tradeoff: History, Physiology, Methodology, and Behavior”. In: *Frontiers in Neuroscience* 8 (2014), p. 150. doi: 10.3389/fnins.2014.00150.
 - [167] Laura Hekanaho, Maija Hirvonen, and Tuomas Virtanen. “Language-Based Machine Perception: Linguistic Perspectives on the Compilation of Captioning Datasets”. In: *Digital Scholarship in the Humanities* 39.3 (Sept. 2024), pp. 864–883. doi: 10.1093/llc/fqae029.
 - [168] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, et al. “Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning”. In: *Journal of Machine Learning Research* 21.248 (2020), pp. 1–43.
 - [169] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, et al. “Localizing Moments in Video With Temporal Language”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii. Association for Computational Linguistics, 2018, pp. 1380–1390. doi: 10.18653/V1/D18-1168.
 - [170] Shawn Hershey, Daniel P W Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, et al. “The Benefit of Temporally-Strong Labels in Audio Event Classification”. In: *ICASSP*. Toronto, ON, Canada: IEEE, June 2021, pp. 366–370. doi: 10.1109/ICASSP39728.2021.9414579.
 - [171] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, et al. *Deep Learning Scaling Is Predictable, Empirically*. Dec. 2017. doi: 10.48550/arXiv.1712.00409.
 - [172] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the Knowledge in a Neural Network”. In: *NIPS 2014 Deep Learning Workshop*. Mar. 2015. doi: 10.48550/arXiv.1503.02531.
 - [173] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, et al. *Training Compute-Optimal Large Language Models*. Mar. 2022. doi: 10.48550/arXiv.2203.15556.
 - [174] *Home - We Sound Effects*. <https://wesoundeffects.com/>. May 2023.
 - [175] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. “Unnatural Instructions: Tuning Language Models With (Almost) No Human Labor”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki. Association for Computational Linguistics, 2023, pp. 14409–14428. doi: 10.18653/V1/2023.ACL-LONG.806.
 - [176] Yuanbo Hou, Qiaoqiao Ren, Andrew Mitchell, Wenwu Wang, Jian Kang, et al. *Soundscape Captioning Using Sound Affective Quality Network and Large Language Model*. June 2024. doi: 10.48550/arXiv.2406.05914.
 - [177] Chi-Hsiang Hsiao, Yi-Cheng Wang, Tzung-Sheng Lin, Yi-Ren Yeh, and Chu-Song Chen. *MegaRAG: Multimodal Knowledge Graph-Based Retrieval Augmented Generation*. Nov. 2025. doi: 10.48550/arXiv.2512.20626.

- [178] Chengming Hu, Xuan Li, Dan Liu, Haolun Wu, Xi Chen, et al. *Teacher-Student Architecture for Knowledge Distillation: A Survey*. Aug. 2023. doi: 10.48550/arXiv.2308.04268.
- [179] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuezhi Li, et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations*. 2022.
- [180] Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, et al. “WavLLM: Towards Robust and Adaptive Speech Large Language Model”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 4552–4572. doi: 10.18653/v1/2024.findings-emnlp.263.
- [181] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, et al. *Make-An-Audio 2: Temporal-Enhanced Text-to-Audio Generation*. May 2023. doi: 10.48550/arXiv.2305.18474.
- [182] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, et al. “Make-An-Audio: Text-to-Audio Generation With Prompt-Enhanced Diffusion Models”. In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 13916–13932.
- [183] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, et al. “AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head”. In: *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*. Ed. by Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan. AAAI Press, 2024, pp. 23802–23804. doi: 10.1609/AAAI.V38I21.30570.
- [184] Zhaolan Huang, Adrien Tousnakhoff, Polina Kozyr, Roman Rehausen, Felix Bießmann, et al. “TinyChirp: Bird Song Recognition Using TinyML Models on Low-Power Wireless Acoustic Sensors”. In: *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10. doi: 10.1109/IS262782.2024.10704131.
- [185] Il-Young Jeong and Jeongsoo Park. “CochlScene: Acquisition of Acoustic Scene Data Using Crowdsourcing”. In: *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Chiang Mai, Thailand: IEEE, Nov. 2022, pp. 17–21. doi: 10.23919/APSIPAASC55919.2022.9979822.
- [186] Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, et al. *AudioEditor: A Training-Free Diffusion-Based Audio Editing Framework*. 2024. doi: 10.48550/arXiv.2409.12466.
- [187] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, et al. *Mixtral of Experts*. Jan. 2024. doi: 10.48550/arXiv.2401.04088.
- [188] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, et al. “MiraData: A Large-Scale Video Dataset with Long Durations and Structured Captions”. In: *NeurIPS 37 (2024)*, pp. 48955–48970. doi: 10.48550/ARXIV.2407.06358.
- [189] Sunghye Jung, Donghun Lee, Shinbok Lee, Gaeun Seo, Daniel Lee, et al. *DiaTool-DPO: Multi-Turn Direct Preference Optimization for Tool-Augmented Large Language Models*. July 2025. doi: 10.48550/arXiv.2504.02882.
- [190] Marek Kadlčík, Adam Hájek, Jürgen Kieslich, and Radosław Winiecki. *A Whisper Transformer for Audio Captioning Trained with Synthetic Captions and Transfer Learning*. May 2023. doi: 10.48550/arXiv.2305.09690.
- [191] Daniel Kahneman. *Thinking, Fast and Slow*. macmillan, 2011.
- [192] Minsung Kang, Sangshin Oh, Hyeongi Moon, Kyungyun Lee, and Ben Sangbae Chon. *FALL-E: A Foley Sound Synthesis Model and Strategies*. Aug. 2023. doi: 10.48550/arXiv.2306.09807.

-
- [193] Taeho Kang, Juhee Kim, and Joonha Lee. “Development of Korean Audio Caption System”. In: *Proceedings of the Korean Society of Broadcast Engineers Conference*. The Korean Institute of Broadcast and Media Engineers, 2020, pp. 364–367.
 - [194] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, et al. *Scaling Laws for Neural Language Models*. 2020. doi: 10.48550/ARXIV.2001.08361.
 - [195] Paul Kay and Willett Kempton. “What Is the Sapir-Whorf Hypothesis?” In: *American Anthropologist* 86 (Mar. 1984), pp. 65–79. doi: 10.1525/aa.1984.86.1.02a00050.
 - [196] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, et al. *Demonstrate-Search-Predict: Composing Retrieval and Language Models for Knowledge-Intensive NLP*. Jan. 2023. doi: 10.48550/arXiv.2212.14024.
 - [197] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, et al. “DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines”. In: *The Twelfth International Conference on Learning Representations*. 2024.
 - [198] Kevin Kilgour, Beat Gfeller, Qingqing Huang, Aren Jansen, Scott Wisdom, et al. “Text-Driven Separation of Arbitrary Sounds”. In: *Proc. Interspeech 2022*. Sept. 2022, pp. 5403–5407. doi: 10.21437/interspeech.2022-11052.
 - [199] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. “AudioCaps: Generating Captions for Audios in The Wild”. In: *Proceedings of the Association for Computational Linguistics* (2019), pp. 119–132. doi: 10.18653/v1/N19-1011.
 - [200] Eungbeom Kim, Jinhee Kim, Yoori Oh, Kyungsu Kim, Minju Park, et al. *Exploring Train and Test-Time Augmentations for Audio-Language Learning*. May 2023. doi: 10.48550/arXiv.2210.17143.
 - [201] Jaewon Kim, Yoon-Ah Park, Jae-Heung Cho, and Joon-Hyuk Chang. “Improving Automated Audio Captioning Fluency Through Data Augmentation and Ensemble Selection”. In: *Proceedings of Conference on Detection and Classification of Acoustic Scenes and Events*. 2023, pp. 86–90.
 - [202] Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo. “EnCLAP: Combining Neural Audio Codec and Audio-Text Joint Embedding for Automated Audio Captioning”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2024, pp. 6735–6739. doi: 10.1109/icassp48485.2024.10446672.
 - [203] Jinhee Kim, Chang-Bin Jeon, Yoori Oh, JoonHyeon Bae, and Kyogu Lee. “Overcoming Data Shortage in Audio-Text Multi-Modal Retrieval: A Tech Report for Dcase 2023 Challenge”. In: *IEEE AASP Challenge Detection Classification Acoust. Scenes Events, Tampere, Finland, Tech. Rep* (2023).
 - [204] Minkyu Kim, Kim Sung-Bin, and Tae-Hyun Oh. “Prefix Tuning for Automated Audio Captioning”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096877.
 - [205] KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, et al. *Kimi-Audio Technical Report*. Apr. 2025. doi: 10.48550/arXiv.2504.18425.
 - [206] Anssi Klapuri and Manuel Davy. “Signal Processing Methods for Music Transcription”. In: (2007).
 - [207] A Sophia Koepke, Andreea-Maria Oncescu, João F Henriques, Zeynep Akata, and Samuel Albanie. “Audio Retrieval With Natural Language Queries: A Benchmark Study”. In: *IEEE Transactions on Multimedia* 25 (2023), pp. 2675–2685. doi: 10.1109/tmm.2022.3149712.
 - [208] Andrew Koh, Fuzhao Xue, and Eng Siong Chng. *Automated Audio Captioning Using Transfer Learning and Reconstruction Latent Space Similarity Regularization*. Aug. 2021. doi: 10.48550/arXiv.2108.04692.
 - [209] Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. *Audio Captioning Using Pre-Trained Large-Scale Language Model Guided by Audio-based Similar Caption Retrieval*. Dec. 2020. doi: 10.48550/arXiv.2012.07331.

- [210] Yuma Koizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. *The NTTD DCASE2020 Challenge Task 6 System: Automated Audio Captioning With Keywords and Sentence Length Estimation*. July 2020. doi: 10.48550/arXiv.2007.00225.
- [211] Tatsuya Komatsu, Yusuke Fujita, Kazuya Takeda, and Tomoki Toda. “Audio Difference Learning for Audio Captioning”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2024, pp. 1456–1460. doi: 10.1109/icassp48485.2024.10448085.
- [212] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, et al. “PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2880–2894. doi: 10.1109/TASLP.2020.3030497.
- [213] Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, et al. “Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities”. In: *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 25125–25148.
- [214] Zhifeng Kong, Sang-gil Lee, Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, et al. “Improving Text-To-Audio Models with Synthetic Captions”. In: *Synthetic Data’s Transformative Role in Foundational Speech Models*. Aug. 2024, pp. 1–5. doi: 10.21437/SynData4GenAI.2024-1.
- [215] Sandeep Kothinti and Dimitra Emmanouilidou. *Investigations in Audio Captioning: Addressing Vocabulary Imbalance and Evaluating Suitability of Language-Centric Performance Metrics*. May 2023. doi: 10.48550/arXiv.2211.06547.
- [216] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. “Efficient Training of Audio Transformers With Patchout”. In: *Interspeech 2022*. Mar. 2022. doi: 10.48550/arXiv.2110.05069.
- [217] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. “Efficient Training of Audio Transformers With Patchout”. In: *Interspeech 2022*. Sept. 2022, pp. 2753–2757. doi: 10.21437/Interspeech.2022-227.
- [218] Thodoris Kouzelis, Grigoris Bastas, Athanasios Katsamanis, and Alexandros Potamianos. *Efficient Audio Captioning Transformer with Patchout and Text Guidance*. Apr. 2023. doi: 10.48550/arXiv.2304.02916.
- [219] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, et al. “AudioGen: Textually Guided Audio Generation”. In: *ICLR. OpenReview.net*, 2022. doi: 10.48550/arxiv.2209.15352.
- [220] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. “Dense-Captioning Events in Videos”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 706–715. doi: 10.1109/ICCV.2017.83.
- [221] Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee. “Understanding Sounds, Missing the Questions: The Challenge of Object Hallucination in Large Audio-Language Models”. In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 4144–4148. doi: 10.21437/Interspeech.2024-1076.
- [222] Jan Kulik, Bartłomiej Zgórzyński, Juliusz Kruk, Ivan Ryzhankow, Anna Ples, et al. “Take It for Granted: Improving Language-Based Audio Retrieval with Large Language Models”. In: *IEEE AASP Challenge Detection and Classification of Acoustic Scenes and Events, Tokyo, Japan, Technical Report* (2024).
- [223] Sonal Kumar, Sreyan Ghosh, Yueqian Lin, Yiran Chen, Ramani Duraiswami, et al. *PolyAudio: Advancing Multi-Audio Analysis & Reasoning in Large Audio Language Models*. 2025.
- [224] Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, et al. “MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. 2026, pp. 22688–22697.

-
- [225] Nikita Kuzmin and Alexander Dyakonov. “Automated Audio Captioning”. In: *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep* (2020).
- [226] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, et al. “Efficient Memory Management for Large Language Model Serving with PagedAttention”. In: *Proceedings of the 29th Symposium on Operating Systems Principles*. SOSP ’23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 611–626. doi: 10.1145/3600006.3613165.
- [227] Étienne Labbé, Thomas Pellegrini, and J. Pinquier. *Killing Two Birds with One Stone: Can an Audio Captioning System Also Be Used for Audio-Text Retrieval?* 2023. doi: 10.48550/arXiv.2308.15090.
- [228] Étienne Labbé, Thomas Pellegrini, and Julien Pinquier. “CoNeTTE: An Efficient Audio Captioning System Leveraging Multiple Datasets With Task Embedding”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 3785–3794. doi: 10.1109/TASLP.2024.3430813.
- [229] Yongquan Lai, Jinsong Pan, and Buxian Chen. “A Resnet-Based Clip Text-to-Audio Retrieval System for Dcase Challenge 2022 Task 6b”. In: *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Tech. Rep* (2022).
- [230] Sabine Langlois, Clara Suied, Thierry Lageat, and Aude Charbonneau. “Cross Cultural Study of Auditory Warnings”. In: *Proceedings of the 14th International Conference on Auditory Display (ICAD2008)*, Paris, France. 2008.
- [231] Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, et al. *Sparks of Large Audio Models: A Survey and Outlook*. Sept. 2023. doi: 10.48550/arXiv.2308.12792.
- [232] Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. “Multimodal Reasoning with Multimodal Knowledge Graph”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 10767–10782. doi: 10.18653/v1/2024.acl-long.579.
- [233] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, et al. “Deduplicating Training Data Makes Language Models Better”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8424–8445. doi: 10.18653/v1/2022.acl-long.577.
- [234] Sang-gil Lee, Zhifeng Kong, Arushi Goel, Sungwon Kim, Rafael Valle, et al. “ETTA: Elucidating the Design Space of Text-to-Audio Models”. In: *Forty-Second International Conference on Machine Learning*. 2025.
- [235] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. “Tvr: A Large-Scale Dataset for Video-Subtitle Moment Retrieval”. In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Vol. 12366. Lecture Notes in Computer Science. Springer, 2020, pp. 447–463. doi: 10.1007/978-3-030-58589-1_27.
- [236] Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, et al. *ViT-Lens: Towards Omni-modal Representations*. Mar. 2024. doi: 10.48550/arXiv.2311.16081.
- [237] James W. Lewis, Frederic L. Wightman, Julie A. Brefczynski, Raymond E. Phinney, Jeffrey R. Binder, et al. “Human Brain Regions Involved in Recognizing Environmental Sounds”. In: *Cerebral Cortex* 14.9 (Sept. 2004), pp. 1008–1021. doi: 10.1093/cercor/bhh061.
- [238] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, et al. *BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension*. Oct. 2019. doi: 10.48550/arXiv.1910.13461.
- [239] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.

