

MECAT: A Multi-Experts Constructed Benchmark for Fine-Grained Audio Understanding Tasks

Yadong Niu^{1*}, Tianzi Wang^{1,2*}, Heinrich Dinkel¹, Xingwei Sun¹, Jiahao Zhou¹, Gang Li¹, Jizhong Liu¹, Xunying Liu¹, Junbo Zhang¹, Jian Luan¹

¹ MiLM Plus, Xiaomi Inc, Beijing, China

² The Chinese University of Hong Kong, Hong Kong, China

Abstract

While large audio-language models have advanced open-ended audio understanding, they still fall short of nuanced human-level comprehension. This gap persists largely because current benchmarks, limited by data annotations and evaluation metrics, fail to reliably distinguish between generic and highly detailed model outputs. To this end, this work introduces MECAT, a Multi-Expert Constructed Benchmark for Fine-Grained Audio Understanding Tasks. Generated via a pipeline that integrates analysis from specialized expert models with Chain-of-Thought large language model reasoning, MECAT provides multi-perspective, fine-grained captions and open-set question-answering pairs. The benchmark is complemented by a novel metric: **DATE** (Discriminative-Enhanced Audio Text Evaluation). This metric penalizes generic terms and rewards detailed descriptions by combining single-sample semantic similarity with cross-sample discriminability. A comprehensive evaluation of state-of-the-art audio models is also presented, providing new insights into their current capabilities and limitations. The data and code are available at <https://github.com/xiaomi-research/mecat>.

Introduction

The human auditory system is highly effective at processing complex acoustic scenes. It can distinguish subtle variations in sound, such as telling a dog’s playful bark from a defensive growl (Plack 2023), and isolate target speech from noisy backgrounds, an ability known as the cocktail party effect.

A central goal of machine hearing is to replicate this auditory intelligence to interpret raw audio signals as semantically rich perception (Lyon 2017). Early works in machine hearing focused on closed-ended tasks such as sound event classification and automatic speech recognition. Large language models (LLM) have spurred the development of large audio-language models (LALMs), which have driven a shift towards more general open-ended tasks like audio captioning and audio question answering (Chu et al. 2023; Du et al. 2023; Hu et al. 2024; Shu et al. 2023; Wang et al. 2023; Tang et al. 2024; Rubenstein et al. 2023; Chen et al. 2023; Huang et al. 2024).

Despite these advances, current LALMs still fall short of achieving the comprehensive understanding that character-

izes human hearing (Sakshi et al. 2025). This work argues that despite ongoing improvements in model architectures and data, a crucial and often-overlooked bottleneck is the existing evaluation benchmark.

The second challenge is rooted in evaluation metrics. Traditional lexical-matching metrics, on the one hand, penalizes semantically correct but lexically different descriptions. Embedding-based metrics, on the other hand, better align with human perception, they often fail to distinguish between generic, vague captions and highly detailed, accurate ones. Even the more recent LLM-as-judge method, while demonstrating strong discriminative ability, is often hindered by practical constraints such as high costs and slow inference speeds, as well as its high dependency on model selection and prompt design.

Current benchmarks inadequately evaluate audio understanding, as they often reward generic captions (e.g., A dog is barking and people are talking) for distinct scenarios (e.g., an excited bark in a park vs. a defensive bark during an argument). This limits their ability to differentiate between models with true perceptual accuracy and those producing vague outputs.

To this end, we introduce MECAT, a Multi-Expert Constructed Benchmark for Fine-Grained Audio Understanding Tasks. By integrating analysis from a series of specialized audio-related experts models, including content-specific models (e.g., for speech, music, and sound events) and content-unrelated models (e.g., for audio quality, reverberation and intensity), followed by Chain-of-Thought (CoT) enhanced LLM reasoning (Guo et al. 2025), MECAT provide fine-grained captions alongside open-set question-answering pairs. The captions primarily focus on providing a comprehensive, multi-perspective description of the acoustic scene, while the QA pairs are designed to probe for specific details and higher-level contextual reasoning that descriptive tasks alone cannot fully assess. Furthermore, we introduce a novel metric **DATE** (Discriminative-Enhanced Audio Text Evaluation), which is designed to better quantify the detail and accuracy of model’s response. It uniquely combines a weighted single-sample semantic similarity that penalizes generic terms while emphasizing discriminative phrases, and a cross-sample discriminability score that explicitly rewards the model’s responses for exceeding general descriptions. This design enables DATE to robustly distin-

*These authors contributed equally.

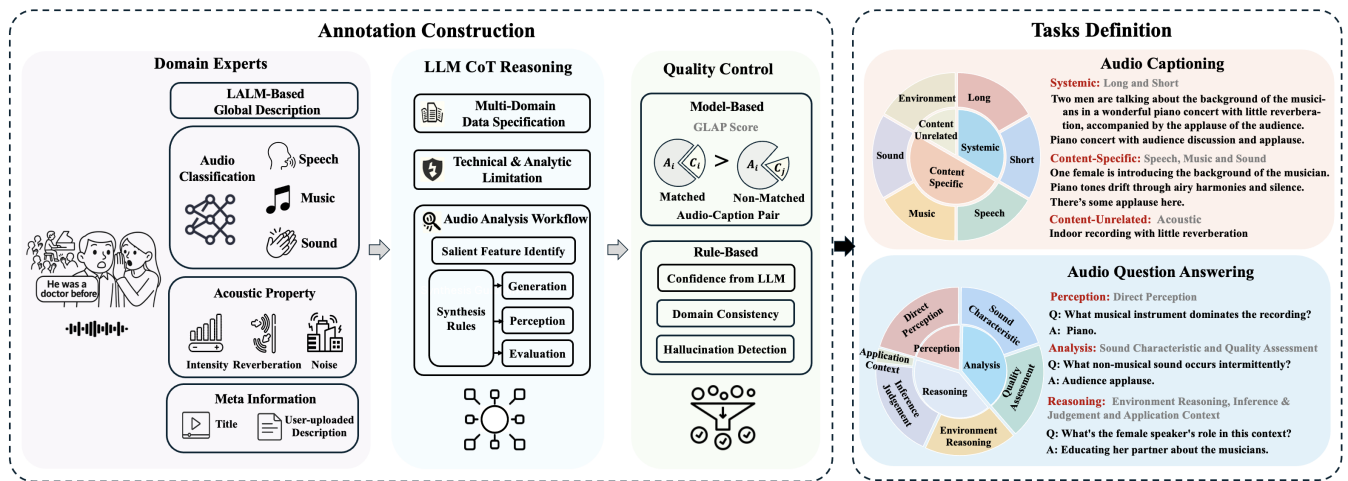


Figure 1: Overview of the MECAT Benchmark.

guish between superficial and context-rich model outputs.

Related works

Audio Captioning Benchmark

Audio captioning benchmarks have been pivotal in advancing audio understanding works (Wu, Dinkel, and Yu 2019; Kim et al. 2019; Drossos, Lipping, and Virtanen 2020; Yuan et al. 2025; Manco et al. 2023; Liu et al. 2024a,b). Early dataset like AudioCaps (Kim et al. 2019) and Clotho (Drossos, Lipping, and Virtanen 2020) primarily relied on manual annotation, where human annotators provide one or more captions for each audio clip. While foundational, these benchmarks face a critical limitation: the coarse-grained nature of their annotations. During the annotation process, human annotators often produce generic, events-level descriptions rather than capturing the nuanced acoustic details of a scene. This results in a gold standard that lacks the specificity needed to evaluate fine-grained understanding.

While newer methods using LLMs for automatic labeling, such as in AutoACD (Sun et al. 2024) and LPMusicCaps (Doh and Nam 2023), have improved scalability, they did not solve the granularity problem. Caption quality suffers from coarse input metadata like titles and tags, perpetuating generic descriptions.

Audio Question-Answering Benchmark

Audio Question Answering (QA) presents a more targeted evaluation of a model’s audio understanding abilities (Lipping et al. 2022; Wang et al. 2025; Li et al. 2022; Sakshi et al. 2025). Datasets like ClothoAQA (Lipping et al. 2022) and MusicAVQA (Li et al. 2022) have been developed with manually crafted question-answer pairs. However, similar to captioning benchmarks, they suffer from limitations that hinder the assessment of detailed understanding.

The main issue is their reliance on close-ended answer formats designed for easier automatic scoring. For example,

many questions in ClothoAQA are limited to “yes/no” answers (Lipping et al. 2022), while other benchmarks like MMAU (Sakshi et al. 2025) utilize a multiple-choice format. While convenient for evaluation, these formats prevent the assessment of a model’s ability to generate detailed, descriptive answers and may encourage models to learn shallow pattern matching rather than deep understanding.

Evaluation Metrics for Audio Caption and QA

The evaluation of open-ended audio caption and QA is critically dependent on the choice of metric. However, existing metrics fail to adequately assess the fine-grained descriptive capabilities of modern generative models.

Traditional metrics, such as BLEU (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016), operate by measuring lexical overlap with reference texts. This reliance on n-gram matching unfairly penalizes novel yet accurate descriptions that do not share the exact wording of the references.

To overcome the limitations of lexical matching, embedding similarity-based metrics were introduced. Approaches like FENSE (Zhang et al. 2022) were specifically designed for audio captioning. However, our experiments found that they still struggle to effectively distinguish between a generic, vague response and a highly detailed and accurate one. More recently, LLM-as-judge has been adopted for evaluating open-ended generation (Wang et al. 2025; Zheng et al. 2023). These methods show strong correlation with human judgment and, as our experiments confirmed, possess a high sensitivity to response specificity. However, LLM-as-judge suffer from practical limitations, such as high computational costs, slow evaluation speeds, and a strong dependency on the choice of the LLM and the design of the prompt (Lee, Park, and Kang 2024; Zheng et al. 2023).

MECAT Benchmark Overview

As illustrated in Figure 1, MECAT is a comprehensive benchmark for fine-grained audio understanding, distin-

Task	Labeling	Dataset	Test Size	Domain	Source
Caption	Manual	AudioCaps (Kim et al. 2019)	~1.0k	Multi-Domain ^{‡,◊}	AudioSet
		Clotho (Drossos, Lipping, and Virtanen 2020)	~1.0k	Multi-Domain ^{‡,◊}	Clotho
		SongDescriber (Manco et al. 2023)	0.7k	Music	MTG-Jamendo
	LLM	AudioCaps-Enhanced (Yuan et al. 2025)	0.9k	Multi-Domain ^{‡,◊}	AudioSet
		AutoACD (Sun et al. 2024)	1.0k	Multi-Domain ^{‡,◊}	AudioSet
		LPMusicCaps-MSD (Doh and Nam 2023)	35k	Music	Song Dataset
		LPMusicCaps-MTT (Doh and Nam 2023)	5k	Music	MagnaTagATune
	MP-LLM [†]	MECAT-Caption (Ours)	20k	Extended Multi-Domain [§]	ACAV100M
QA	Manual	ClothoAQA (Lipping et al. 2022)	2k	Multi-Domain ^{‡,◊}	Clotho
		WavCaps-QA (Wang et al. 2025)	0.3k	Multi-Domain ^{‡,◊}	AudioSet and 2 others
		MusicAVQA (Li et al. 2022)	6k	Music	YouTube
		Audiocaps-QA (Wang et al. 2025)	0.3k	Multi-Domain ^{‡,◊}	AudioSet
		MMAU (Sakshi et al. 2025)	10k	Multi-Domain [‡]	AudioSet and 12 others
	LLM	EvalSIFT (Pandey et al. 2025)	30k	Speech	Open-source ASR
	MP-LLM [†]	MECAT-QA (Ours)	20k	Extended Multi-Domain [§]	ACAV100M

Table 1: Comparison of MECAT with Recent General Sound Evaluation Benchmark Datasets. [†] MP-LLM: Multiple Experts Models and LLM; [‡] Multi-Domain: This includes speech, music and sound-events ([◊] denotes that domain were not elaborated in detail); [§] Extended Multi-Domain: This includes speech, music, sound-events, combinations thereof, and silence.

guished by its unique data sources, broad domain coverage, and two core evaluation tasks: MECAT-Caption and MECAT-QA.

Dataset Description

To ensure data source novelty, MECAT is constructed from a carefully selected subset of ACAV100M (Lee et al. 2021). This approach contrasts with benchmarks, such as AudioCaps (Kim et al. 2019), Clotho (Drossos, Lipping, and Virtanen 2020), and WavCaps-QA (Wang et al. 2025), which predominantly draw from a limited pool of sources such as AudioSet (Gemmeke et al. 2017) and Clotho (Drossos, Lipping, and Virtanen 2020) (see Table 1). The dataset comprises approximately 20,000 Creative Commons-licensed audio clips, each with a maximum duration of 10 seconds.

Based on this unique data foundation, MECAT encompasses eight distinct audio domains designed to comprehensively represent real-world acoustic scenarios. These categories include four *Pure* domains: silence (000), speech (S00), sound events (00A), and music (0M0), as well as all four possible combinations of *Mixed* domains that reflect the complexity of natural auditory environments (SM0, S0A, 0MA, and SMA). This extended multi-domain coverage, with its distribution detailed in Figure 2, enables a nuanced evaluation of models on complex acoustic scenes, such as those that combine piano music with spoken discussion and audience applause.

Tasks Definition

As illustrated in Figure 1, the MECAT-Caption task delivers multi-perspective annotations for comprehensive evaluation. Each audio clip is annotated with a rich set of captions organized into three categories, which together comprise six distinct sub-categories. The first category, *Systemic Captions*, consists of two sub-categories: a concise short cap-

tion focused on primary audio content and a detailed long caption encompassing contextual details and event interactions. The second category, *Content-Specific Captions*, includes three sub-categories for the independent analysis of speech, music, and sound events. Crucially, to assess model performance across different levels of acoustic complexity, the evaluation for each content type is performed on corresponding pure domains (e.g., pure speech - S00) and all mixed domains. Notably, these captions also explicitly state when a corresponding domain is absent. The final category is a single *Content-Unrelated Caption* that focuses exclusively on acoustic characteristics like audio quality and reverberation. For each of these six sub-categories, three synonymous reference captions are provided, yielding a total of 18 reference captions per clip and creating a significantly richer vocabulary than existing datasets (see Appendix A for more details).

The final score $\text{Score}_{\text{Cap}}$ for the MECAT-Caption task is calculated as a weighted average of the three main categories:

$$\text{Score}_{\text{Cap}} = 0.4 \cdot S_{\text{Systemic}} + 0.4 \cdot S_{\text{Content-Specific}} + 0.2 \cdot S_{\text{Content-Unrelated}},$$

where the category scores are themselves weighted sums of their sub-categories:

$$S_{\text{Systemic}} = 0.8 \cdot S_{\text{Long}} + 0.2 \cdot S_{\text{Short}},$$

$$S_{\text{Content-Specific}} = 0.6 \cdot S_{\text{Speech}} + 0.3 \cdot S_{\text{Music}} + 0.1 \cdot S_{\text{Sound}}.$$

The score for each content type (S_{Speech} , S_{Music} , S_{Sound}) is calculated as the unweighted mean of its performance on the corresponding pure domains (e.g., S00, 0M0, 00A) and all mixed domains.

Complementing the captioning task, MECAT-QA facilitates evaluation through targeted, probing questions. Each audio clip is paired with five question-answer pairs that span

different cognitive skills, resulting in over 100,000 QA pairs in total. These pairs are organized into three cognitive categories. The first, *Perception*, focuses on the direct identification of audio content through its Direct Perception (DP) sub-category. The second, *Analysis*, delves into acoustic properties via two sub-categories: Sound Characteristics (SC), for examining properties like pitch, and Quality Assessment (QAS), for evaluating technical fidelity. The final and most complex category, *Reasoning*, targets higher-level cognitive skills through three sub-categories: Environment Reasoning (ER), requiring acoustic scene inference; Inference & Judgement (IJ), involving logical deductions; and Application Context (AC), testing the understanding of practical scenarios.