- [240] Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, et al. *Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering*. May 2025. doi: 10.48550/arXiv.2503.11197.
- [241] Guangyao Li, Yixin Xu, and Di Hu. “Multi-Scale Attention for Audio Question Answering”. In: *Proceedings of Interspeech 2023*. Aug. 2023, pp. 3442–3446. doi: 10.21437/interspeech.2023-1606.
- [242] Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, et al. “See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation”. In: *Conference on Robot Learning*. PMLR, 2023, pp. 1368–1378.
- [243] J Li. “Towards Robust Large-Scale Audio/Visual Learning”. PhD thesis. Pittsburgh: Carnegie Mellon University, 2023.
- [244] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. “BLIP-2: Bootstrapping Language-Image Pre-Training With Frozen Image Encoders and Large Language Models”. In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 19730–19742.
- [245] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. “BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation”. In: *International Conference on Machine Learning*. PMLR, 2022, pp. 12888–12900.
- [246] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, et al. *VideoChat: Chat-Centric Video Understanding*. Jan. 2024. doi: 10.48550/arXiv.2305.06355.
- [247] Yizhi Li, Ge Zhang, Yi Ma, Ruibin Yuan, Kang Zhu, et al. *OmniBench: Towards The Future of Universal Omni-Language Models*. 2024. doi: 10.48550/arXiv.2409.15272.
- [248] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, et al. “From System 1 to System 2: A Survey of Reasoning Large Language Models”. In: *CoRR abs/2502.17419* (Feb. 2025). doi: 10.48550/arXiv.2502.17419.
- [249] Jinhua Liang, Xubo Liu, Wenwu Wang, Mark D. Plumbley, Huy Phan, et al. “Acoustic Prompt Tuning: Empowering Large Language Models With Audition Capabilities”. In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 949–961. doi: 10.1109/TASLPRO.2025.3533375.
- [250] Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, et al. *BATON: Aligning Text-to-Audio Model With Human Preference Feedback*. Feb. 2024. doi: 10.48550/arXiv.2402.00744.
- [251] Chin-Yew Lin. “Rouge: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81.
- [252] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, et al. “Egocentric Video-Language Pretraining”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 7575–7586. doi: 10.48550/arxiv.2206.01670.
- [253] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, et al. “Microsoft COCO: Common Objects in Context”. In: *European Conference on Computer Vision*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Vol. 8693. Springer, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.
- [254] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. “Clotho-AQA: A Crowdsourced Dataset for Audio Question Answering”. In: *EUSIPCO*. IEEE, 2022, pp. 1140–1144.
- [255] Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, et al. *Voxtral*. July 2025. doi: 10.48550/arXiv.2507.13264.

-
- [256] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, et al. “AudioLDM: Text-to-Audio Generation With Latent Diffusion Models”. In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 21450–21474.
 - [257] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, et al. *AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining*. May 2024. doi: 10.48550/arXiv.2308.05734.
 - [258] Huadai Liu, Rongjie Huang, Yang Liu, Hengyuan Cao, Jiale Wang, et al. “AudioLCM: Efficient and High-Quality Text-to-Audio Generation With Minimal Inference Steps”. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. Melbourne VIC Australia: ACM, Oct. 2024, pp. 7008–7017. doi: 10.1145/3664647.3681072.
 - [259] Siqi Liu, Zhenhai Zhu, Ning Ye, S. Guadarrama, and K. Murphy. “Improved Image Captioning via Policy Gradient Optimization of SPIDER”. In: *2017 IEEE International Conference on Computer Vision (ICCV) (2016)*, pp. 873–881. doi: 10.1109/ICCV.2017.100.
 - [260] Weizhuo Liu and Zhe Gao. “Improving Automated Audio Captioning with LLM Decoder and Beats Audio Encoder”. In: *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*. 2024, pp. 252–259.
 - [261] Xubo Liu, Qiushi Huang, Xinhao Mei, Tom Ko, H Lilian Tang, et al. “Cl4AC: A Contrastive Loss for Audio Captioning”. In: *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, Technical Report (2021)*.
 - [262] Xubo Liu, Qiushi Huang, Xinhao Mei, Haohe Liu, Qiuqiang Kong, et al. “Visually-Aware Audio Captioning With Adaptive Audio-Visual Attention”. In: *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*. Ed. by Naomi Harte, Julie Carson-Berndsen, and Gareth Jones. ISCA, 2023, pp. 2838–2842. doi: 10.21437/INTERSPEECH.2023-914.
 - [263] Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, et al. *Separate Anything You Describe*. Dec. 2024. doi: 10.48550/arXiv.2308.05037.
 - [264] Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, et al. “Separate What You Describe: Language-Queried Audio Source Separation”. In: *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*. Ed. by Hanseok Ko and John H. L. Hansen. ISCA, 2022, pp. 1801–1805. doi: 10.21437/INTERSPEECH.2022-10894.
 - [265] Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, et al. *WavJourney: Compositional Audio Creation With Large Language Models*. Nov. 2023. doi: 10.48550/arXiv.2307.14335.
 - [266] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, et al. *RoBERTa: A Robustly Optimized Bert Pretraining Approach*. 2019.
 - [267] Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. “Fully Automatic Semantic Mt Evaluation”. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 243–252.
 - [268] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. 2019.
 - [269] Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, et al. *DeSTA2.5-Audio: Toward General-Purpose Large Audio Language Model With Self-Generated Cross-Modal Alignment*. July 2025. doi: 10.48550/arXiv.2507.02768.
 - [270] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, et al. “Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 26439–26455.

- [271] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, et al. “Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 2507–2521. doi: 10.48550/arXiv.2209.09513.
- [272] Sasha Luccioni, Yacine Jernite, and Emma Strubell. “Power Hungry Processing: Watts Driving the Cost of AI Deployment?” In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024, pp. 85–99. doi: 10.1145/3630106.3658542.
- [273] Robin Lund. *Do Not Use ‘Freetousesounds.Com’* | Robin Lund. Feb. 2023.
- [274] Kaiyi Luo, Xulong Zhang, Jianzong Wang, Huaxiong Li, Ning Cheng, et al. “Contrastive Latent Space Reconstruction Learning for Audio-Text Retrieval”. In: *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. Atlanta, GA, USA: IEEE, Nov. 2023, pp. 913–917. doi: 10.1109/ICTAI59109.2023.00137.
- [275] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, et al. *Valley: Video Assistant With Large Language Model Enhanced Ability*. 2023.
- [276] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, et al. “DETRs Beat YOLOs on Real-time Object Detection”. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 16965–16974. doi: 10.1109/CVPR52733.2024.01605.
- [277] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, et al. *Macaw-LLM: Multi-Modal Language Modeling With Image, Audio, Video, and Text Integration*. June 2023. doi: 10.48550/arXiv.2306.09093.
- [278] Hao Ma, Zhiyuan Peng, Xu Li, Mingjie Shao, Xixin Wu, et al. “CLAPSep: Leveraging Contrastive Pre-Trained Model for Multi-Modal Query-Conditioned Target Sound Extraction”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 4945–4960. doi: 10.1109/TASLP.2024.3497586.
- [279] Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. *Audio-CoT: Exploring Chain-of-Thought Reasoning in Large Audio Language Model*. Jan. 2025. doi: 10.48550/arXiv.2501.07246.
- [280] Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, et al. “MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix”. In: *Advances in Neural Information Processing Systems*. Vol. 38. 2026. doi: 10.48550/arXiv.2505.13032.
- [281] Kevin Madigan. *Mid-Year Review: AI Copyright Case Developments in 2025*. Aug. 2025.
- [282] R Mahfuz, Y Guo, and E Visser. *Resource-Efficient Reference-Free Evaluation of Audio Captions*. 2024. doi: 10.48550/ARXIV.2409.08489.
- [283] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, et al. “Tango 2: Aligning Diffusion-Based Text-to-Audio Generations Through Direct Preference Optimization”. In: *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*. Ed. by Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, et al. ACM, 2024, pp. 564–572. doi: 10.1145/3664647.3681688.
- [284] Potsawee Manakul, Guangzhi Sun, Warit Sirichotedumrong, Kasima Tharnpipitchai, and Kunat Pipatanakul. *Enhancing Low-Resource Language and Instruction Following Capabilities of Audio Language Models*. May 2025. doi: 10.48550/arXiv.2409.10999.
- [285] Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, et al. “TAVGBench: Benchmarking Text to Audible-Video Generation”. In: *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*. Ed. by Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, et al. ACM, 2024, pp. 6607–6616. doi: 10.1145/3664647.3680612.
- [286] Marcel and Libby. *Free to Use Sounds*. <https://www.freetousesounds.com/>.

-
- [287] Christian Marinoni, Riccardo Fosco Gramaccioni, Changan Chen, and Danilo Comminiello. *L3DAS23 - 2023 IEEE ICASSP Grand Challenge*.
- [288] Irene Martin-Morato and Annamaria Mesaros. "What Is the Ground Truth? Reliability of Multi-Annotator Data for Audio Tagging". In: *2021 29th European Signal Processing Conference (EUSIPCO)*. Dublin, Ireland: IEEE, Aug. 2021, pp. 76–80. doi: 10.23919/EUSIPCO54536.2021.9616087.
- [289] Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. "A Review of Deep Learning Techniques for Speech Processing". In: *Information Fusion* 99 (Nov. 2023), p. 101869. doi: 10.1016/j.inffus.2023.101869.
- [290] Xinhao Mei, Qiushi Huang, Xubo Liu, Gengyun Chen, Jingqian Wu, et al. *An Encoder-Decoder Based Audio Captioning System With Transfer and Reinforcement Learning*. Aug. 2021. doi: 10.48550/arXiv.2108.02752.
- [291] Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. *Audio Captioning Transformer*. July 2021. doi: 10.48550/arXiv.2107.09817.
- [292] Xinhao Mei, Xubo Liu, Haohe Liu, Jianyuan Sun, Mark D Plumbley, et al. "Language-Based Audio Retrieval with Pre-Trained Models". In: *DCASE 2022 Challenge Technical Report*. 2022.
- [293] Xinhao Mei, Xubo Liu, Mark D Plumbley, and Wenwu Wang. "Automated Audio Captioning: An Overview of Recent Progress and New Challenges". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2022.1 (Oct. 2022), p. 26. doi: 10.1186/s13636-022-00259-2.
- [294] Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D. Plumbley, and Wenwu Wang. *Diverse Audio Captioning via Adversarial Training*. Mar. 2022. doi: 10.48550/arXiv.2110.06691.
- [295] Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D. Plumbley, and Wenwu Wang. "On Metric Learning for Audio-Text Cross-Modal Retrieval". In: *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*. Ed. by Hanseok Ko and John H. L. Hansen. ISCA, 2022, pp. 4142–4146. doi: 10.21437/INTERSPEECH.2022-11115.
- [296] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, et al. "WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 3339–3354. doi: 10.1109/TASLP.2024.3419446.
- [297] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. "Metrics for Polyphonic Sound Event Detection". In: *Applied Sciences* 6.162 (2016). doi: 10.3390/app6060162.
- [298] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. "TUT Database for Acoustic Scene Classification and Sound Event Detection". In: *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, Aug. 2016, pp. 1128–1132. doi: 10.1109/eusipco.2016.7760424.
- [299] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. *TUT Sound Events 2017, Development Dataset*. 2017. doi: 10.5281/zenodo.814831.
- [300] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, and Mark D Plumbley. "Sound Event Detection: A Tutorial". In: *IEEE Signal Processing Magazine* 38.5 (2021), pp. 67–83. doi: 10.1109/MSP.2021.3090678.
- [301] Meta. *The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation*. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>. May 2025.
- [302] Zhongqi Miao, Benjamin Elizalde, Soham Deshmukh, Justin Kitzes, Huaming Wang, et al. "Zero-Shot Transfer for Wildlife Bioacoustics Detection". Aug. 2023.
- [303] Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, et al. *Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-Loras*. Mar. 2025. doi: 10.48550/arXiv.2503.01743.