The scoring for MECAT-QA is designed to ensure equal contribution from each cognitive skill. The overall score is the unweighted arithmetic mean of the scores from all six individual sub-categories (DP, SC, QAS, ER, IJ, and AC):

$$\text{Score}_{\text{QA}} = \frac{S_{\text{DP}} + S_{\text{SC}} + S_{\text{QAS}} + S_{\text{ER}} + S_{\text{IJ}} + S_{\text{AC}}}{6}.$$

Annotation Construction

This section details the MECAT annotation construction pipeline. As illustrated in Figure 1, the process starts with a audio classification stage identifying the domain of each audio clip. Based on the resulting domains, the clip is then processed by a series of specialized expert models.

The structured outputs from these experts are subsequently synthesized using LLM CoT reasoning to generate fine-grained captions and open-set QA pairs. The pipeline concludes with a rigorous quality control stage to ensure the reliability of all final annotations. The complete list of the used models is available in Appendix B.

Domain Experts For each audio clip, we first use Audio Flamingo 2 (Ghosh et al. 2025) to generate a global, event-level summarization in natural language. Furthermore, we apply a series of domain expert models for more detailed analysis.

Audio Classification For each audio clip, we use CED-Base (Dinkel et al. 2024a) to predict AudioSet (Gemmeke

et al. 2017) labels for every 2-second, non-overlapping interval. This process results in a sequence of multi-label predictions for each clip. Based on the CED prediction, we categorize each clip into one of eight distinct domains: 000, 00A, 0M0, S00, SM0, 0MA, S0A, SMA, as detailed in Dataset Description Section.

Speech-focused Analysis For speech-domain clips (S00, S0A, SM0, SMA), we employ a speech-focused analysis pipeline (Figure 3-I). The pipeline consists of automatic speech recognition, language identification, and speaker diarization. Using the temporal boundaries from diarization, we extract each speaker’s attributes, including gender, age, emotion, and English accent. The probabilities of these results are also utilized for subsequent LLM reasoning.

Music-focused Analysis For music-domain clips (0M0, SM0, 0MA, SMA), a music-focused analysis pipeline is employed (Figure 3-II). It consists of LALM-based global description of music content (Audio Flamingo 2 (Ghosh et al. 2025)), musical attribute analysis, and music separation. Musical attribute analysis provides a series of perceptual and technical attributes such as emotions and tempo. The music separation module isolates vocal tracks from the instrumental background, which are then routed to the speech analysis pipeline.

Sound Events-focused Analysis For audios in 00A, we directly utilize the events labels predicted by the CED-Base model during the audio classification stage.

Acoustic Properties Analysis To extract fundamental signal characteristics and assess the recording environment, we apply a universal acoustic property analysis pipeline to all audio clips (Figure 3-III). The analysis content includes signal intensity, speech quality assessment, and reverberation. Signal intensity is quantified via Root Mean Square (RMS). For audio quality, we conduct both DNSMOS (Reddy 2021) and NISQA2 (Mittag et al. 2021) assessments to measure signal distortion, background noise, and perceptual quality. We also characterize the acoustic environment by estimating the reverberation time of the recording space.

LLM CoT Reasoning Our pipeline employs a Chain-of-Thought (CoT) guided LLM (Deepseek-R1: Guo et al., 2025) to synthesize a set of rich annotations. The model is instructed to reason over the outputs from all preceding analyses and the metadata. This reasoning process weighs evidence from various sources to resolve inconsistencies and identifying salient features. The final output consists of captions and corresponding QA pairs, where each item is annotated with a confidence level. The complete prompt is shown in Appendix C.

Quality Control The model-based filtering use GLAP (Dinkel et al. 2025) to compute the cosine similarity between audio clip and its systemic long caption embeddings. A sample is kept only if the similarity of its correct audio-caption pair exceeds its average similarity with a set of 6 other randomly selected captions by an empirically set threshold of 6.

We further apply rule-based filtering including LLM confidence thresholding, domain consistency between au-

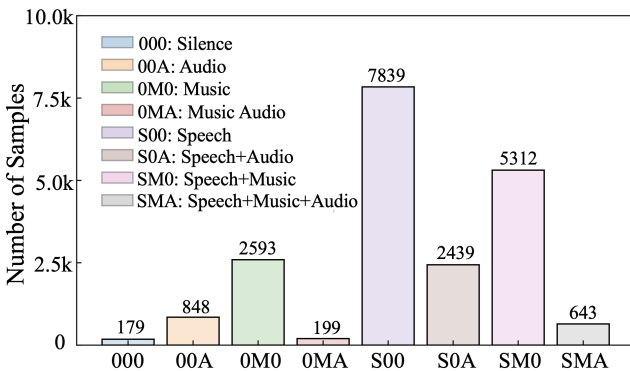


Figure 2: Distribution of audio samples across extended multi-domains in the MECAT, including speech, music, audio, combinations thereof, and silence.

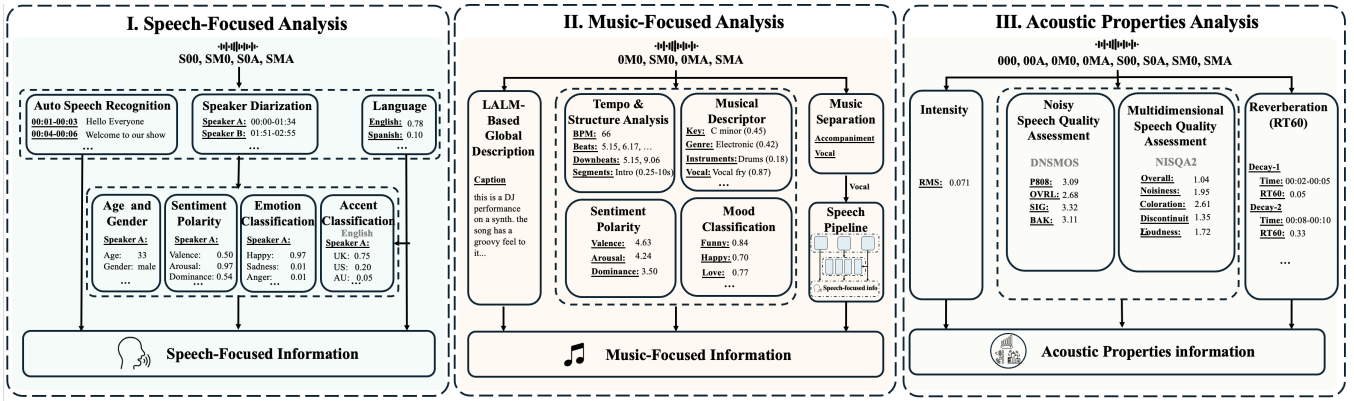


Figure 3: Domain Experts for Speech, Music, and Acoustic Properties.

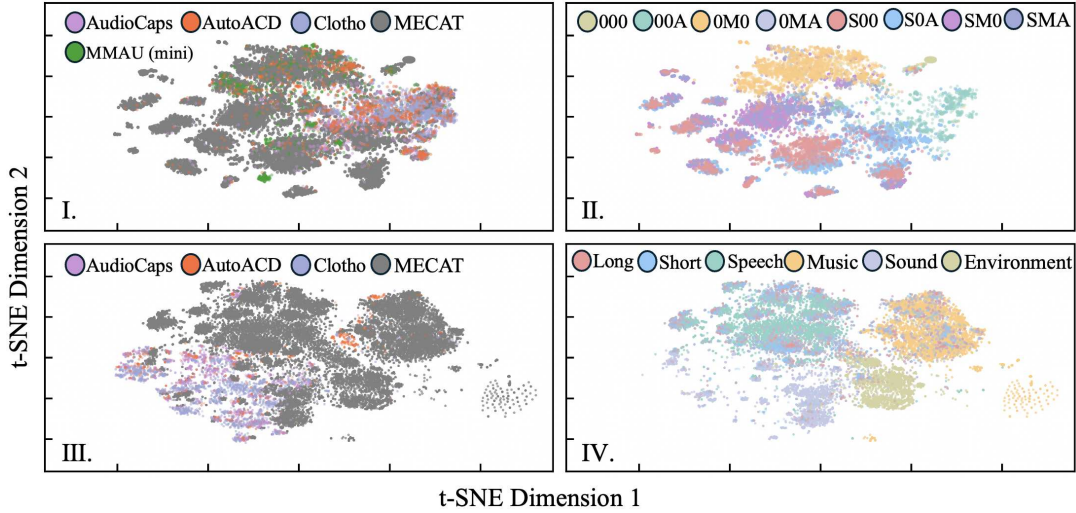


Figure 4: t-SNE plots of MECAT audio embeddings compared to other benchmarks I), further clustered by domain II). Caption embeddings are visualized in III) and clustered by categories in IV). Audio embeddings and captions embeddings are extracted from Dasheng-Base (Dinkel et al. 2024b) and Sentence-BERT (Reimers and Gurevych 2019) respectively.

dio classification and LLM output, and hallucination removal (Barański et al. 2025).

Metric Design

Existing evaluation metrics demonstrate significant limitations when evaluating fine-grained, detailed descriptions. To address this, we propose DATE, a metric built on Sentence-BERT (Reimers and Gurevych 2019) that improves semantic assessment by combining single-sample semantic similarity and cross-sample discriminability score.

Single-Sample Semantic Similarity We apply Term Frequency-Inverse Document Frequency (tf-idf) weighting to token embeddings from Sentence-BERT to emphasize tokens that are frequent within a single sample but rare across the dataset. The weighted embedding vector \mathbf{v}_T for a given sentence T is computed as:

$$\mathbf{v}_T = \sum_{t \in T} (\text{tf}(t, T) \cdot \text{idf}(t)) \cdot E(t),$$

where t is a token in T . The term $\text{tf}(t, T)$, $\text{idf}(t)$, and $E(t)$ are the frequency, inverse document frequency, and the Sentence-BERT embedding, respectively. The single-sample semantic similarity, $S_{\text{sim}, i}$, is the cosine similarity between the weighted embeddings of the candidate and reference text. $S_{\text{sim}, i} = (\mathbf{v}_{\text{cand}} \cdot \mathbf{v}_{\text{ref}}) / (\|\mathbf{v}_{\text{cand}}\| \|\mathbf{v}_{\text{ref}}\|)$.

Cross-Sample Discriminability An ideal description should be clearly distinguishable from descriptions of other audio samples. We construct a cross-sample similarity matrix, \mathcal{M} , where each element $M_{i,j}$ is the score between the reference description for audio i and the candidate description for audio j . For each sample i , we rank the correctly matched score $M_{i,i}$ against all candidate scores $\{M_{i,j}\}_{j=1}^N$. Denoting this rank as r_i , the discriminability score is defined as:

$$S_{\text{dis}, i} = 1 - \frac{r_i}{N}.$$

This rewards candidates that rank highly for their correct reference, approaching 1 for top ranks and 0 for bottom ranks.

DATE To ensure a balanced evaluation for both descriptive accuracy and uniqueness, the DATE score for each sample DATE_i is defined as the harmonic mean of its semantic similarity ($S_{\text{sim},i}$) and discriminability ($S_{\text{dis},i}$):

$$\text{DATE}_i = \frac{2 \cdot S_{\text{sim},i} \cdot S_{\text{dis},i}}{S_{\text{sim},i} + S_{\text{dis},i}} \in [0, 1].$$

The DATE score of a dataset is computed as $\text{DATE} = \frac{1}{N} \sum_{i=1}^N \text{DATE}_i$.

Result and Discussion

This section presents an analysis of MECAT’s data diversity, the analysis of the DATE metric, and a comprehensive evaluation of state-of-the-art models on MECAT.

Data Diversity Analysis

The T-distributed stochastic neighbor embedding (t-SNE) in Figure 4 reveals two key findings:

First, the audio in MECAT exhibits both broad external coverage and a well-structured internal distribution. As shown in Figure 4-I and Figure 4-II, embeddings from other benchmarks cluster densely, aligning primarily with MECAT’s sound-event domain. In contrast, MECAT’s embeddings are widely distributed across the feature space. Internally, the pure domains (S00, 0M0, and 00A) form distinct and well-separated clusters, while other mixed domains occupy the intermediate spaces. Second, as shown in Figure 4-III, the caption embeddings span a much diverse space

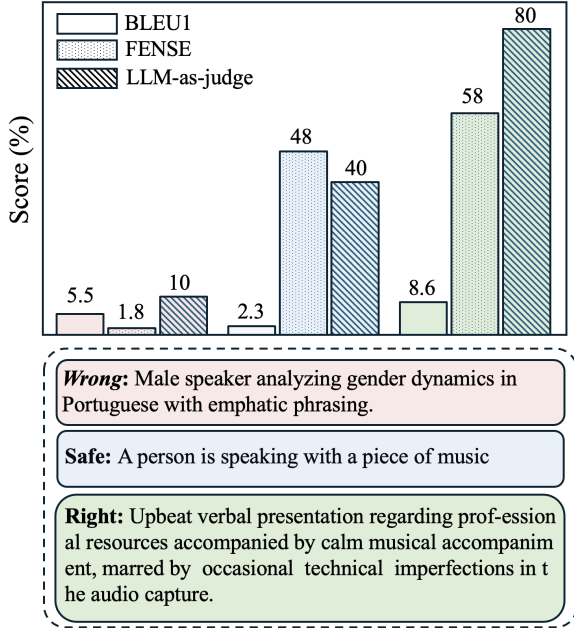


Figure 5: Example predictions from different models on MECAT-Caption dataset. Reference: “An animated woman’s voice shares information about learning materials while melodic instruments play quietly underneath, with persistent low-quality artifacts in the recording.”

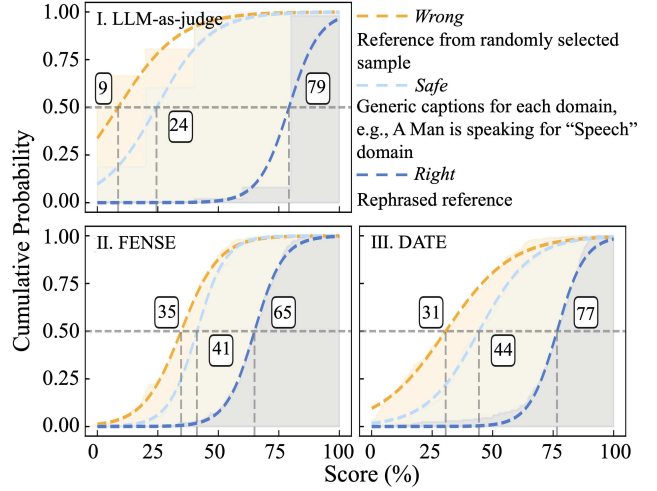


Figure 6: Cumulative Distribution Functions (CDF) of LLM-as-judge, FENSE, and DATE on Caption (left) and QA (right). Larger distances between CDF curves indicate better discriminative ability of the metric.

than other datasets. This diversity is driven by MECAT’s distinct caption categories. Specifically, speech, music and sound events focused captions and content-unrelated captions form separate clusters. In contrast, systematic long and short captions are dispersed throughout the entire space (see Figure 4-IV).

Metric Analysis

This section validates the proposed metric, DATE. We conduct a comparative analysis of DATE against the strong baseline FENSE, using the LLM-as-judge method as an upper-bound performance reference. The prompts used in LLM-as-judge method and the reliability analysis of its score could be found in Appendix D and E.

The case study in Figure 5 reveals key issues of existing metrics: lexical based metrics like BLEU-1 are semantically unreliable (score of Wrong > Safe). While FENSE improves the score, it struggles to differentiate high-quality (Right) from vague (Safe) captions with only small score gap of 10, which stands in sharp contrast to the clear separation provided by the LLM-as-judge score.

Figure 6 plots the CDF curves for each metric, larger distances between CDF curves indicate better discriminative ability of the metric. It can be found that DATE demonstrates a superior discriminative ability over FENSE, which is evident in its larger median score spans on both the Right vs. Wrong (DATE: 46 vs. FENSE: 30) and the Right vs. Safe (DATE: 33 vs. FENSE: 24).