- [304] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, et al. "HowTo100m: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips". In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 2630–2640. doi: 10.1109/ICCV.2019.00272.
- [305] George A Miller. "WordNet: A Lexical Database for English". In: *Communications of the ACM* 38.11 (Nov. 1995), pp. 39–41. doi: 10.1145/219717.219748.
- [306] Mistral Team. *Introducing Mistral 3 | Mistral AI*. <https://mistral.ai/news/mistral-3>. Dec. 2025.
- [307] Mistral Team. *Mistral Small 3 | Mistral AI*. <https://mistral.ai/news/mistral-small-3>.
- [308] Anton Mitrofanov, Sergey Novoselov, Tatiana Prisyach, Vladislav Marchevskiy, Arseniy Karelin, et al. "Cryfish: On Deep Audio Analysis with Large Language Models". In: *Proceedings of Interspeech 2025*. 2025, pp. 3249–3253. doi: 10.21437/Interspeech.2025-2109.
- [309] S Mo, J Shi, and Y Tian. *DiffAVA: Personalized Text-to-Audio Generation with Visual Alignment*. 2023. doi: 10.48550/ARXIV.2305.12903.
- [310] Md Mohaimenuzzaman, Christoph Bergmeir, Ian West, and Bernd Meyer. "Environmental Sound Classification on the Edge: A Pipeline for Deep Acoustic Networks on Extremely Resource-Constrained Devices". In: *Pattern Recognition* 133 (2023), p. 109025. doi: 10.1016/j.patcog.2022.109025.
- [311] Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. "The Parrot Dilemma: Human-Labeled vs. LLM-augmented Data in Classification Tasks". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 179–192. doi: 10.18653/v1/2024.eacl-short.17.
- [312] Christine Moon, Hugo Lagercrantz, and Patricia K Kuhl. "Language Experienced in Utero Affects Vowel Perception after Birth: A Two-Country Study". In: *Acta Paediatrica* 102.2 (2013), pp. 156–160. doi: 10.1111/apa.12098.
- [313] Irene Martin Morato and Annamaria Mesaros. "Diversity and Bias in Audio Captioning Datasets". In: *Detection and Classification of Acoustic Scenes and Events*. 2021, pp. 90–94.
- [314] Youssef Mroueh, Nicolas Dupuis, Brian Belgodere, Apoorva Nitsure, Mattia Rigotti, et al. *Revisiting Group Relative Policy Optimization: Insights into on-Policy and Off-Policy Training*. 2025. doi: 10.48550/arXiv.2505.22257.
- [315] Siyuan Mu and Sen Lin. *A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications*. 2025. doi: 10.48550/arXiv.2503.07137.
- [316] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, et al. *S1: Simple Test-Time Scaling*. Mar. 2025. doi: 10.48550/arXiv.2501.19393.
- [317] Nicolas Müller, Pavel Czepin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. "Does Audio Deepfake Detection Generalize?" In: *Proc. Interspeech 2022*. 2022, pp. 2783–2787. doi: 10.21437/Interspeech.2022-108.
- [318] Hokuto Munakata, Takehiro Imamura, Taichi Nishimura, and Tatsuya Komatsu. *CASTELLA: Long Audio Dataset with Captions and Temporal Boundaries*. Nov. 2025. doi: 10.48550/arXiv.2511.15131.
- [319] Hokuto Munakata, Taichi Nishimura, Shota Nakada, and Tatsuya Komatsu. *Language-Based Audio Moment Retrieval*. 2024. doi: 10.48550/arXiv.2409.15672.
- [320] Mihai Nadaş, Laura Dioşan, and Andreea Tomescu. "Synthetic Data Generation Using Large Language Models: Advances in Text and Code". In: *IEEE Access* 13 (2025), pp. 134615–134633. doi: 10.1109/ACCESS.2025.3589503.

-
- [321] Indrani Nag. *Systematic Literature Review of Deep Visual and Audio Captioning*. Tampere, Finland: Tampere University, 2021.
- [322] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, et al. "Learning Audio-Video Modalities From Image Captions". In: *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*. Ed. by Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner. Vol. 13674. Lecture Notes in Computer Science. Springer, 2022, pp. 407–426. doi: 10.1007/978-3-031-19781-9_24.
- [323] Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. "Are LLMs Better Than Reported? Detecting Label Errors and Mitigating Their Effect on Model Performance". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025, pp. 26770–26797. doi: 10.18653/v1/2025.emnlp-main.1360.
- [324] Taishi Nakamura, Satoki Ishikawa, Masaki Kawamura, Takumi Okamoto, Daisuke Nohara, et al. *Optimal Sparsity of Mixture-of-Experts Language Models for Reasoning Tasks*. Sept. 2025. doi: 10.48550/arXiv.2508.18672.
- [325] Chaitanya Prasad Narisetty, Tomoki Hayashi, Ryunosuke Ishizaki, Shinji Watanabe, and Kazuya Takeda. "Leveraging State-of-the-Art Asr Techniques to Audio Captioning." In: *Detection and Classification of Acoustic Scenes and Events*. 2021, pp. 160–164.
- [326] Arbi Haza Nasution and Aytuğ Onan. "ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks". In: *IEEE Access* 12 (2024), pp. 71876–71900. doi: 10.1109/ACCESS.2024.3402809.
- [327] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, et al. *A Comprehensive Overview of Large Language Models*. Oct. 2024. doi: 10.48550/arXiv.2307.06435.
- [328] Jiquan Ngiam, Daiyi Peng, Vijay Vasudevan, Simon Kornblith, Quoc V. Le, et al. *Domain Adaptive Transfer Learning with Specialist Models*. Dec. 2018. doi: 10.48550/arXiv.1811.07056.
- [329] Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, et al. "SeaLLMs - Large Language Models for Southeast Asia". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Ed. by Yixin Cao, Yang Feng, and Deyi Xiong. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 294–304. doi: 10.18653/v1/2024.acl-demos.28.
- [330] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. "BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation". In: *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*. IEEE, 2021, pp. 1–8. doi: 10.1109/IJCNN52387.2021.9534474.
- [331] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, Masahiro Yasuda, et al. "M2D-CLAP: Masked Modeling Duo Meets Clap for Learning General-Purpose Audio-Language Representation". In: *Proc. Interspeech 2024*. June 2024, pp. 57–61. doi: 10.48550/arXiv.2406.02032.
- [332] Evgenii Nikishin, Pavel Izmailov, Ben Athiwaratkun, Dmitrii Podoprikin, Timur Garipov, et al. "Improving Stability in Deep Reinforcement Learning With Weight Averaging". In: *Uncertainty in Artificial Intelligence Workshop*. 2018.
- [333] Taichi Nishimura, Shota Nakada, and Masayoshi Kondo. *On the Audio Hallucinations in Large Audio-Video Language Models*. 2024. doi: 10.48550/ARXIV.2401.09774.
- [334] Kari A. Noriy, Xiaosong Yang, Marcin Budka, and Jian Jun Zhang. *CLARA: Multilingual Contrastive Learning for Audio Representation Acquisition*. Nov. 2023. doi: 10.48550/arXiv.2310.11830.
- [335] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks". In: *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021. doi: 10.48550/arXiv.2103.14749.

- [336] Eduardo C. Nunes. *Anomalous Sound Detection with Machine Learning: A Systematic Review*. Feb. 2021. doi: 10.48550/arXiv.2102.07820.
- [337] Nvidia, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, et al. *Cosmos World Foundation Model Platform for Physical AI*. July 2025. doi: 10.48550/arXiv.2501.03575.
- [338] ODEON Cinematic Sound Effects | Vfx Motion - Paramount Motion. <https://www.paramountmotion.com/odeon-sound-effects>.
- [339] Y Oh, Y Han, and K Lee. “Distance Sampling-based Paraphraser Leveraging ChatGPT for Text Data Manipulation”. In: *Proceedings of the 47th International ACM SIGIR ...* (July 2024), pp. 2363–2367. doi: 10.1145/3626772.3657976.
- [340] Yuki Okamoto, Kanta Shimonishi, Keisuke Imoto, Kota Dohi, Shota Horiguchi, et al. “CAPTDURE: Captioned Sound Dataset of Single Sources”. In: *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*. Ed. by Naomi Harte, Julie Carson-Berndsen, and Gareth Jones. ISCA, 2023, pp. 1683–1687. doi: 10.21437/INTERSPEECH.2023-577.
- [341] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, et al. *Olmo 3*. 2025. doi: 10.48550/arXiv.2512.13961.
- [342] Michel Olvera, Changhong Wang, Paraskevas Stamatiadis, Gaël Richard, and Slim ESSID. “iKnow-Audio: Integrating Knowledge Graphs with Audio-Language Models”. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 34683–34700. doi: 10.18653/v1/2025.emnlp-main.1759.
- [343] Andreea-Maria Oncescu, João F. Henriques, and A. Sophia Koepke. *Dissecting Temporal Understanding in Text-to-Audio Retrieval*. Sept. 2024. doi: 10.48550/arXiv.2409.00851.
- [344] Andreea-Maria Oncescu, João F. Henriques, Andrew Zisserman, Samuel Albanie, and A. Sophia Koepke. “A Sound Approach: Using Large Language Models to Generate Audio Descriptions for Egocentric Text-Audio Retrieval”. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 7300–7304. doi: 10.1109/ICASSP48485.2024.10448486.
- [345] Kenneth Ooi, Zhen-Ting Ong, Karn N. Watcharasupat, Bhan Lam, J. Hong, et al. “ARAUS: A Large-Scale Dataset and Baseline Models of Affective Responses to Augmented Urban Soundscapes”. In: *IEEE Transactions on Affective Computing* 15.1 (Jan. 2024), pp. 105–120. doi: 10.1109/TAFFC.2023.3247914.
- [346] OpenAI. *Function Calling and Other Api Updates*. <https://openai.com/index/function-calling-and-other-api-updates/>. July 2023.
- [347] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, et al. *GPT-4o System Card*. Oct. 2024. doi: 10.48550/arXiv.2410.21276.
- [348] OpenAI Team. *GPT-5 System Card*. Aug. 2025.
- [349] OpenAI Team. *Introducing OpenAI O1*. <https://openai.com/o1/>. Sept. 2024.
- [350] Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, et al. “Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs”. In: *2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Hybrid, Miami, United States of America, Nov 12 2024-Nov 16 2024*. Association for Computational Linguistics (ACL), 2024, pp. 9340–9366.
- [351] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, et al. “Training Language Models to Follow Instructions with Human Feedback”. In: *NeurIPS* 35 (2022), pp. 27730–27744. doi: 10.48550/ARXIV.2203.02155.

-
- [352] Francesco Paissan and Elisabetta Farella. “tinyCLAP: Distilling Constrastive Language-Audio Pretrained Models”. In: *Interspeech 2024*. Sept. 2024, pp. 1685–1689. doi: 10.21437/Interspeech.2024-193.
 - [353] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, et al. “Unifying Large Language Models and Knowledge Graphs: A Roadmap”. In: *IEEE Transactions on Knowledge and Data Engineering* 36.7 (July 2024), pp. 3580–3599. doi: 10.1109/TKDE.2024.3352100.
 - [354] Sharnil Pandya and Hemant Ghayvat. “Ambient Acoustic Event Assistive Framework for Identification, Detection, and Recognition of Unknown Acoustic Events of a Residence”. In: *Advanced Engineering Informatics* 47 (2021), p. 101238. doi: 10.1016/j.aei.2020.101238.
 - [355] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “Bleu: A Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. doi: 10.3115/1073083.1073135.
 - [356] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, et al. “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition”. In: *Interspeech 2019*. Sept. 2019, pp. 2613–2617. doi: 10.21437/Interspeech.2019-2680.
 - [357] Hyeoncheol Park, Jiyoung Seo, Jaewon Mun, Hogun Park, Wonmin Byeon, et al. *M3-KG-RAG: Multi-hop Multimodal Knowledge Graph-enhanced Retrieval-Augmented Generation*. Dec. 2025. doi: 10.48550/arXiv.2512.20136.
 - [358] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, et al. “Automatic Differentiation in Pytorch”. In: *Nips-w*. 2017.
 - [359] Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, et al. “The Berkeley Function Calling Leaderboard (Bfcl): From Tool Use to Agentic Evaluation of Large Language Models”. In: *ICML*. 2025.
 - [360] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, et al. *Carbon Emissions and Large Neural Network Training*. Apr. 2021. doi: 10.48550/arXiv.2104.10350.
 - [361] Thomas Pellegrini. “Language-Based Audio Retrieval With Textual Embeddings of Tag Names”. In: *Workshop on Detection and Classification of Acoustic Scenes and Events (Workshop DCASE 2022)*. 2022.
 - [362] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. *Instruction Tuning with GPT-4*. Apr. 2023. doi: 10.48550/arXiv.2304.03277.
 - [363] Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. *Introducing Gemini 2.0: Our New Ai Model for the Agentic Era*. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>. Dec. 2024.
 - [364] Karol J. Piczak. “ESC: Dataset for Environmental Sound Classification”. In: *Proceedings of the 23rd Annual ACM Conference on Multimedia*. Brisbane, Australia: ACM Press, Oct. 2015, pp. 1015–1018. doi: 10.1145/2733373.2806390.
 - [365] Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. “End-to-End Speech Recognition: A Survey”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 325–351. doi: 10.1109/TASLP.2023.3328283.
 - [366] Douglas Price. *Pro Sound Effects - Bring Your Ideas to Life With Sound*. <https://www.prosoundeffects.com/>.
 - [367] Paul Primus, Khaled Koutini, and Gerhard Widmer. *Advancing Natural-Language Based Audio Retrieval with PaSST and Large Audio-Caption Data Sets*. Aug. 2023. doi: 10.48550/arXiv.2308.04258.
 - [368] Paul Primus, Florian Schmid, and Gerhard Widmer. *TACOS: Temporally-Aligned Audio Captions for Language-Audio Pretraining*. May 2025. doi: 10.48550/arXiv.2505.07609.