Model Performance on MECAT

We evaluated a range of publicly available models with distinct specializations on the MECAT-Caption task (Table 2): the traditional caption-only models EnClap (Kim et al. 2024) and Pengi (Deshmukh et al. 2023), the speech-focused Kimi-Audio (7B) (KimiTeam et al. 2025), the non-

Type	Model	Systemic		Content-Specific						Content Unrelated	Score _{Cap}
		Long	Short	Speech		Music		Sound			
				Pure	Mixed	Pure	Mixed	Pure	Mixed	Env	
Caption-Only	Pengi	43.5	46.8	27.2	29.5	29.3	13.1	42.8	14.6	7.1	30.6
	EnClap	48.6	53.1	30.2	31.8	17.9	15.9	48.8	15.2	6.8	33.3
LALM	Kimi-Audio	49.5	54.2	30.0	31.3	27.7	16.9	43.1	16.2	7.0	34.3
	Audio Flamingo 2	48.6	49.7	30.5	34.3	28.8	25.6	41.2	18.5	17.5	35.6
	Qwen2.5-Omni 3B	56.4	55.2	42.5	41.3	46.6	29.7	52.9	23.9	19.4	42.6
	Qwen2.5-Omni 7B	61.1	56.5	39.9	40.9	32.1	30.9	50.7	23.8	17.9	43.0

Table 2: Model performance (DATE %) on MECAT-Caption.

speech-focused Audio Flamingo 2 (3B) (Ghosh et al. 2025), and the general-purpose Qwen2.5-Omni (3B and 7B) (Xu et al. 2025). The prompts used in LALM could be found in Appendix F. Several key trends can be found:

i) A clear performance hierarchy exists (Table 2, Overall), with LALMs significantly outperforming classic models. Among LALMs, the general-purpose Qwen-Omni series leads, driven by their strong perception and domain-focused capabilities.

ii) The ability to generate long, detailed captions is a key differentiator for advanced models (e.g., comparing Qwen-7B with EnClap, the performance gap shrinks from 12.5 on Long to 3.4 on Short).

iii) Domain complexity presents a challenge on non-speech-focused domains, with all models performing worse on mixed domains than on the pure ones.

iv) All models exhibit a limited capability on the Content-Unrelated task (ranging from 6.8 to 19.4), which indicates a strong bias towards event recognition over describing acoustic properties.

Sub-Category	Kimi-Audio	Audio Flamingo 2	Qwen2.5-Omni 3B	Qwen2.5-Omni 7B
DP [†]	45.6	45.1	55.7	57.8
SC [‡]	39.2	46.3	53.2	52.9
QAS [‡]	18.7	34.9	38.6	39.1
ER [§]	34.6	37.5	41.1	44.0
IJ [§]	48.9	44.0	51.8	53.2
AC [§]	41.2	42.4	50.8	50.8
Score _{QA}	38.0	41.7	48.5	49.6

Table 3: Model Performance (DATE %) on MECAT-QA. Category: [†]Perception; [‡]Analysis; [§]Reasoning. Sub-Category (DP: Direct Perception; SC: Sound Characteristics; QAS: Quality Assessment; ER: Environment Reasoning; IJ: Inference & Judgment; AC: Application Context).

On the MECAT-QA task, we only evaluated the LALMs (Table 3). Several key insights can be found:

i) The overall performance hierarchy mirrors the captioning task, with the general-purpose Qwen-Omni models

again achieving the top scores;

ii) All models score highest on Direct Perception tasks, which rely on identifying explicit events. In contrast, performance is lower on most Analysis and Reasoning tasks. This suggests that while current models can perceive primary audio events, with less capable when tasks demand a underlying acoustic characteristics or higher-level reasoning.

iii) The fine-grained breakdown reveals model specializations. For instance, the speech-focused Kimi-Audio struggles with Quality Assessment (18.7), whereas the non-speech-focused Audio Flamingo 2 performs notably better on the same task (34.9). This contrast highlights the MECAT’s ability to diagnose these underlying model biases.

Besides, our evaluation reveals two overarching findings: First, model performance does not scale linearly with parameter count, indicating that architectural innovation and training strategy are more critical. The Qwen2.5-Omni 3B model performs nearly on par with the 7B version (e.g., QA: 48.5 vs. 49.6), and it also outperforms the larger 7B speech-focused Kimi-Audio. Second, the accuracy of even the best-performing models on many of MECAT’s fine-grained subtasks is low, ranging from 20-60%. This indicates a substantial gap remains between current LALMs and nuanced, human-level audio understanding.

Conclusion and Future Work

In this work, we introduce MECAT, a Multi-Experts Constructed Benchmark for Fine-Grained Audio Understanding Tasks. Through the integration of specialized expert audio models and LLM with CoT reasoning, MECAT constructs a large-scale, multi-perspective dataset covering both Audio Captioning and Audio Question-Answering tasks. The benchmark is complemented by a novel metric DATE, which is designed to penalize generic terms and reward detailed, discriminative descriptions. Our comprehensive evaluation of state-of-the-art audio models reveals significant challenges, with even the best-performing models achieving only 20-60 score across different subtasks.

Future work should build upon MECAT by addressing its current limitations, such as incorporating a deeper analysis of fundamental acoustic properties and expanding beyond single audio module, to guide the development of more robust audio AI systems.

References

- Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European conference on computer vision*, 382–398. Springer.
- Barański, M.; Jasiński, J.; Bartolewska, J.; Kacprzak, S.; Witkowski, M.; and Kowalczyk, K. 2025. Investigation of whisper asr hallucinations induced by non-speech audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Chen, F.; Han, M.; Zhao, H.; Zhang, Q.; Shi, J.; Xu, S.; and Xu, B. 2023. X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages. *arXiv preprint arXiv:2305.04160*.
- Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *arXiv preprint arXiv:2311.07919*.
- Deshmukh, S.; Elizalde, B.; Singh, R.; and Wang, H. 2023. Pengi: An Audio Language Model for Audio Tasks. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 18090–18108. Curran Associates, Inc.
- Dinkel, H.; Wang, Y.; Yan, Z.; Zhang, J.; and Wang, Y. 2024a. CED: Consistent ensemble distillation for audio tagging. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 291–295. IEEE.
- Dinkel, H.; Yan, Z.; Wang, T.; Wang, Y.; Sun, X.; Niu, Y.; Liu, J.; Li, G.; Zhang, J.; and Luan, J. 2025. GLAP: General contrastive audio-text pretraining across domains and languages. *arXiv:arXiv preprint arXiv:2506.11350*.
- Dinkel, H.; Yan, Z.; Wang, Y.; Zhang, J.; Wang, Y.; and Wang, B. 2024b. Scaling up masked audio encoder learning for general audio classification. In *Proceedings of the 25th Interspeech Conference (interspeech)*, 547–551.
- Doh, S.; and Nam, J. 2023. LP-MusicCaps: LLM-Based Pseudo Music Captioning. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*. International Society for Music Information Retrieval Conference.
- Drossos, K.; Lipping, S.; and Virtanen, T. 2020. Clotho: an Audio Captioning Dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 736–740. IEEE.
- Du, Z.; Wang, J.; Chen, Q.; Chu, Y.; Gao, Z.; Li, Z.; Hu, K.; Zhou, X.; Xu, J.; Ma, Z.; et al. 2023. LauraGPT: Listen, Attend, Understand, and Regenerate Audio with GPT. *arXiv preprint arXiv:2310.04673*.
- Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An Ontology and Human-labeled Dataset for Audio Events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. IEEE.
- Ghosh, S.; Kong, Z.; Kumar, S.; Sakshi, S.; Kim, J.; Ping, W.; Valle, R.; Manocha, D.; and Catanzaro, B. 2025. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 1–48.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Hu, S.; Zhou, L.; Liu, S.; Chen, S.; Meng, L.; Hao, H.; Pan, J.; Liu, X.; Li, J.; Sivasankaran, S.; et al. 2024. WavLLM: Towards Robust and Adaptive Speech Large Language Model. In *Proceedings of the Findings of the Association for Computational Linguistics (EMNLP)*, 4552–4572.
- Huang, R.; Li, M.; Yang, D.; Shi, J.; Chang, X.; Ye, Z.; Wu, Y.; Hong, Z.; Huang, J.; Liu, J.; et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23802–23804.
- Kim, C. D.; Kim, B.; Lee, H.; and Kim, G. 2019. Audio-caps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 119–132.
- Kim, J.; Jung, J.; Lee, J.; and Woo, S. H. 2024. 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)*, 6735–6739.
- KimiTeam; Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; Tang, H.; Wang, Z.; Wei, C.; Xin, Y.; Xu, X.; Yu, J.; Zhang, Y.; Zhou, X.; Charles, Y.; Chen, J.; Chen, Y.; Du, Y.; He, W.; Hu, Z.; Lai, G.; Li, Q.; Liu, Y.; Sun, W.; Wang, J.; Wang, Y.; Wu, Y.; Wu, Y.; Yang, D.; Yang, H.; Yang, Y.; Yang, Z.; Yin, A.; Yuan, R.; Zhang, Y.; and Zhou, Z. 2025. Kimi-Audio Technical Report. *arXiv:2504.18425*.
- Lee, S.; Chung, J.; Yu, Y.; Kim, G.; Breuel, T.; Chechik, G.; and Song, Y. 2021. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10274–10284.
- Lee, Y.; Park, I.; and Kang, M. 2024. FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3732–3746.
- Li, G.; Wei, Y.; Tian, Y.; Xu, C.; Wen, J.-R.; and Hu, D. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19108–19118.
- Lipping, S.; Sudarsanam, P.; Drossos, K.; and Virtanen, T. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, 1140–1144. IEEE.