- [369] Paul Primus and Gerhard Widmer. *Improving Natural-Language-based Audio Retrieval with Transfer Learning and Audio & Text Augmentations*. Oct. 2022. doi: 10.48550/arXiv.2208.11460.
- [370] Prodigy · Prodigy · an Annotation Tool for Ai, Machine Learning & Nlp. <https://prodi.gy/>.
- [371] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. *Qwen2.5 Technical Report*. Jan. 2025. doi: 10.48550/arXiv.2412.15115.
- [372] Qwen Team. *QwQ: Reflect Deeply on the Boundaries of the Unknown*. Nov. 2024.
- [373] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *ICML*. PmlR, 2021, pp. 8748–8763. doi: 10.48550/arxiv.2103.00020.
- [374] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. Dec. 2022. doi: 10.48550/arXiv.2212.04356.
- [375] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, et al. “Language Models Are Unsupervised Multitask Learners”. In: *OpenAI blog* 1.8 (2019), p. 9.
- [376] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, et al. *Exploring the Limits of Transfer Learning With a Unified Text-to-Text Transformer*. Sept. 2023. doi: 10.48550/arXiv.1910.10683.
- [377] Ksheeraja Raghavan, Samiran Gode, Ankit Shah, Surabhi Raghavan, Wolfram Burgard, et al. “Did You Hear That? Introducing AADG: A Framework for Generating Benchmark Data in Audio Anomaly Detection”. In: *ICLR 2025 Workshop SynthData*. 2025.
- [378] Robi Rahman and David Owen. *The Size of Datasets Used to Train Language Models Doubles Approximately Every Six Months*. 2024.
- [379] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. “Zero: Memory Optimizations Toward Training Trillion Parameter Models”. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC ’20. Atlanta, Georgia: IEEE Press, Nov. 2020, pp. 1–16.
- [380] Muhammad Mamunur Rashid, Guiqing Li, and Chengrui Du. “Nonspeech7k Dataset: Classification and Analysis of Human Non-speech Sound”. In: *IET Signal Processing* 17.6 (June 2023), e12233. doi: 10.1049/si12.12233.
- [381] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/d19-1410.
- [382] Paul Robert, Robert Zatorre, Akanksha Gupta, Julien Sein, Jean-Luc Anton, et al. *Auditory Hemispheric Asymmetry as a Specialization for Actions and Objects*. Apr. 2023. doi: 10.1101/2023.04.19.537361.
- [383] D Robinson, A Robinson, and L Akrapongpisak. “Transferable Models for Bioacoustics with Human Language Supervision”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Apr. 2024), pp. 1316–1320. doi: 10.1109/icassp48485.2024.10447250.
- [384] Dominik Roblek, Kevin Kilgour, Matt Sharifi, and Mauricio Zuluaga. “Fr’echet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms”. In: *Proc. Interspeech*. 2019, pp. 2350–2354.
- [385] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Sikandar Amin, et al. “Coherent Multi-Sentence Video Description With Variable Level of Detail”. In: *Pattern Recognition: 36th German Conference, GCPR 2014, M{u}nster, Germany, September 2-5, 2014, Proceedings 36*. Vol. 8753. Springer, 2014, pp. 184–195. doi: 10.1007/978-3-319-11752-2_15.

-
- [386] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. *A Dataset for Movie Description*. Jan. 2015. doi: 10.48550/arXiv.1501.02530.
- [387] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, et al. “Movie Description”. In: *International Journal of Computer Vision* 123.1 (May 2017), pp. 94–120. doi: 10.1007/s11263-016-0987-1.
- [388] Robin San Roman, Yossi Adi, Antoine Deleforge, Romain Serizel, Gabriel Synnaeve, et al. “From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion”. In: *Thirty-Seventh Conference on Neural Information Processing Systems (2023)*. Nov. 2023. doi: 10.48550/arXiv.2308.02560.
- [389] Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, et al. *Omni-R1: Do You Really Need Audio to Fine-Tune Your Audio Llm?* June 2025. doi: 10.48550/arXiv.2505.09439.
- [390] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. “SDR – Half-baked or Well Done?” In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2019, pp. 626–630. doi: 10.1109/ICASSP.2019.8683855.
- [391] Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, et al. “A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Association for Computational Linguistics, 2024, pp. 11709–11724. doi: 10.18653/v1/2024.FINDINGS-EMNLP.685.
- [392] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, et al. “NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for Each Benchmark”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10776–10787. doi: 10.18653/v1/2023.findings-emnlp.722.
- [393] S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, et al. “MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark”. In: *The Thirteenth International Conference on Learning Representations*. 2025.
- [394] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. “A Dataset and Taxonomy for Urban Sound Research”. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. MM ’14. New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 1041–1044. doi: 10.1145/2647868.2655045.
- [395] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. “Scaper: A Library for Soundscape Synthesis and Augmentation”. In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, Oct. 2017, pp. 344–348. doi: 10.1109/waspaa.2017.8170052.
- [396] Leonard Salewski, Stefan Fauth, A. Sophia Koepke, and Zeynep Akata. *Zero-Shot Audio Captioning with Audio-Language Model Guidance and Audio Context Keywords*. Nov. 2023. doi: 10.48550/arXiv.2311.08396.
- [397] Arunodhayan Sampathkumar and Danny Kowerko. “Automated Audio Captioning”. In: *Proceedings of Conference on Detection and Classification of Acoustic Scenes and Events*. 2020.
- [398] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, et al. *How2: A Large-Scale Dataset for Multimodal Language Understanding*. Dec. 2018. doi: 10.48550/arXiv.1811.00347.
- [399] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, et al. “Multitask Prompted Training Enables Zero-Shot Task Generalization”. In: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

- [400] Omar Sanseviero and Ian Ballantyne. *Introducing Gemma 3n: The Developer Guide - Google Developers Blog*. <https://developers.googleblog.com/en/introducing-gemma-3n-developer-guide/>. June 2025.
- [401] George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, et al. *Granite-Speech: Open-Source Speech-Aware LLMs With Strong English ASR Capabilities*. May 2025. doi: 10.48550/arXiv.2505.08699.
- [402] Edward Sapir. "The Status of Linguistics as a Science". In: *Language* 5.4 (1929), pp. 207–214. doi: 10.2307/409588.
- [403] Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, et al. *Grounded Reinforcement Learning for Visual Reasoning*. May 2025. doi: 10.48550/arXiv.2505.23678.
- [404] Joseph Sardin. *Free Sound Effects Library - Bigsoundbank*. <https://bigsoundbank.com/>.
- [405] Timothy Schaumlöffel, Martina G Vilas, and Gemma Roig. "PEACS: Prefix Encoding for Auditory Caption Synthesis". In: *Proceedings of the Conference on Detection and Classification of Acoustic Scenes and Events Challenge*. 2023, pp. 1–3.
- [406] Thibault Sellam, Dipanjan Das, and Ankur Parikh. "Bleurt: Learning Robust Metrics for Text Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7881–7892. doi: 10.18653/v1/2020.acl-main.704.
- [407] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models With Monolingual Data". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/V1/P16-1009.
- [408] Tal Shaharabany, Ariel Shaulov, and Lior Wolf. *Zero-Shot Audio Captioning via Audibility Guidance*. Sept. 2023. doi: 10.48550/arXiv.2309.03884.
- [409] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. 2024. doi: 10.48550/arXiv.2402.03300.
- [410] Lloyd S Shapley. "A Value for N-Person Games". In: *Contributions to the Theory of Games II*. Ed. by Harold W. Kuhn and Albert W. Tucker. Princeton: Princeton University Press, 1953, pp. 307–317.
- [411] Sandhya Sharma, Kazuhiko Sato, and Bishnu Prasad Gautam. "A Methodological Literature Review of Acoustic Wildlife Monitoring Using Artificial Intelligence Tools and Techniques". In: *Sustainability* 15.9 (2023), p. 7128. doi: 10.3390/su15097128.
- [412] Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, et al. *Fine-Grained Audible Video Description*. Mar. 2023. doi: 10.48550/arXiv.2303.15616.
- [413] Anna Shi and ShuangFeng First. *Audio Captioning With the Transformer*. Tech. rep. DCASE2020 Challenge Technical Report, 2020.
- [414] Yangyang Shi, Gael Le Lan, Varun Nagaraja, Zhaoheng Ni, Xinhao Mei, et al. *Enhance Audio Generation Controllability through Representation Similarity Regularization*. Sept. 2023. doi: 10.48550/arXiv.2309.08773.
- [415] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, et al. *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism*. Mar. 2020. doi: 10.48550/arXiv.1909.08053.
- [416] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. *Audio-Visual LLM for Video Understanding*. 2023. doi: 10.48550/ARXIV.2312.06720.
- [417] Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. "UniVAL: Unified Model for Image, Video, Audio and Language Tasks". In: *Transactions on Machine Learning Research* 2835-8856 (Dec. 2023). doi: 10.48550/arXiv.2307.16184.

-
- [418] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, et al. “AI Models Collapse When Trained on Recursively Generated Data”. In: *Nature* 631.8022 (July 2024), pp. 755–759. doi: 10.1038/s41586-024-07566-y.
- [419] Jaeheon Sim, Eungbeom Kim, and Kyogu Lee. “Label-Refined Sequential Training With Noisy Data for Automated Audio Captioning”. In: *Detection and Classification of Acoustic Scenes and Events*. 2023.
- [420] Joren Six and Marc Leman. “Panako: A Scalable Acoustic Fingerprinting System Handling Time-Scale and Pitch Modification”. In: *International Society for Music Information Retrieval, Proceedings*. Vol. Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR 2014). 2014.
- [421] SONNISS | Professional Sound Libraries (Royalty-Free). <https://sonniss.com>.
- [422] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. “Curriculum Learning: A Survey”. In: *International Journal of Computer Vision* 130.6 (June 2022), pp. 1526–1565. doi: 10.1007/s11263-022-01611-x.
- [423] Dilara Soyulu, Christopher Potts, and Omar Khattab. “Fine-Tuning and Prompt Optimization: Two Great Steps That Work Better Together”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 10696–10710. doi: 10.18653/v1/2024.emnlp-main.597.
- [424] Karen Sparck Jones. “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”. In: *Journal of Documentation* 28.1 (Jan. 1972), pp. 11–21. doi: 10.1108/eb026526.
- [425] Sreyashi Chatterjee. *Epidemic Sound Review: Is It Worth the Hype?* July 2024.
- [426] Stanley Smith Stevens, John Volkman, and Edwin Broomell Newman. “A Scale for the Measurement of the Psychological Magnitude Pitch”. In: *Journal of the Acoustical Society of America* 8.3 (1937), pp. 185–190. doi: 10.1121/1.1915893.
- [427] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbly. “Detection and Classification of Acoustic Scenes and Events”. In: *IEEE Transactions on Multimedia* 17.10 (Oct. 2015), pp. 1733–1746. doi: 10.1109/TMM.2015.2428998.
- [428] Emma Strubell, Ananya Ganesh, and Andrew McCallum. “Energy and Policy Considerations for Deep Learning in NLP”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3645–3650. doi: 10.18653/v1/P19-1355.
- [429] Bob L. Sturm. “The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval”. In: *Journal of New Music Research* 43.2 (Apr. 2014), pp. 147–172. doi: 10.1080/09298215.2014.894533.
- [430] Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, et al. *Crossing the Reward Bridge: Expanding RL with Verifiable Rewards Across Diverse Domains*. Apr. 2025. doi: 10.48550/arXiv.2503.23829.
- [431] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, et al. *PandaGPT: One Model to Instruction-Follow Them All*. May 2023. doi: 10.48550/arXiv.2305.16355.
- [432] Parthasaarathy Sudarsanam and Tuomas Virtanen. “Attention-Based Methods for Audio Question Answering”. In: *EUSIPCO. IEEE*, 2023, pp. 750–754. doi: 10.23919/EUSIPCO58844.2023.10289751.
- [433] Jianyuan Sun, Xubo Liu, Xinhao Mei, Mark D. Plumbly, Volkan Kilic, et al. *Automated Audio Captioning via Fusion of Low- and High- Dimensional Features*. Oct. 2022. doi: 10.48550/arXiv.2210.05037.
- [434] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, et al. “Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph”. In: *The Twelfth International Conference on Learning Representations*. 2024.

- [435] Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. “Auto-ACD: A Large-scale Dataset for Audio-Language Representation Learning”. In: *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*. Ed. by Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, et al. ACM, 2024, pp. 5025–5034. doi: 10.1145/3664647.3681472.
- [436] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, et al. *ERNIE 3.0: Large-Scale Knowledge Enhanced Pre-Training for Language Understanding and Generation*. July 2021. doi: 10.48550/arXiv.2107.02137.
- [437] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. Dec. 2014. doi: 10.48550/arXiv.1409.3215.
- [438] Richard Sutton. *The Bitter Lesson*. Mar. 2019.
- [439] Modan TAILLEUR, Junwon Lee, Mathieu Lagrange, Keunwoo Choi, Laurie M. Heller, et al. *Correlation of Fréchet Audio Distance With Human Perception of Environmental Audio Is Embedding Dependant*. 2024. doi: 10.48550/arXiv.2403.17508.
- [440] Daiki Takeuchi, Yuma Koizumi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. “Effects of Word-Frequency Based Pre- and Post- Processings for Audio Captioning”. In: *Workshop on Detection and Classification of Acoustic Scenes and Events*. 2020. doi: 10.48550/arXiv.2009.11436.
- [441] Daiki Takeuchi, Yasunori Ohishi, Daisuke Niizumi, Noboru Harada, and Kunio Kashino. *Audio Difference Captioning Utilizing Similarity-Discrepancy Disentanglement*. Aug. 2023. doi: 10.48550/arXiv.2308.11923.
- [442] Daiki Takeuchi, Yasunori Ohishi, Daisuke Niizumi, Noboru Harada, and Kunio Kashino. “Introducing Auxiliary Text Query-Modifier to Content-Based Audio Retrieval”. In: *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*. Ed. by Hanseok Ko and John H. L. Hansen. ISCA, 2022, pp. 4197–4201. doi: 10.21437/INTERSPEECH.2022-11428.
- [443] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, et al. “Let Me Speak Freely? A Study On The Impact Of Format Restrictions On Large Language Model Performance”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Miami, Florida, US: Association for Computational Linguistics, 2024, pp. 1218–1236. doi: 10.18653/v1/2024.emnlp-industry.91.
- [444] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, et al. “SALMONN: Towards Generic Hearing Abilities for Large Language Models”. In: *International Conference on Learning Representations*. OpenReview.net, 2023. doi: 10.48550/arXiv.2310.13289.
- [445] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. “Any-to-Any Generation via Composable Diffusion”. In: *Advances in Neural Information Processing Systems 36 (2023)*, pp. 16083–16099. doi: 10.48550/arXiv.2305.11846.
- [446] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, et al. *Alpaca: A Strong, Replicable Instruction-Following Model*. 2023.
- [447] Tiago Fernandes TAVARES, Fabio José Ayres, Zhepei Wang, and Paris Smaragdis. “On Class Separability Pitfalls In Audio-Text Contrastive Zero-Shot Learning”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Hyderabad, India: IEEE, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10888245.
- [448] *Tavily - the Web Access Layer for AI Agents*. <https://www.tavily.com/>.
- [449] GLM-4 5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, et al. *GLM-4.5: Agentic, Reasoning, and Coding (Arc) Foundation Models*. Aug. 2025. doi: 10.48550/arXiv.2508.06471.
- [450] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, et al. *Kimi K2: Open Agentic Intelligence*. July 2025. doi: 10.48550/arXiv.2507.20534.

-
- [451] Lyria Team, Antoine Caillon, Brian McWilliams, Cassie Tarakajian, Ian Simon, et al. *Live Music Models*. Nov. 2025. doi: 10.48550/arXiv.2508.04651.
- [452] RAPIDS Development Team. *Rapids: Libraries for End to End Gpu Data Science*. Manual. 2023.
- [453] Reka Team, Aitor Ormazabal, Che Zheng, Cyprien de Masson d’Autume, Dani Yogatama, et al. *Reka Core, Flash, and Edge: A Series of Powerful Multimodal Language Models*. Apr. 2024. doi: 10.48550/arXiv.2404.12387.
- [454] Xiaomi LLM-Core Team, Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, et al. *MiMo-Audio: Audio Language Models Are Few-Shot Learners*. Dec. 2025. doi: 10.48550/arXiv.2512.23808.
- [455] Ryan Teknum, Roger Jin, Jai Suphavadeeprasit, Dakota Mahan, Jeffrey Quesnelle, et al. *Hermes 4 Technical Report*. Sept. 2025. doi: 10.48550/arXiv.2508.18255.
- [456] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. “Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing”. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Vol. 12348. Cham: Springer International Publishing, 2020, pp. 436–454.
- [457] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. “Audio-Visual Event Localization in Unconstrained Videos”. In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11206. Lecture Notes in Computer Science. Springer, 2018, pp. 252–268. doi: 10.1007/978-3-030-01216-8_16.
- [458] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. “Learning From Between-Class Examples for Deep Sound Recognition”. In: *International Conference on Learning Representations*. arXiv, Feb. 2018. doi: 10.48550/arXiv.1711.10282.
- [459] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. *Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research*. Mar. 2015. doi: 10.48550/arXiv.1503.01070.
- [460] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. July 2023. doi: 10.48550/arXiv.2307.09288.
- [461] Andreas Triantafyllopoulos, Iosif Tsangko, Alexander Gebhard, Annamaria Mesaros, Tuomas Virtanen, et al. *Computer Audition: From Task-Specific Machine Learning to Foundation Models*. 2024. doi: 10.48550/arXiv.2407.15672.
- [462] Ivo Trowitzsch, Jalil Taghia, Youssef Kashef, and Klaus Obermayer. *The NIGENS General Sound Events Database*. 2019. doi: 10.48550/arxiv.1902.08314.
- [463] Shunsuke Tsubaki, Yohei Kawaguchi, Tomoya Nishida, Keisuke Imoto, Yuki Okamoto, et al. “Audio-Change Captioning to Explain Machine-Sound Anomalies”. In: *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*. Tampere, Finland, Sept. 2023, pp. 201–205.
- [464] Shunsuke Tsubaki, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, et al. *Refining Knowledge Transfer on Audio-Image Temporal Agreement for Audio-Text Cross Retrieval*. 2024. doi: 10.48550/arXiv.2403.10756.
- [465] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, et al. *Hear: Holistic Evaluation of Audio Representations*. May 2022. doi: 10.48550/arXiv.2203.03022.
- [466] Vishaal Udandaraao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H. S. Torr, et al. *“Zero-Shot” Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance*. Oct. 2024. doi: 10.48550/arXiv.2404.04125.
- [467] Frenk van Harreveld, Eric-Jan Wagenmakers, and Han L. J. van der Maas. “The Effects of Time Pressure on Chess Skill: An Investigation into Fast and Slow Processes Underlying Expert Performance”. In: *Psychological Research* 71.5 (Sept. 2007), pp. 591–597. doi: 10.1007/s00426-006-0076-0.

- [468] Christina M Vanden Bosch der Nederlanden, Joel S Snyder, and Erin E Hannon. "Children Use Object-Level Category Knowledge to Detect Changes in Complex Auditory Scenes". In: *Developmental Psychology* 52.11 (2016), p. 1867. doi: 10.1037/dev0000211.
- [469] Jurgen Vandendriessche, Nick Wouters, Bruno da Silva, Mimoun Lamrini, Mohamed Yassin Chkouri, et al. "Environmental Sound Recognition on Embedded Systems: From FPGAs to TPUs". In: *Electronics* 10.21 (2021), p. 2622. doi: 10.3390/electronics10212622.
- [470] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. "Attention Is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17*. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010. doi: 10.5555/3295222.3295349.
- [471] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. "CIDEr: Consensus-Based Image Description Evaluation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4566–4575.
- [472] Shashanka Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yuki M. Asano, and Yannis Avrithis. "Is ImageNet Worth 1 Video? Learning Strong Image Encoders From 1 Long Unlabelled Video". In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- [473] Ashvala Vinay and Alexander Lerch. "Evaluating Generative Audio Systems and Their Metrics". In: *ISMIR 2022 Hybrid Conference*. 2022.
- [474] E. Vincent, R. Gribonval, and C. Fevotte. "Performance Measurement in Blind Audio Source Separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), pp. 1462–1469. doi: 10.1109/TSA.2005.858005.
- [475] Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, et al. "AudioBench: A Universal Benchmark for Audio Large Language Models". In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, 2025, pp. 4297–4316. doi: 10.18653/v1/2025.naacl-long.218.
- [476] Chung-Che Wang, Jiawei Du, and Jyh-Shing Roger Jang. *Dcase 2023 Task 6b: Text-to-audio Retrieval Using Pretrained Models*. Tech. rep. DCASE2023 Challenge, Tech. Rep, 2023.
- [477] Helin Wang, Jiarui Hai, Yen-Ju Lu, Karan Thakkar, Mounya Elhilali, et al. *SoloAudio: Target Sound Extraction with Language-oriented Audio Diffusion Transformer*. Jan. 2025. doi: 10.48550/arXiv.2409.08425.
- [478] Helin Wang, Bang Yang, Yuexian Zou, and Dading Chong. "Automated Audio Captioning With Temporal Attention". In: *DCASE2020 Challenge, Tech. Rep.* (2020).
- [479] Helin Wang, Yuexian Zou, and Wenwu Wang. "SpecAugment++: A Hidden Space Data Augmentation Method for Acoustic Scene Classification". In: *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*. Ed. by Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, et al. ISCA, 2021, pp. 551–555. doi: 10.21437/INTERSPEECH.2021-140.
- [480] Hualei Wang, Yiming Li, Shuo Ma, Hong Liu, and Xiangdong Wang. *Listening Between the Frames: Bridging Temporal Gaps in Large Audio-Language Models*. Dec. 2025. doi: 10.48550/arXiv.2511.11039.
- [481] Hualei Wang, Jianguo Mao, Zhifang Guo, Jiarui Wan, Hong Liu, et al. "Leveraging Language Model Capabilities for Sound Event Detection". In: *Proceedings of Interspeech 2024*. 2024, pp. 4803–4807. doi: 10.21437/Interspeech.2024-112.
- [482] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, et al. "Generalizing to Unseen Domains: A Survey on Domain Generalization". In: *IEEE Transactions on Knowledge and DATA Engineering* 35.8 (2022), pp. 8052–8072.

-
- [483] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, et al. *ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities*. May 2023. doi: 10.48550/arXiv.2305.11172.
- [484] Rui Wang, Hongru Wang, Boyang Xue, Jianhui Pang, Shudong Liu, et al. “Harnessing the Reasoning Economy: A Survey of Efficient Reasoning for Large Language Models”. In: *CoRR* (2025). doi: 10.48550/arXiv.2503.24377.
- [485] Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, et al. *Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions*. Sept. 2024. doi: 10.48550/arXiv.2308.14263.
- [486] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, et al. “CogVLM: Visual Expert for Pretrained Language Models”. In: *Advances in Neural Information Processing Systems 37* (2023), pp. 121475–121499. doi: 10.48550/arXiv.2311.03079.
- [487] Wenwu Wang. *Machine Audition: Principles, Algorithms and Systems: Principles, Algorithms and Systems*. IGI global, 2010.
- [488] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, et al. *VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research*. June 2020. doi: 10.48550/arXiv.1904.03493.
- [489] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, et al. “Self-Instruct: Aligning Language Models With Self-Generated Instructions”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki. Association for Computational Linguistics, 2023, pp. 13484–13508. doi: 10.18653/V1/2023.ACL-LONG.754.
- [490] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, et al. “Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ Nlp Tasks”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, 2022, pp. 5085–5109. doi: 10.18653/V1/2022.EMNLP-MAIN.340.
- [491] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, et al. “AUDIT: Audio Editing by Following Instructions With Latent Diffusion Models”. In: *Advances in Neural Information Processing Systems 36* (2023), pp. 71340–71357. doi: 10.48550/arXiv.2304.00830.
- [492] Yuanyuan Wang, Hangting Chen, Dongchao Yang, Zhiyong Wu, Helen Meng, et al. *Audio-Composer: Towards Fine-grained Audio Generation with Natural Language Descriptions*. 2024. doi: 10.48550/arXiv.2409.12560.
- [493] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, et al. “Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis”. In: *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.
- [494] Zhepei Wang, Cem Subakan, Krishna Subramani, Junkai Wu, Tiago Tavares, et al. “Unsupervised Improvement of Audio-Text Cross-Modal Representations”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2023, New Paltz, NY, USA, October 22-25, 2023*. IEEE, 2023, pp. 1–5. doi: 10.1109/WASPAA58266.2023.10248180.
- [495] Zixuan Wang, Chi-Keung Tang, and Yu-Wing Tai. *Audio-Agent: Leveraging LLMs For Audio Generation, Editing and Composition*. 2024. doi: 10.48550/arXiv.2410.03335.
- [496] Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, György Fazekas, et al. “MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models”. In: *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*. Nov. 2024. doi: 10.5281/ZENODO.14877459.

- [497] Benno Weck and Xavier Serra. “Data Leakage in Cross-Modal Retrieval Training: A Case Study”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10094617.
- [498] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, et al. *Finetuned Language Models Are Zero-Shot Learners*. Feb. 2022. doi: 10.48550/arXiv.2109.01652.
- [499] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, et al. Vol. 35. Curran Associates, Inc., 2022, pp. 24824–24837. doi: 10.48550/arXiv.2201.11903.
- [500] Jason W. Wei and Kai Zou. “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computational Linguistics, 2019, pp. 6381–6387. doi: 10.18653/V1/D19-1670.
- [501] Shengyun Wei, Shun Zou, Feifan Liao, and Weimin Lang. “A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification”. In: *Journal of Physics: Conference Series* 1453.1 (Jan. 2020), p. 12085. doi: 10.1088/1742-6596/1453/1/012085.
- [502] Xiangyi Wei, Haotian Zhang, Xinyi Cao, Siyu Xie, Weifeng Ge, et al. *Audio-VLA: Adding Contact Audio Perception to Vision-Language-Action Model for Robotic Manipulation*. Nov. 2025. doi: 10.48550/arXiv.2511.09958.
- [503] Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. *SARI: Structured Audio Reasoning via Curriculum-Guided Reinforcement Learning*. Apr. 2025. doi: 10.48550/arXiv.2504.15900.
- [504] Hairu Wen, Yang Chen, Ze Yang, Lei Fu, Zeyu Wang, et al. “A Dynamic Adaptive Data Filtering and Sampling Framework for Optimized Fine-Tuning of Large Language Models”. In: *2025 5th International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. 2025, pp. 600–604. doi: 10.1109/ICCECE65250.2025.10984505.
- [505] Yutong Wen, Ke Chen, Prem Seetharaman, Oriol Nieto, Jiaqi Su, et al. *PromptSep: Generative Audio Separation via Multimodal Prompting*. Nov. 2025. doi: 10.48550/arXiv.2511.04623.
- [506] Janet F. Werker and Takao K. Hensch. “Critical Periods in Speech Perception: New Directions”. In: *Annual Review of Psychology* 66.1 (Jan. 2015), pp. 173–196. doi: 10.1146/annurev-psych-010814-015104.
- [507] Benjamin Lee Whorf. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Ed. by John B. Carroll. MIT Press, 1956.
- [508] Gijs Wijngaard, Elia Formisano, Michel Dumontier, and Jenia Jitsev. *AudioToolAgent: An Agentic Framework for Audio-Language Models*. Feb. 2026. doi: 10.48550/arXiv.2510.02995.
- [509] Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. “Audio-Language Datasets of Scenes and Events: A Survey”. In: *IEEE Access: Practical Innovations, Open Solutions* 13 (Jan. 2025), pp. 20328–20360. doi: 10.1109/ACCESS.2025.3534621.
- [510] Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. “AudSemThinker: Enhancing Audio-Language Models through Reasoning over Semantics of Sound”. In: *Neural Information Processing Systems (NeurIPS)*. San Diego, Dec. 2025.
- [511] Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. *Data-Balanced Curriculum Learning for Audio Question Answering*. July 2025. doi: 10.48550/arXiv.2507.06815.
- [512] Gijs Wijngaard, Elia Formisano, Bruno L. Giordano, and Michel Dumontier. “ACES: Evaluating Automated Audio Captioning Models on the Semantics of Sounds”. In: *2023 31st European Signal Processing Conference (EUSIPCO)*. Helsinki, Finland: IEEE, Sept. 2023, pp. 770–774. doi: 10.23919/EUSIPCO58844.2023.10289793.