- Liu, J.; Li, G.; Zhang, J.; Dinkel, H.; Wang, Y.; Yan, Z.; Wang, Y.; and Wang, B. 2024a. Enhancing Automated Audio Captioning via Large Language Models with Optimized Audio Encoding. In *Proceedings of the 25th Interspeech Conference (interspeech)*, 1135–1139.
- Liu, J.; Li, G.; Zhang, J.; Liu, C.; Dinkel, H.; Wang, Y.; Yan, Z.; Wang, Y.; and Wang, B. 2024b. Leveraging ced encoder and large language models for automated audio captioning. *Proceedings of the DCASE Challenge*, 1–4.
- Lyon, R. F. 2017. *Human and machine hearing*. Cambridge University Press.
- Manco, I.; Weck, B.; Doh, S.; Won, M.; Zhang, Y.; Bogdanov, D.; Wu, Y.; Chen, K.; Tovstogan, P.; Benetos, E.; et al. 2023. The Song Describer Dataset: a Corpus of Audio Captions for Music-and-Language Evaluation. *arXiv preprint arXiv:2311.10057*.
- Mittag, G.; Naderi, B.; Chehadi, A.; and Möller, S. 2021. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In *Proceedings of the 22nd Interspeech Conference (interspeech)*, 2127–2131.
- Pandey, P.; Swaminathan, R. V.; Girish, K.; Sen, A.; Xie, J.; Strimel, G. P.; and Schwarz, A. 2025. SIFT-50m: A large-scale multilingual dataset for speech instruction fine-tuning. *arXiv preprint arXiv:2504.09081*.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- Plack, C. J. 2023. *The sense of hearing*. Routledge.
- Reddy, C. K. e. a. 2021. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6493–6497. IEEE.
- Reimers, N.; and Gurevych, I. 2019. 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.
- Rubenstein, P. K.; Asawaroengchai, C.; Nguyen, D. D.; Bapna, A.; Borsos, Z.; Quitry, F. d. C.; Chen, P.; Badawy, D. E.; Han, W.; Kharitonov, E.; et al. 2023. AudioPaLM: A Large Language Model That Can Speak and Listen. *arXiv preprint arXiv:2306.12925*.
- Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2025. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 1–36.
- Shu, Y.; Dong, S.; Chen, G.; Huang, W.; Zhang, R.; Shi, D.; Xiang, Q.; and Shi, Y. 2023. LLASM: Large Language and Speech Model. *arXiv preprint arXiv:2308.15930*.
- Sun, L.; Xu, X.; Wu, M.; and Xie, W. 2024. Auto-ACD: A large-scale dataset for audio-language representation learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5025–5034.
- Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; MA, Z.; and Zhang, C. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *Proceedings of the 20th International Conference on Learning Representations (ICLR)*, 1–23.
- Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
- Wang, B.; Zou, X.; Lin, G.; Sun, S.; Liu, Z.; Zhang, W.; Liu, Z.; Aw, A.; and Chen, N. 2025. 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)*, 4297–4316.
- Wang, C.; Liao, M.; Huang, Z.; Lu, J.; Wu, J.; Liu, Y.; Zong, C.; and Zhang, J. 2023. BLSP: Bootstrapping Language-Speech Pre-Training via Behavior Alignment of Continuation Writing. *arXiv preprint arXiv:2309.00916*.
- Wu, M.; Dinkel, H.; and Yu, K. 2019. Audio caption: Listen and tell. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 830–834. IEEE.
- Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- Yuan, Y.; Jia, D.; Zhuang, X.; Chen, Y.; Chen, Z.; Wang, Y.; Wang, Y.; Liu, X.; Kang, X.; Plumbley, M. D.; et al. 2025. Sound-VECaps: Improving Audio Generation with Visually Enhanced Captions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Zhang, T.; Yu, Y.; Mao, X.; Lu, Y.; Li, Z.; and Wang, H. 2022. FENSE: A feature-based ensemble modeling approach to cross-project just-in-time defect prediction. *Empirical Software Engineering*, 27(7): 162.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.

Appendix

A. Vocabulary Size of Audio Captioning TestSet

This section introduces vocabulary size comparison to demonstrate the lexical diversity of MECAT-Caption. The following table indicates that the vocabulary size of MECAT-Caption contains about 4-17 times more words than the existing dataset.

Dataset	# Vocab
AudioCaps	5,581
AudioCaps-Enhanced	1,260
AutoACD	3,517
Clotho	1,852
StrongDescriber	2,726
LPMusicCaps-MTT	1,666
MECAT-Caption	22,595

B. Deployed Acoustic Models in Processing Pipeline

This section introduces the acoustic models deployed in our processing pipeline. These models are categorized into Content-Specific models (including Speech, Music, and Sound analysis) and Content-Unrelated models (Environment analysis), each designed to handle different aspects of audio understanding tasks.

Category	Subcategory	Model	Analysis Task
Content Specific	Speech	Speechbrain-ECAPA(Ravanelli et al. 2021)	Language Recognition
		Whisper Large v2 (Radford et al. 2023)	Auto Speech Recognition
		Pyannote-SD 3.1 (Bredin 2023)	Speaker Diarization
		Emotion2Vec (Ma et al. 2023)	Speaker Emotion Recognition
		Audeering-DSER (Wagner et al. 2023)	Dimensional Speaker Emotion Recognition
		Audeering-AGR (Burkhardt et al. 2023)	Age and Gender Recognition
		CommonAccent (Zuluaga-Gomez et al. 2023)	English Accent Recognition
	Music	Music Structure Analyzer (Kim and Nam 2023)	Tempo & Structure
		Music2Emo (Kang and Herremans 2025)	Emotion (Sentiment Polarity and Mood)
		MERT (Yizhi et al. 2023)	Musical Descriptor
		ByteSep (Kong et al. 2021)	Music Separation
Content Unrelated	Sound	Audio Flamingo 2 (Ghosh et al. 2025)	AudioLLM
		CED (Dinkel et al. 2024)	Sound Event Recognition
	Environment	DNSMOS (Reddy 2021, 2022)	Noisy Speech Quality Assessment
		NISQA V2.0 (Mittag et al. 2021)	Multidimensional Speech Quality Assessment
		SHAART (Hawley 2023)	Reverberation