-
- [513] Julia Wilkins, Justin Salamon, Magdalena Fuentes, Juan Pablo Bello, and Oriol Nieto. “Bridging High-Quality Audio and Video via Language for Sound Effects Retrieval From Visual Queries”. In: *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA: IEEE, Oct. 2023, pp. 1–5. doi: 10.1109/WASPAA58266.2023.10248113.
 - [514] Brandon T. Willard and Rémi Louf. *Efficient Guided Generation for Large Language Models*. Aug. 2023. doi: 10.48550/arXiv.2307.09702.
 - [515] Jonathan Winawer, Nathan Witthoft, Michael C. Frank, Lisa Wu, Alex R. Wade, et al. “Russian Blues Reveal Effects of Language on Color Discrimination”. In: *Proceedings of the National Academy of Sciences* 104.19 (2007), pp. 7780–7785. doi: 10.1073/pnas.0701644104.
 - [516] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, et al. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proc. Conference on Empirical Methods in NLP*. Online: ACL, Oct. 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.
 - [517] Hyejin Won, Baekseung Kim, Il-Youp Kwak, and Changwon Lim. “Transfer Learning Followed by Transformer for Automated Audio Captioning”. In: *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop*. 2021, pp. 221–225.
 - [518] Woods Hole Oceanographic Institution. *Watkins Marine Mammal Sound Database*. <https://whoicf2.whoi.edu/science/B/whalesounds/index.cfm>.
 - [519] Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, et al. *Step-Audio 2 Technical Report*. Aug. 2025. doi: 10.48550/arXiv.2507.16632.
 - [520] Ho-Hsiang Wu, Oriol Nieto, Juan Pablo Bello, and Justin Salamon. “Audio-Text Models Do Not yet Leverage Natural Language”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10097117.
 - [521] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. *Wav2CLIP: Learning Robust Audio Representations From Clip*. Feb. 2022. doi: 10.48550/arXiv.2110.11499.
 - [522] Mengyue Wu, Heinrich Dinkel, and Kai Yu. “Audio Caption: Listen and Tell”. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2019, pp. 830–834. doi: 10.1109/ICASSP.2019.8682377.
 - [523] Shih-Lun Wu, Xuankai Chang, Gordon Wichern, Jee-weon Jung, François Germain, et al. “Improving Audio Captioning Models with Fine-Grained Audio Features, Text Embedding Supervision, and LLM Mix-Up Augmentation”. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2024, pp. 316–320. doi: 10.1109/icaassp48485.2024.10447215.
 - [524] Wenxun Wu, Yuanyang Li, Guhan Chen, Linyue Wang, and Hongyang Chen. *Tool-Augmented Policy Optimization: Synergizing Reasoning and Adaptive Tool Use with Reinforcement Learning*. Oct. 2025. doi: 10.48550/arXiv.2510.07038.
 - [525] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, et al. “Large-Scale Contrastive Language-Audio Pretraining With Feature Fusion and Keyword-to-Caption Augmentation”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, June 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095969.
 - [526] Yusong Wu, Christos Tsirigotis, Ke Chen, Cheng-Zhi Anna Huang, Aaron Courville, et al. “FLAM: Frame-Wise Language-Audio Modeling”. In: *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025.
 - [527] Yusong Wu, Tianyu Zhang, and Ke Chen. “Text-to-Audio Retrieval via Large-Scale Contrastive Training”. In: *IEEE AASP Challenge Detection Classification Acoust. Scenes Events, Nancy, France (2022)*.

- [528] Feiyang Xiao, Jian Guan, Qiaoxi Zhu, Xubo Liu, Wenbo Wang, et al. *A Reference-free Metric for Language-Queried Audio Source Separation Using Contrastive Language-Audio Pretraining*. Jan. 2025. doi: 10.48550/arXiv.2407.04936.
- [529] Feiyang Xiao, Qiaoxi Zhu, Jian Guan, Xubo Liu, Haohe Liu, et al. *Synth-AC: Enhancing Audio Captioning with Synthetic Supervision*. Sept. 2023. doi: 10.48550/arXiv.2309.09705.
- [530] Feiyang Xiao, Qiaoxi Zhu, Haiyan Lan, Wenwu Wang, and Jian Guan. “Ensemble Systems With Contrastive Language-Audio Pretraining and Attention-Based Audio Features for Audio Captioning and Retrieval”. In: *IEEE AASP Challenge Detection Classification Acoust. Scenes Events, Tampere, Finland, Tech. Rep* (2023).
- [531] Yang Xiao and Rohan Kumar Das. *WildDESED: An LLM-Powered Dataset for Wild Domestic Environment Sound Event Detection System*. Oct. 2024. doi: 10.48550/arXiv.2407.03656.
- [532] Huang Xie, Khazar Khorrami, Okko Räsänen, and Tuomas Virtanen. *Crowdsourcing and Evaluating Text-Based Audio Retrieval Relevances*. Aug. 2023. doi: 10.48550/arXiv.2306.09820.
- [533] Huang Xie, Khazar Khorrami, Okko Räsänen, and Tuomas Virtanen. *Integrating Continuous and Binary Relevances in Audio-Text Relevance Learning*. Aug. 2024. doi: 10.48550/arXiv.2408.14939.
- [534] Huang Xie, Samuel Lipping, and Tuomas Virtanen. *Language-Based Audio Retrieval Task in Dcase 2022 Challenge*. Oct. 2022. doi: 10.48550/arXiv.2209.09967.
- [535] Huang Xie, Okko Räsänen, and Tuomas Virtanen. “On Negative Sampling for Contrastive Audio-Text Retrieval”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, June 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095319.
- [536] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. “Unsupervised Data Augmentation for Consistency Training”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*. Ed. by Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. arXiv, 2020.
- [537] Z Xie, X Xu, Z Wu, and M Wu. *PicoAudio: Enabling Precise Timestamp and Frequency Controllability of Audio Events in Text-to-audio Generation*. 2024. doi: 10.48550/arXiv.2407.02869.
- [538] Zeyu Xie, Baihan Li, Xuenan Xu, Zheng Liang, Kai Yu, et al. “FakeSound: Deepfake General Audio Detection”. In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 112–116. doi: 10.21437/Interspeech.2024-1703.
- [539] Zeyu Xie, Xuenan Xu, Mengyue Wu, and Kai Yu. “Enhance Temporal Relations in Audio Captioning With Sound Event Detection”. In: *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 4179–4183. doi: 10.21437/Interspeech.2023-1614.
- [540] Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. “AudioTime: A Temporally-aligned Audio-text Benchmark Dataset”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Hyderabad, India: IEEE, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10889879.
- [541] Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, et al. *Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models*. Sept. 2025. doi: 10.48550/arXiv.2503.02318.
- [542] Cheng Xu, Shuhao Guan, Derek Greene, and M.-Tahar Kechadi. *Benchmark Data Contamination of Large Language Models: A Survey*. June 2024. doi: 10.48550/arXiv.2406.04244.
- [543] David Xu. *AudioSetMix: Enhancing Audio-Language Datasets With LLM-Assisted Augmentations*. June 2024. doi: 10.48550/arXiv.2405.11093.
- [544] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, et al. *Qwen2.5-Omni Technical Report*. Mar. 2025. doi: 10.48550/arXiv.2503.20215.

-
- [545] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. “MSR-VTT: A Large Video Description Dataset for Bridging Video and Language”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 5288–5296. doi: 10.1109/CVPR.2016.571.
- [546] M Xu, C Li, X Tu, Y Ren, R Fu, et al. *Towards Diverse and Efficient Audio Captioning via Diffusion Models*. 2024. doi: 10.48550/ARXIV.2409.09401.
- [547] Xinmeng Xu, Yiqun Zhang, Yuhong Yang, and Weiping Tu. “Leveraging Sound Local and Global Features for Language-Queried Target Sound Extraction”. In: *International Conference on Neural Information Processing* 14450 (2024). Ed. by Biao Luo, Long Cheng, Zheng-Guang Wu, Hongyi Li, and Chaojie Li, pp. 367–379. doi: 10.1007/978-981-99-8070-3_28.
- [548] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. “A CRNN-GRU Based Reinforcement Learning Approach to Audio Captioning.” In: *Detection and Classification of Acoustic Scenes and Events*. 2020.
- [549] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. “Audio Caption in a Car Setting With a Sentence-Level Loss”. In: *12th International Symposium on Chinese Spoken Language Processing, ISCSLP 2021, Hong Kong, January 24-27, 2021*. IEEE, 2021, pp. 1–5. doi: 10.1109/ISCSLP49672.2021.9362117.
- [550] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. “Text-to-Audio Grounding: Building Correspondence Between Captions and Sound Events”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 606–610. doi: 10.1109/ICASSP39728.2021.9414834.
- [551] Xuenan Xu, Ziyang Ma, Mengyue Wu, and Kai Yu. *Towards Weakly Supervised Text-to-Audio Grounding*. July 2024. doi: 10.48550/arXiv.2401.02584.
- [552] Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kai Yu. “Beyond the Status Quo: A Contemporary Survey of Advances and Challenges in Audio Captioning”. In: *IEEE ACM Trans. Audio Speech Lang. Process.* 32 (2024), pp. 95–112. doi: 10.1109/TASLP.2023.3321968.
- [553] Xuenan Xu, Xiaohang Xu, Zeyu Xie, Pingyue Zhang, Mengyue Wu, et al. “A Detailed Audio-Text Data Simulation Pipeline Using Single-Event Sounds”. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 1091–1095. doi: 10.1109/ICASSP48485.2024.10448441.
- [554] Xuenan Xu, Zhiling Zhang, Zelin Zhou, Pingyue Zhang, Zeyu Xie, et al. “BLAT: Bootstrapping Language-Audio Pre-Training Based on Audioset Tag-Guided Synthetic Data”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. Ottawa ON Canada: ACM, Oct. 2023, pp. 2756–2764. doi: 10.1145/3581783.3613820.
- [555] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, et al. “Advancing High-Resolution Video-Language Representation With Large-Scale Video Transcriptions”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 5026–5035. doi: 10.1109/CVPR52688.2022.00498.
- [556] J Xue, Y Deng, Y Gao, and Y Li. *Auffusion: Leveraging the Power of Diffusion and Large Language Models for Text-to-Audio Generation*. 2024. doi: 10.1109/TASLP.2024.3485485.
- [557] Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, et al. *SimpleTIR: End-to-End Reinforcement Learning for Multi-Turn Tool-Integrated Reasoning*. Sept. 2025. doi: 10.48550/arXiv.2509.02479.
- [558] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, et al. *Learning to Reason under Off-Policy Guidance*. June 2025. doi: 10.48550/arXiv.2504.14945.
- [559] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. *Qwen3 Technical Report*. May 2025. doi: 10.48550/arXiv.2505.09388.
- [560] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, et al. *Large Language Models as Optimizers*. Apr. 2024. doi: 10.48550/arXiv.2309.03409.

- [561] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, et al. “UniAudio: Towards Universal Audio Generation with Large Language Models”. In: *Forty-First International Conference on Machine Learning*. 2024.
- [562] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, et al. *DiffSound: Discrete Diffusion Model for Text-to-Sound Generation*. Apr. 2023. doi: 10.48550/arXiv.2207.09983.
- [563] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, et al. “AVQA: A Dataset for Audio-Visual Question Answering on Videos”. In: *Proceedings of ACM International Conference on Multimedia*. Mm ’22. ACM, Oct. 2022, pp. 3480–3491. doi: 10.1145/3503161.3548291.
- [564] Qi Yang, Binjie Mao, Zili Wang, Xing Nie, Pengfei Gao, et al. *Draw an Audio: Leveraging Multi-Instruction for Video-to-Audio Synthesis*. 2024. doi: 10.48550/arXiv.2409.06135.
- [565] Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, et al. “AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Association for Computational Linguistics, 2024, pp. 1979–1998. doi: 10.18653/v1/2024.acl-long.109.
- [566] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, et al. “ReAct: Synergizing Reasoning and Acting in Language Models”. In: *ICLR*. 2022. doi: 10.48550/ARXIV.2210.03629.
- [567] Zhongjie Ye, Helin Wang, Dongchao Yang, and Yuexian Zou. *Improving the Performance of Automated Audio Captioning Via Integrating the Acoustic and Semantic Information*. Oct. 2021. doi: 10.48550/arXiv.2110.06100.
- [568] Zhongjie Ye, Yuqing Wang, Helin Wang, Dongchao Yang, and Yuexian Zou. “FeatureCut: An Adaptive Data Augmentation for Automated Audio Captioning”. In: *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Nov. 2022, pp. 313–318. doi: 10.23919/APSIPAASC55919.2022.9980325.
- [569] Han Yin, Jisheng Bai, Yang Xiao, Hui Wang, Siqi Zheng, et al. “Exploring Text-Queried Sound Event Detection with Audio Source Separation”. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, pp. 1–5. doi: 10.1109/icassp49660.2025.10889789.
- [570] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, et al. “A Survey on Multimodal Large Language Models”. In: *National Science Review* 11.12 (Nov. 2024), nwae403. doi: 10.1093/nsr/nwae403.
- [571] Xinlei Yin, Xiulian Peng, Xue Jiang, Zhiwei Xiong, and Yan Lu. “Text-Queried Audio Source Separation via Hierarchical Modeling”. In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 4928–4940. doi: 10.1109/TASLPRD.2025.3633078.
- [572] Yi Yuan, Zhuo Chen, Xubo Liu, Haohe Liu, Xuenan Xu, et al. “T-CLAP: Temporal-Enhanced Contrastive Language-Audio Pretraining”. In: *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*. London, United Kingdom: IEEE, Sept. 2024, pp. 1–6. doi: 10.1109/MLSP58920.2024.10734763.
- [573] Yi Yuan, Dongya Jia, Xiaobin Zhuang, Yuanzhe Chen, Zhengxi Liu, et al. “Sound-VECaps: Improving Audio Generation with Visual Enhanced Captions”. In: *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*. 2024.
- [574] Yi Yuan, Xubo Liu, Haohe Liu, Mark D. Plumbley, and Wenwu Wang. “FlowSep: Language-Queried Sound Separation with Rectified Flow Matching”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Hyderabad, India: IEEE, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10890129.
- [575] Fan Zhang and Michael Gienger. *Learning Robot Manipulation from Audio World Models*. Dec. 2025. doi: 10.48550/arXiv.2512.08405.

-
- [576] Hang Zhang, Xin Li, and Lidong Bing. “Video-LLaMA: An Instruction-Tuned Audio-Visual Language Model for Video Understanding”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*. Ed. by Yansong Feng and Els Lefever. Association for Computational Linguistics, 2023, pp. 543–553. doi: 10.18653/V1/2023.EMNLP-DEMO.49.
- [577] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. “Mixup: Beyond Empirical Risk Minimization”. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
- [578] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. “PEGASUS: Pre-Training With Extracted Gap-Sentences for Abstractive Summarization”. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 11328–11339.
- [579] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. “BERTScore: Evaluating Text Generation With Bert”. In: *International Conference on Learning Representations*. Feb. 2020. doi: 10.48550/arXiv.1904.09675.
- [580] Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. *Beyond Random Sampling: Efficient Language Model Pretraining via Curriculum Learning*. June 2025. doi: 10.48550/arXiv.2506.11300.
- [581] Yiming Zhang, Ruoyi Du, Zheng-Hua Tan, Wenwu Wang, and Zhanyu Ma. “Generating Accurate and Diverse Audio Captions Through Variational Autoencoder Framework”. In: *IEEE Signal Processing Letters* 31 (2024), pp. 2520–2524. doi: 10.1109/LSP.2024.3409212.
- [582] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, et al. *LLaVA-Video: Video Instruction Tuning With Synthetic Data*. Aug. 2025. doi: 10.48550/arXiv.2410.02713.
- [583] Zihan Zhang, Xize Cheng, Zhennan Jiang, Dongjie Fu, Jingyuan Chen, et al. *MARS-Sep: Multimodal-Aligned Reinforced Sound Separation*. Oct. 2025. doi: 10.48550/arXiv.2510.10509.
- [584] Hang Zhao, Yifei Xin, Zhesong Yu, Bilei Zhu, Lu Lu, et al. “MINT: Boosting Audio-Language Model via Multi-Target Pre-Training and Instruction Tuning”. In: *Interspeech* (June 2024). doi: 10.48550/arXiv.2402.07485.
- [585] J Zhao, X Qian, Y Xu, H Liu, Y Cao, et al. “Text-Queried Target Sound Event Localization”. In: *EUSIPCO 2024*. Aug. 2024, pp. 261–265. doi: 10.23919/eusipco63174.2024.10715199.
- [586] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, et al. *A Survey of Large Language Models*. Oct. 2024. doi: 10.48550/arXiv.2303.18223.
- [587] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, et al. *BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs*. 2023. doi: 10.48550/ARXIV.2307.08581.
- [588] Yanpeng Zhao, Jack Hessel, Youngjae Yu, Ximing Lu, Rowan Zellers, et al. “Connecting the Dots Between Audio and Text Without Parallel Data Through Visual Knowledge Transfer”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*. Association for Computational Linguistics, 2022, pp. 4492–4507. doi: 10.18653/V1/2022.NAAACL-MAIN.333.
- [589] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, et al. *ChatBridge: Bridging Modalities With Large Language Model as a Language Catalyst*. May 2023. doi: 10.48550/arXiv.2305.16103.
- [590] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, et al. “Judging LLM-as-a-Judge With MT-Bench and Chatbot Arena”. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, et al. 2023.

- [591] Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, et al. “BAT: Learning to Reason About Spatial Sounds With Large Language Models”. In: *Forty-First International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- [592] Zhuo Zhi, Chen Feng, Adam Daneshmendi, Mine Orlu, Andreas Demosthenous, et al. *Seeing and Reasoning With Confidence: Supercharging Multimodal LLMs With an Uncertainty-Aware Agentic Framework*. Mar. 2025. doi: 10.48550/arXiv.2503.08308.
- [593] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. “Learning Deep Features for Scene Recognition Using Places Database”. In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014.
- [594] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. *Domain Generalization with MixStyle*. 2021. doi: 10.48550/ARXIV.2104.02008.
- [595] Luwei Zhou, Chenliang Xu, and Jason Corso. “Towards Automatic Learning of Procedures From Web Instructional Videos”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). doi: 10.1609/aaai.v32i1.12342.
- [596] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, et al. *Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models*. 2024. doi: 10.48550/arXiv.2405.13872.
- [597] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, et al. “BERT Loses Patience: Fast and Robust Inference with Early Exit”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 18330–18341.
- [598] Y Zhou, C Guo, X Wang, Y Chang, and Y Wu. *A Survey on Data Augmentation in Large Model Era*. 2024. doi: 10.48550/ARXIV.2401.15422.
- [599] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, et al. “Mixture-of-Experts with Expert Choice Routing”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, et al. Vol. 35. Curran Associates, Inc., 2022, pp. 7103–7114.
- [600] Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, et al. “Can Audio Captions Be Evaluated With Image Caption Metrics?” In: *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. arXiv, Jan. 2022. doi: 10.1109/ICASSP43922.2022.9746427.
- [601] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, et al. “Languagebind: Extending Video-Language Pretraining to N-Modality by Language-Based Semantic Alignment”. In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- [602] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. “MiniGPT-4: Enhancing Vision-Language Understanding With Advanced Large Language Models”. In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- [603] Ge Zhu, Jordan Darefsky, and Zhiyao Duan. “Cacophony: An Improved Contrastive Audio-Text Model”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 4867–4879. doi: 10.1109/TASLP.2024.3485170.
- [604] Xinfa Zhu, Wenjie Tian, Xinsheng Wang, Lei He, Xi Wang, et al. *CosyAudio: Improving Audio Generation with Confidence Scores and Synthetic Captions*. Jan. 2025. doi: 10.48550/arXiv.2501.16761.
- [605] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, et al. “Cross-Task Weakly Supervised Learning From Instructional Videos”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 3532–3540. doi: 10.1109/CVPR.2019.00365.

-
- [606] Noah Ziem, Dilara Soylu, Lakshya A. Agrawal, Isaac Miller, Liheng Lai, et al. *Multi-Module GRPO: Composing Policy Gradients and Prompt Optimization for Language Model Programs*. Aug. 2025. doi: 10.48550/arXiv.2508.04660.

Summary



This dissertation explores how computers can learn to understand sound and describe it using language. The goal is to approach human capabilities. When a person hears a dog barking, they can instantly recognize it and describe what they hear. The research in this dissertation focuses on building and improving computer systems that move toward similar capabilities. The systems analyze which sound is present, what action is involved, where it occurs, and which objects are involved.

The foundation of teaching computers to understand sound lies in showing them examples. Just as children learn by being exposed to examples and learn by doing, these computer systems learn from large collections of sounds, paired with text descriptions. The first part of this work provides a survey of 69 datasets used in this field and examines where they come from and what they contain. There are challenges in the use of audio data. For example, there is substantial overlap between datasets, which can give a misleading impression of how well these systems perform when tested.

Understanding these data problems led to the second contribution: improving the way computers learn from these data. Training a computer system is not simply about showing all available examples at once. Research shows that starting with easier examples and gradually introducing more difficult ones leads to better learning outcomes. This approach, inspired by how curriculum is designed in education, was applied to the task of answering questions about audio. It introduces examples in order of difficulty. Other approaches include filtering out duplicate data and balancing overrepresented sound categories. The resulting system achieved improvements in performance without requiring more computational resources.

The third major contribution draws inspiration from how humans perceive and describe sounds. When people describe what they hear, they naturally organize their descriptions around key questions: Who or what is making the sound? How is it produced? Where and when is it happening? Cognitive neuroscience research suggests that recognizing sound sources, understanding how sounds are made, and placing them in context involve specific mental processes. Building on this, the AudSemThinker model was developed. AudSemThinker explicitly reasons about these semantic aspects of sound before generating its final answer, which is similar to how a person might think through what they are hearing before speaking. To train AudSemThinker, a completely new dataset called AudSem was created from YouTube

video subtitles. It was designed to avoid overlap with existing test sets. The resulting model matches or exceeds the performance of strong existing systems on multiple benchmarks.

Improving these systems also requires better ways to evaluate them. Traditional evaluation methods often focus on word matching. The fourth contribution introduces a new evaluation metric, called ACES, which judges computer-generated descriptions based on whether they capture the essential semantic elements of the sound. Instead of simply comparing words, ACES examines whether descriptions correctly identify sound entities, actions, and other contexts. The idea is that this metric aligns more closely with how humans judge the quality of audio descriptions.

The final contribution takes a different approach to audio understanding by combining multiple specialized systems. Like how a team of human experts with different specializations can solve problems by collaborating, the AudioToolAgent framework coordinates several specialized computer programs. A central reasoning system acts as a coordinator, deciding which specialized tools to consult for different aspects of an audio task. For example, when processing audio that contains speech, the coordinator delegates transcription to a speech recognition specialist, while consulting other tools for environmental sounds or music. This collaborative approach achieves strong results on multiple challenging benchmarks. It remains modular and cost-effective by reusing existing pre-trained models.

These contributions advance the field of audio-language learning by: (1) addressing challenges in data quality and availability; (2) improving audio question answering via data-balanced curricula; (3) introducing human-inspired reasoning mechanisms; (4) developing evaluation methods that focus on semantic content; and (5) demonstrating how coordinated specialist systems can outperform single general-purpose models. These developments move computers closer to understanding the acoustic world and communicating that understanding through natural language.

Samenvatting



Titel: Data en Semantische Redenatie in Audio-Taal Modellen

Dit proefschrift onderzoekt hoe computers kunnen leren om geluid te begrijpen en het te beschrijven met taal, vergelijkbaar met hoe mensen dat doen. Wanneer je bijvoorbeeld een hond hoort blaffen, kun je onmiddellijk herkennen en beschrijven wat het is dat je hoort. Het uiteindelijke doel met dit proefschrift is computersystemen te bouwen en te verbeteren, zodat zij geluiden even goed kunnen herkennen en duiden als de mens.

Computers leren door middel van data. Deze data worden gebruikt om een computer te trainen. Net zoals kinderen die leren door blootgesteld te worden aan voorbeelden, leren deze computersystemen van grote verzamelingen geluiden gekoppeld aan tekst. Het eerste deel van dit werk biedt een uitgebreid overzicht van 69 datasets die in dit vakgebied worden gebruikt, waarbij wordt onderzocht waar ze vandaan komen en wat ze bevatten. Hieruit blijkt dat het merendeel van deze data afkomstig is van platforms zoals YouTube en Freesound. Problemen zijn er ook, zoals aanzienlijke onderlinge overlap, wat een verkeerde indruk kan geven van hoe goed deze systemen daadwerkelijk presteren wanneer ze worden getest.

Om deze dataproblemen te beperken, richt dit proefschrift zich op hoe computers van data leren. Een aanpak is curriculum learning: het model krijgt eerst makkelijke voorbeelden te zien en daarna geleidelijk moeilijkere. Dit verbetert leerresultaten bij het beantwoorden van vragen over geluiden. Andere technieken, zoals het zorgvuldig uifilteren van dubbele data en het beperken van oververtegenwoordigde geluidscategorieën, verbeteren de prestaties verder.

De derde hoofdbijdrage haalt inspiratie uit hoe mensen geluiden waarnemen en beschrijven. Wanneer mensen beschrijven wat ze horen, organiseren ze hun beschrijvingen van nature rond kernvragen: Wie of wat maakt het geluid? Hoe wordt het geproduceerd? Waar en wanneer gebeurt het? Cognitief neurowetenschappelijk onderzoek suggereert dat het herkennen van geluidsbronnen, het begrijpen hoe geluiden worden gemaakt, en het plaatsen ervan in context allemaal specifieke mentale processen zijn. Voortbouwend op dit inzicht werd het model AudSemThinker ontwikkeld. Dit systeem redeneert expliciet over deze “semantische” aspecten van geluid voordat het zijn definitieve antwoord genereert, vergelijkbaar met hoe een persoon zou kunnen nadenken over wat hij hoort alvorens te spreken. Om dit systeem te trainen, werd een volledig nieuwe dataset genaamd AudSem gecreëerd uit YouTube

video ondertitels, zorgvuldig ontworpen om overlap met bestaande data te vermijden. Het resulterende model presteerde beter dan bijna alle vergelijkbare systemen op meerdere benchmarks.

Het verbeteren van deze systemen vereist ook betere manieren om ze te evalueren. Traditionele evaluatiemethoden richten zich vaak op oppervlakkige woordovereenkomsten. De vierde bijdrage introduceert een nieuwe evaluatiemethode genaamd ACES, die door computers gegenereerde beschrijvingen beoordeelt op basis van of ze de essentiële semantische elementen van het geluid vastleggen. In plaats van simpelweg woorden te vergelijken, onderzoekt ACES of beschrijvingen geluidsentiteiten, acties en andere contexten correct identificeren. De gedachte erachter is dat deze metriek nauwer aansluit bij hoe mensen de kwaliteit van audiobeschrijvingen beoordelen.

De laatste bijdrage hanteert een andere benadering van het herkennen van audio. In plaats van een enkel model, worden er meerdere gespecialiseerde systemen gecombineerd. Vergelijkbaar met hoe een team van menselijke experts met verschillende specialisaties complexe problemen kan oplossen door samen te werken, coördineert het AudioToolAgent-framework verschillende computerprogramma's. Een centraal systeem fungeert als coördinator en beslist welke gespecialiseerde tools geraadpleegd moeten worden voor verschillende aspecten voor het herkennen van de audio, waarbij de coördinator zelf niet de audio kan luisteren. Bij het verwerken van audio die spraak bevat, delegeert de coördinator transcriptie aan een spraakherkenningspecialist, terwijl hij andere tools raadpleegt voor omgevingsgeluiden of muziek. Door samen te werken met verschillende modellen, en een overkoepelende model alle denkstappen te laten maken, behaalt deze methode de beste resultaten op meerdere uitdagende evaluaties. Een groot voordeel hiervan is de besparing in rekentijd en middelen; bestaande modellen worden hergebruikt in plaats van het laten leren van een geheel nieuw model.

Gezamenlijk bevorderen deze bijdragen het vakgebied van de combinatie van audio en taal, door: (1) belangrijke uitdagingen in datakwaliteit en beschikbaarheid te duiden; (2) op mensen geïnspireerde leermethodieken te testen; (3) nieuwe redeneermechanismen op basis van hoe mensen leren te introduceren; (4) betere evaluatiemethoden te ontwikkelen; en (5) aan te tonen hoe gecoördineerde specialistische systemen beter kunnen presteren dan enkele modellen. Dit proefschrift maakt computers beter in staat om de rijke wereld van geluid om ons heen te begrijpen en dat te communiceren via natuurlijke taal.

Valorization



This dissertation studies how computers learn to listen to everyday sounds and describe them in natural language. The main purpose of the research is to make audio-language models more broadly useful. The work focuses on two themes. First, it improves the collection of example sounds and text descriptions on which these models are trained. Second, it designs models that reason about sounds in a way that is closer to human thinking. Across the chapters, the central conclusion is that better data and better reasoning jointly lead to audio systems that perform more accurately on benchmarks and generalize better to new situations.