References

- Bredin, H. 2023. pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proceedings of the 24th Interspeech Conference (interspeech)*, 1983–1987. ISCA.
- Burkhardt, F.; Wagner, J.; Wierstorf, H.; Eyben, F.; and Schuller, B. 2023. Speech-based age and gender prediction with transformers. In *Speech Communication; 15th ITG Conference*, 46–50. VDE.
- Dinkel, H.; Wang, Y.; Yan, Z.; Zhang, J.; and Wang, Y. 2024. CED: Consistent ensemble distillation for audio tagging. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 291–295. IEEE.
- Ghosh, S.; Kong, Z.; Kumar, S.; Sakshi, S.; Kim, J.; Ping, W.; Valle, R.; Manocha, D.; and Catanzaro, B. 2025. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 1–48.
- Hawley, S. H. 2023. SHAART: Speech and Hearing in Audio and Real Time. <https://github.com/drscotthawley/SHAART>.
- Kang, J.; and Herremans, D. 2025. Towards Unified Music Emotion Recognition across Dimensional and Categorical Models. arXiv:2502.03979.
- Kim, T.; and Nam, J. 2023. All-In-One Metrical And Functional Structure Analysis With Neighborhood Attentions on Demixed Audio. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.
- Kong, Q.; Cao, Y.; Liu, H.; Choi, K.; and Wang, Y. 2021. Decoupling Magnitude and Phase Estimation with Deep ResUNet for Music Source Separation. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 342–349. Citeseer.
- Ma, Z.; Zheng, Z.; Ye, J.; Li, J.; Gao, Z.; Zhang, S.; and Chen, X. 2023. Emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. *arXiv preprint arXiv:2312.15185*.
- Mittag, G.; Naderi, B.; Chehadi, A.; and Möller, S. 2021. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In *Proceedings of the 22nd Interspeech Conference (interspeech)*, 2127–2131.
- Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 28492–28518.
- Ravanelli, M.; Parcollet, T.; Plantinga, P.; Rouhe, A.; Cornell, S.; Lugosch, L.; Subakan, C.; Dawalatabad, N.; Heba, A.; Zhong, J.; Chou, J.-C.; Yeh, S.-L.; Fu, S.-W.; Liao, C.-F.; Rastorgueva, E.; Grondin, F.; Aris, W.; Na, H.; Gao, Y.; Mori, R. D.; and Bengio, Y. 2021. SpeechBrain: A General-Purpose Speech Toolkit. ArXiv:2106.04624, arXiv:2106.04624.
- Reddy, C. K. e. a. 2021. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6493–6497. IEEE.
- Reddy, C. K. e. a. 2022. DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 886–890. IEEE.
- Wagner, J.; Triantafyllopoulos, A.; Wierstorf, H.; Schmitt, M.; Burkhardt, F.; Eyben, F.; and Schuller, B. W. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10745–10759.

Yizhi, L.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; Yin, H.; Xiao, C.; Lin, C.; Ragni, A.; Benetos, E.; et al. 2023. MERT: Acoustic music understanding model with large-scale self-supervised training. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 1–24.

Zuluaga-Gomez, J.; Ahmed, S.; Visockas, D.; and Subakan, C. 2023. CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice. In *Proceedings of the 24th Interspeech Conference (interspeech)*, 5291–5295. ISCA.

C. LLM Audio Analysis Synthesis Prompts

Act as an expert audio analysis synthesizer to process multi-model JSON outputs through this workflow

Step 1: Multi-Domain Data Specifications

1.1 Multi-Domain Input Integration

- a) Speech: Speech recognition, speech emotion, speaker diarization and so on
- b) Music: Structure analysis, technical descriptors, emotion and so on
- c) Sound: Event detection timestamps, classifications
- d) Environment: Acoustic characteristics, interference markers
- e) Meta-info: Title and description of original video where audio clip was extracted

1.2 Data Integrity Challenges

- a) Missing fields
- b) Contradictory model outputs
- c) Confidence score variances

Step 2: Technical and Analytical Limitations

2.1 Model and System Constraints

- a) No speech recognition ability in audio captioning models (e.g., audio-flamingo variants)
- b) Accuracy disparities across the analyzed domains
- c) Potential conflicting information between models

2.2 Audio Content Heterogeneity

- a) Hybrid audio types (e.g., speech, music, sound-event, environment)
- b) Variable audio properties (e.g., clip lengths or quality)
- c) Reliable topic or domain, but absent or non-relevant details in Meta-info

Step 3: Audio Analysis Workflow

3.1 Salient Feature Identification

3.1.1 Identify dominant characteristics of this audio:

What makes this specific audio clip unique according to the analysis? Examples include:

- a) Specific spoken phrases
- b) Dominant musical styles or moods
- c) Significant sound events
- d) The overall acoustic scene
- e) Notable quality issues
- f) Complex interplay of elements

3.1.2 Supporting Evidence Extraction:

Gather the key details describing these salient features from the relevant JSON fields

3.2 Synthesis Rules

3.2.1 Generation Rules:

- a) Critically weigh evidence from different fields, considering inaccuracies or conflicts and accounting for domain-specific limitations
- b) Prioritize information most reliable or central to the audio's character based on overall data patterns
- c) Carefully identify conflicting information between fields and avoid mentioning conflicting aspects in the final caption. Focus only on consistent and unopposed information. Do not invent details not present in the data
- d) Crucial Constraint:
 - The final generated text must strictly describe only the analyzed content of the audio segment itself
 - It must not refer to the topic, title, description, or inferred subject matter from the overall video metadata
 - Avoid phrases like "in a clip from a video about..." or similar references to the source video's topic
 - Prohibit using parentheses to provide detailed explanation in any output, e.g., Moderate tempo (88 BPM)

3.2.2 Perspective Rules:

ALL answers must be created from the perspective of someone who ONLY LISTENED to the audio without any technical/model references or quantitative metrics (e.g., BPM, MOS, etc.)

3.2.3 Evaluation Rules:

Assign a confidence level (High or Low) based on the following aspects:

- a) Consistency: Are the different analyses in the JSON generally consistent or contradictory? High consistency increases confidence
- b) Completeness: Is key information present? (Fewer gaps = higher confidence)
- c) Clarity: How clearly does the consistent data point to the audio's nature? (Less ambiguity in reliable data = higher confidence)
- d) Metadata Context Usefulness: How relevant and useful was the overall video metadata in confirming or contextualizing findings from the clip's direct analysis?

3.3 Caption Development Framework

3.3.1 Systematic Caption

- a) Short (< 15 words):
 - Protocol: Primary domain characteristics + Most prevalent characteristic from cross-model correlation
 - Example: Blues guitar performance at live concert with audience reactions
- b) Long (1-2 sentences):
 - Protocol: Primary domain + significant secondary elements + notable quality factors
 - Example: A live concert recording featuring guitar with crowd cheers, despite occasional microphone static

3.3.2 Content-Focused Caption

a) Speech:

- Protocol: ASR content + paralinguistic context
- Example: Two speakers discussing jazz history, with piano accompaniment

b) Music:

- Protocol: Technical descriptors + performance context
- Example: Upbeat electronic track with distant traffic noise

c) Sound:

- Protocol: Event taxonomy + spatial relationships
- Example: Office environment with printer hum and keyboard typing, mild echo present

3.3.3 Content-Unrelated Caption

a) Environment:

- Protocol: Acoustic properties + interference profile
- Example: Studio recording with noticeable background interference

3.3.4 Caption Variants

- a) Lexical substitution (WordNet-based synonyms)
- b) Structural reordering (active/passive voice)
- c) Descriptive equivalence ('crowd cheers' → 'audience applause')

3.3.5 Null Handling

When no domain-specific elements are detected:

- a) Use explicit 'None' declaration in content field
- b) Generate null statement variants (e.g., 'No discernible speech content', 'Musical elements appear absent')

3.4 Question-Answering Design

3.4.1 Content Categories

Include questions across:

- a) Direct Perception (sound type, volume, duration)
- b) Sound Characteristics (timbre, rhythm, frequency characteristics)
- c) Environmental Perception (recording setting, echo, background noise)
- d) Quality Assessment (clarity, interference factors)
- e) Inference and Judgment (sound source, generation method, object properties)
- f) Application Context (use cases, semantic meaning)

3.4.2 Difficulty Levels

Include a mix of:

- a) Basic: Direct descriptive questions (e.g., 'What sound is heard?')
- b) Intermediate: Analytical questions (e.g., 'What are the characteristics of this sound?')
- c) Advanced: Inferential questions (e.g., 'In what environment was this recorded?')