The scientific contribution of this work lies in clarifying how data and reasoning shape audio-language learning. In the short term, the survey of datasets, training procedures, metric, semantic reasoning research, and the agent framework give researchers concrete tools to build and assess new models. The analysis of dataset overlap and bias informs how future benchmarks should be designed, including the need for cleaner test sets and more diverse sound sources. The curriculum pipeline provides a template for how to structure training for question answering tasks in a way that avoids data leakage. The AudSem and AudSemThinker work links ideas from neuroscientific research and human-inspired reasoning to machine learning model design. ACES adds a semantics-aware evaluation to the field, and AudioToolAgent offers a framework for agentic systems that treat audio models as tools that can be combined and that can co-operate together. In this way, this research also supports follow-up research in machine learning, audio signal processing, natural language processing, and cognitive science.

The long-term scientific impact goes beyond audio alone. The methods developed here give insight into how multimodal models could reason over richer sensory inputs in a structured way. Semantic frameworks and evaluation methods can inspire similar work for vision and robotics. The agentic setup in AudioToolAgent will be seen more in future systems where multiple specialized models collaborate under the guidance of a general reasoning agent. Such systems are relevant for research on world models that combine hearing with vision, touch, and other senses. The open-source code and datasets serve as material for courses on machine listening, multimodal learning, and evaluation methods, helping new students gain hands-on experience with state-of-the-art techniques.

The potential impact on society builds on these scientific advances but remains mostly indirect and long term. The dissertation focuses on fundamental methods and public

resources rather than finished consumer products. However, the ability to interpret sounds and explain them in plain language is closely related to everyday needs. People who are deaf or hard of hearing rely on visual information and vibrations to detect events in their surroundings. Systems based on audio-language models make it possible to translate alarms, doorbells, approaching vehicles, and other important sounds into readable text or haptic feedback on a phone or wearable device. In public spaces, richer subtitles and audio descriptions can make online videos more accessible. In health care and elderly care, sound monitoring can detect patterns such as coughing, calls for help, or unusual silence, and can alert caregivers when needed.

Environmental monitoring is another domain where audio-language systems are relevant. Many animal species are easier to track by sound than by sight. Microphones placed in forests, wetlands, or cities record continuous soundscapes that are difficult to inspect manually. Models that recognize and describe bird calls, insect activity, whale sounds, or human noise provide a scalable way to monitor biodiversity, detect illegal hunting, or track changes in the acoustic environment of cities. Similar ideas apply in industrial and transport settings. Machines and vehicles have characteristic sound signatures. Early changes in these signatures can signal wear or malfunction. Audio-language systems that explain these changes in simple terms help technicians and operators respond in time, which supports safety and reduces unexpected downtime.

The work is also relevant to the sectors that develop digital assistants and robots. Sound understanding improves safety and usability in homes. Assistants that understand speech and background sounds can react to events more easily. In robotics, world models that incorporate hearing alongside vision and touch enable robots to operate more safely in environments where sight alone is not enough. For example, a household robot that hears a kettle boiling or a pot overflowing can react earlier than one that relies only on images.

An important step toward bringing this research into practice is a grant received by the supervisor prof. Elia Formisano. This grant supports bringing the fundamental research of this dissertation into practice. This grant is being applied at the Brightlands Institute of Smart Sciences in Heerlen. Within this program, business developers created one-pagers with project descriptions that they use to translate the core ideas of this dissertation into concrete application sketches, such as patient monitoring through sound, acoustic safety systems in homes and cities, predictive maintenance for machines, and sound-based support for drivers and vehicles. These activities are ongoing and business developers are contacting potential partners for the techniques developed in this dissertation.

Several groups form the main audience for the findings of this dissertation. Within academia, the research targets scientists in fields such as machine learning, audio

signal processing, computational linguistics, human-computer interaction, and cognitive neuroscience. These communities also benefit from public datasets and code published with each chapter when designing and evaluating new models. Beyond academia, a potential audience for this research is developers in technology companies and start-ups who particularly work on tools for hearing support, environmental sensing, smart homes, and robotics. Other groups include Health-care providers, patient organizations, environmental agencies, and city planners. For them, the work provides better performing models that future monitoring tools may rely on. Lastly, people with hearing impairments, field biologists, maintenance engineers, and first responders are potential indirect target users.

The dissertation follows an open and FAIR approach: All major components developed in the chapters are released as open-source code and, when licensing allows, as public datasets. This allows researchers and practitioners to reuse and extend the methods and reproduce the experiments. The repositories include documentation and example scripts that are easy to use. For the scientific community, results are presented at conferences and in journals.

Acknowledgments



First of all, I would like to express my gratitude to my promoter and mentor, Elia Formisano. I greatly appreciate that he offered me this position and believed in me from the start. Elia helped me during challenging times when dealing with short deadlines and mountains of work. I sincerely enjoyed working with him and the total freedom he gave me to shape my PhD into something that I could be proud of. He did not require or force me to pursue specific directions, but he provided the opportunity to follow my own research ideas while still offering guidance and feedback.

I would also like to thank my co-supervisor, Michel Dumontier. He taught me the best practices to become a good scientist and researcher and offered many ideas on interesting, novel research directions to pursue. He often joined me in enthusiasm and encouraged me to pursue my own interests in the process of becoming a researcher.

I am incredibly fortunate to have had the Aud2Sem group, which shares my passion for this research subject and has interesting questions and ideas for things to explore. Especially the group in Maastricht, particularly Michele, as well as Maria, Tim, and Tonio, were extremely supportive and great colleagues and friends. The group in Marseille, particularly Bruno Giordano, Giorgio, Christian, and Marie, has been immensely helpful by providing interesting opinions and perspectives. Along the way, Martin, Yeni, and Chris have helped and supported me in the process.

I would also like to express my appreciation for my IDS colleagues, who have been a great support and encouragement in my development as a researcher. I enjoyed the fun times and events we shared, as well as the stimulating discussions about our work. Especially thanks to Chang, Komal, Visara, Thales, Yaru, Ensar, Vincent, Yuyang, Shervin, Maryam, Jinzhou, Özge, Remzi, Anda, Mikołaj, Dong, Kun, Seun, Louis, Parveen, Rishabh, Rohan, Xu, Christopher, Carlos, and Rena.

I would like to express my gratitude to Jan Scholtes, with whom I truly enjoyed collaborating in teaching, along with all his ideas and delightful stories about sailing and wine. I would like to thank Aki Harma for the interesting talks about the field and for providing me with valuable advice. I would also like to thank my colleagues at DACS, including Abderrahmane, Pawel, Subilal, Lars, Linda, Jerry, Bulat, Mado, Cedric, Matus, and Ellen.

Thank you to Job for being such a great friend here in Maastricht and for showing me around and introducing me to your large group of friends. Thanks to my wonderful friends Mark, Dirk, John, Elias, Beau, and Luuk. Thanks to my parents, my

sister, my grandparents, and my extended family for their unlimited support and encouragement in my life.

And of course, my greatest thanks to my girlfriend Mariëlle for standing by me through these 4 years of my PhD. I know the long working nights were not the most fun for you, but I know you were there for me every step of the way.

All content was written by me and, where applicable, by my supervisors. Generative AI tools such as ChatGPT, Gemini, and Claude were occasionally used to help me rephrase specific sentences or overcome writer's block, but they did not produce any final text without substantial editing by me. For code editing and bug fixing, AI tools such as Cursor, Claude Code, and Codex were very helpful.

Gijs Wijngaard
Maastricht
June

Published work

Gijs Wijngaard, Elia Formisano, Bruno L. Giordano, and Michel Dumontier. “ACES: Evaluating Automated Audio Captioning Models on the Semantics of Sounds”. In: *2023 31st European Signal Processing Conference (EUSIPCO)*. Helsinki, Finland: IEEE, Sept. 2023, pp. 770–774. doi: 10.23919/EUSIPCO58844.2023.10289793

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. *Data-Balanced Curriculum Learning for Audio Question Answering*. July 2025. doi: 10.48550/arXiv.2507.06815

Gijs Wijngaard, Elia Formisano, Michel Dumontier, and Jenia Jitsev. *AudioToolAgent: An Agentic Framework for Audio-Language Models*. Feb. 2026. doi: 10.48550/arXiv.2510.02995

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. “Audio-Language Datasets of Scenes and Events: A Survey”. In: *IEEE Access: Practical Innovations, Open Solutions* 13 (Jan. 2025), pp. 20328–20360. doi: 10.1109/ACCESS.2025.3534621

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. “AudSem-Thinker: Enhancing Audio-Language Models through Reasoning over Semantics of Sound”. In: *Neural Information Processing Systems (NeurIPS)*. San Diego, Dec. 2025

About the author



Gijs Wijngaard was born in 1996 in Den Bosch, the Netherlands. He grew up in a small town called Bilthoven, near Utrecht. From an early age, he was fascinated by computers and sounds/music. Having started making websites at age 12 and founded his own business at 14, his interest in programming and artificial intelligence developed early.

In high school, he was fascinated by artificial intelligence through science fiction movies like *Her*, *Transcendence*, and *Ex Machina*, and he decided that AI would be a great and fascinating field to work in. He pursued a Bachelor's degree in Artificial Intelligence at Utrecht University (2015-2020), focusing on machine learning techniques. During his bachelor studies, he wrote his thesis on clustering soccer players using unsupervised learning methods, supervised by dr. Tejaswini Deoskar. Seeking to broaden his horizons, he spent a semester at Monash University in Melbourne, Australia (2019), where he took a minor in informatics, focusing on data science and data analytics.

Following his bachelor's degree, Gijs completed a Master's degree in Information Science with a specialization in Data Science at the University of Amsterdam (2020-2021). His master's thesis explored unsupervised approaches for analyzing COVID-19 conspiracies, supervised by prof. dr. Antal van den Bosch and dr. Maarten Marx. During his master's studies, he also worked as a research intern at the Meertens Institute (April-July 2021), where he researched NLP techniques for analyzing diversity in language and culture in the Netherlands as part of the BETTER-MODS project funded by the Dutch Research Council (NWO).

In December 2021, Gijs began his PhD in Machine Learning at Maastricht University under the supervision of prof. dr. Elia Formisano and prof. dr. Michel Dumontier.



His research focused on audio-language models for scenes and events, particularly advancing automated audio captioning and audio question answering through human-inspired reasoning and data. His work bridged cognitive neuroscience and artificial intelligence, drawing on how humans perceive and describe sounds to develop more interpretable and effective audio-language systems.

Throughout his PhD, Gijs has been motivated to make his research accessible and applicable. He has developed multiple open-source software packages as part of his PhD research, including ACES (Audio Captioning Evaluation on Semantics of Sound), which has been adopted by the research community for evaluating audio captioning systems. His work on audio-language datasets has provided insights into data overlap and quality issues affecting the field. He has actively contributed to the research community by releasing code, datasets, and pretrained models through platforms like GitHub and Hugging Face to facilitate reproducibility and enable other researchers to build upon his work.

Gijs is passionate about the intersection of entrepreneurship and research. Before his PhD, he co-founded Duwtje (2018-2020), an AI-powered news application, where he served as CTO and was responsible for app development and data engineering. The startup participated in the Dutch Journalism Fund accelerator and received €75K in funding. He also worked as a front-end developer at a startup called Channable (2017-2018, 2020), contributing to the development of a SaaS platform for managing online marketing.

His research has been recognized through multiple awards and grants, including three Dutch Research Council compute grants for research on the Snellius supercomputer totaling €250K, an OpenAI Research Grant, and a NeurIPS Scholar Award. He has published extensively as a first author in top-tier venues, with accepted papers at NeurIPS and EUSIPCO, . His survey paper on audio-language datasets has been published in IEEE Access.

In January 2026, Gijs started a position at the Scalable Learning and Multi-Purpose AI Lab (SLAMPAI) at the Juelich Supercomputing Center (JSC), Helmholtz Research Center Juelich (FZJ), in Cologne, Germany. There he works on large-scale training of audio-language models, including pretraining and collecting audio data, deriving scaling laws, and evaluating these models, while continuing to publish at top-tier conferences. In May 2026, he joined Microsoft Research in Redmond, WA, USA, as a summer research intern, working on spatial audio-language models.

Gijs is committed to advancing the field of audio-language learning by developing more effective models and datasets. He believes that model training and the development of new modeling techniques form only a small part of what matters: understanding which data are used to train the models, how the models are evaluated, and ensuring that the results are reproducible and transparent are just as important.

Awards & Grants

- **2026** - Google Gemini Pro Research Grant - Grant: \$10K
- **2025** - NeurIPS Scholar Award - Travel Grant
- **2022-2025** - Three Dutch Research Council Compute grants for PhD research. Total: €250K
- **2024** - OpenAI Research Grant for work on AudSemThinker. Grant: \$1000
- **2022** - DCASE Travel Grant
- **2021** - 2nd Place in the Course-wide Kaggle Challenge for Image Classification (out of 180+ students)
- **2019** - Kairos Society Fellow - Fellow of a global community of top students and global leaders who aim to solve the world's greatest challenges
- **2018** - Winner of the Xomnia Datathon
- **2018** - AWS Cloud Practitioner Essentials Certified
- **2018** - Dutch Journalism Fund (Stimuleringsfonds voor de Journalistiek). Grant: €75K
- **2017-2022** - T500 Member - Member of The Next Web's list of the most ambitious entrepreneurs under the age of 26
- **2017** - Winner of the ADEhack Hackathon

Data and Semantic Reasoning in Audio-Language Learning

- 01 Curating training and evaluation data is just as important for the training of machine learning models as architectural innovation. THIS THESIS
- 02 Zero-shot evaluation in audio-language learning must contain no overlap between its training and evaluation dataset; otherwise it should be called *weak* zero-shot evaluation. THIS THESIS
- 03 Reasoning with human-inspired semantic descriptors improves the performance of audio-language models. THIS THESIS
- 04 Evaluating automated audio captioning models on semantic content results in better quality judgments than relying on word overlap metrics alone. THIS THESIS
- 05 Audio-language models that can listen, reason, invoke tools, and use the web will be seen more in future systems. THIS THESIS
- 06 A modular agent that coordinates specialized audio and speech tools can match or surpass audio-language models in performance. THIS THESIS
- 07 Scaling laws will hold true for a long time, and model performance will continue to improve with data size. THIS THESIS
- 08 We are moving from the age of scaling to the age of research. – *Ilya Sutskever*
- 09 To get all the way to something like AGI may just require one or two more new breakthroughs. – *Demis Hassabis*
- 10 Artificial intelligence requires not just artificial intelligence.