- d) Complex: Comprehensive judgment questions (e.g., 'Based on the sound, what is the most likely material?')

3.4.3 Question Distribution

Basic (25%) — Intermediate (35%) — Advanced (25%) — Complex (15%)

3.4.4 Question Variety

Include:

- a) Ensure questions cover all listed categories
- b) Avoid repetitive question patterns or formats
- c) Include both yes/no questions and open-ended questions
- d) Include some questions about what is NOT present in the audio
- e) Include some comparative questions (e.g., 'Does this sound more like X or Y?')

3.4.5 Cognitive Levels

Include:

- a) Include questions requiring simple recognition
- b) Include questions requiring analysis of components
- c) Include questions requiring synthesis of information
- d) Include questions requiring evaluation or judgment

Step 4: Structured Output Specification (JSON Format)

Confidence: High/Low

Possible Conflicts: None or list of conflicting fields

Reasoning: 2-3 line evaluation considering model consensus and data quality

Short-Caption: Single-sentence essence

Short-Caption-Variants-1: Paraphrased version 1

Short-Caption-Variants-2: Paraphrased version 2

Main-Caption: Integrated summary

Main-Caption-Variants-1: Paraphrased version 1

Main-Caption-Variants-2: Paraphrased version 2

Speech-Captions: Speech-focused analysis or NONE

Speech-Caption-Variants-1: Paraphrased version 1

Speech-Caption-Variants-2: Paraphrased version 2

Music-Captions: Music-focused analysis or NONE

Music-Caption-Variants-1: Paraphrased version 1

Music-Caption-Variants-2: Paraphrased version 2

Sound-Captions: Sound-focused analysis or NONE

Sound-Caption-Variants-1: Paraphrased version 1

Sound-Caption-Variants-2: Paraphrased version 2

Environment-Caption: Environment-focused analysis

Environment-Caption-Variants-1: Paraphrased version 1

Environment-Caption-Variants-2: Paraphrased version 2

QA-Pair-1-id: 1 or None

QA-Pair-1-difficulty: basic, intermediate, advanced, or complex

QA-Pair-1-category: direct perception, sound characteristics, environmental perception, quality assessment, inference judgment, or application context

QA-Pair-1-question: question content
QA-Pair-1-answer: answer content
QA-Pair-2-id: 2 or None
QA-Pair-2-difficulty: basic, intermediate, advanced, or complex
QA-Pair-2-category: direct perception,sound characteristics, environmental perception,quality assessment, inference judgment,or application context
QA-Pair-2-question: question content
QA-Pair-2-answer: answer content
// ... 3 more QA pairs following the same pattern

D. LLM-as-Judge Prompts in evaluation

This section provides the prompt template required for LLM-as-Judge method. The evaluation *tasks* primarily include audio captioning and audio question-answering. In the template, the *description*, *subtask*, and *scoring_aspects* parameters can be referenced from the corresponding columns in the task table above, while *ref_texts* represents the samples to be evaluated.

Evaluation Prompt Template

You are tasked with evaluating if a set of candidate {tasks} responses accurately addresses the same audio as a reference set of answers. You will focus on the {description} for the subtask '{subtask}'.

Evaluation Steps:

- a) First, carefully compare the candidate answers with the reference answers
- b) Assess the accuracy and precision of how the audio characteristics are captured in the responses, then provide a 0-10 fine-grained score:
 - 10 = perfect match with the reference content
 - 0 = completely wrong
- c) Provide detailed scoring reasoning, explaining why you gave this score

Scoring Aspects: {scoring_aspects}

Score rubric (0-10 Scale):

- points 9-10: Excellent - Highly accurate, comprehensive, well-expressed
- points 8: Very Good - Accurate with minor gaps, clear expression
- points 7: Good - Mostly accurate, some missing details
- points 6: Acceptable - Basic accuracy, meets minimum HIGH standard
- points 4-5: Below Standard - Some correct elements but major issues
- points 2-3: Poor - Limited accuracy, significant problems
- points 0-1: Very Poor - Major errors or completely incorrect

You need to evaluate the following sample: {ref_texts}

Please return JSON-formatted evaluation results for the sample.

Return format (strict JSON array):

sample_id: sample ID

subtask: subtask_name

fine_score: <numerical value 0-10>

reasoning: detailed scoring rationale, including comparative analysis with reference answers

Table D.1: Category, subcategory, descriptions and scoring aspects of captioning evaluation with LLM-as-Judge method

Category	Subcategory	Description & Scoring Aspects
Systemic	Short	Description: quality of short audio descriptions Scoring Aspects: <ul style="list-style-type: none"> a) accuracy of core content capture (most important) b) conciseness and completeness of expression c) semantic consistency with reference descriptions
	Long	Description: quality of detailed audio descriptions Scoring Aspects: <ul style="list-style-type: none"> a) comprehensiveness and richness of description details b) accuracy of detailed descriptions c) logical structure and expression coherence
Content-Specific	Speech	Description: accuracy of speech content recognition Scoring Aspects: <ul style="list-style-type: none"> a) accuracy rate of speech content recognition b) accurate description of speaker characteristics (gender, accent, etc.) c) description of speech quality and environment
	Music	Description: quality of music content description Scoring Aspects: <ul style="list-style-type: none"> a) accuracy of music type, style, and rhythm identification b) identification of instruments and musical elements c) description of musical emotion and atmosphere
	Sound	Description: accuracy of sound event identification Scoring Aspects: <ul style="list-style-type: none"> a) accurate identification and classification of sound sources b) description of sound occurrence timing and duration c) description of sound intensity, pitch and other characteristics
Content-Unrelated	Environment	Description: accuracy of environment and recording quality description Scoring Aspects: <ul style="list-style-type: none"> a) identification of recording environment (indoor/outdoor, space size, etc.) b) assessment of audio technical quality (distortion, noise, etc.) c) description of environmental atmosphere and background characteristics

Table D.2: Category, subcategory, descriptions and scoring aspects of question-answering evaluation with LLM-as-Judge method

Category	Subcategory	Description & Scoring Aspects
Perception	Direct Perception	Description: accuracy of direct audio content identification Scoring Aspects: <ul style="list-style-type: none"> a) correct identification of primary audio elements (most important) b) accurate detection of presence/absence of specific sounds c) precise recognition of obvious audio features and events
	Sound Characteristics	Description: quality of sound property analysis Scoring Aspects: <ul style="list-style-type: none"> a) accurate description of sound attributes (pitch, volume, timbre, etc.) b) correct identification of sound sources and their properties c) precise characterization of audio dynamics and patterns
Analysis	Quality Assessment	Description: accuracy of audio quality evaluation Scoring Aspects: <ul style="list-style-type: none"> a) correct assessment of technical audio quality (clarity, distortion, etc.) b) accurate evaluation of recording conditions and fidelity c) appropriate judgment of audio production quality
	Environment Reasoning	Description: quality of environmental context inference Scoring Aspects: <ul style="list-style-type: none"> a) accurate inference of recording location and setting b) correct identification of spatial and acoustic properties c) logical deduction of environmental factors affecting audio
Reasoning	Inference Judgment	Description: accuracy of complex audio analysis and reasoning Scoring Aspects: <ul style="list-style-type: none"> a) correct interpretation of implicit audio information b) accurate temporal reasoning and sequence understanding c) logical inference of causality and relationships between audio elements
	Application Context	Description: relevance and appropriateness of contextual understanding Scoring Aspects: <ul style="list-style-type: none"> a) accurate understanding of audio’s intended purpose or context b) appropriate application of domain-specific knowledge c) correct interpretation of cultural, social, or professional context

E. Validation of LLM-as-Judge as a reference metric

To validate the effectiveness of LLM-as-Judge as a reference metric, we assessed its performance on three distinct sets of responses with varying quality levels: Right (detailed and accurate rephrasings of the ground-truth reference), Safe (generic, vague descriptions, e.g., "A man is speaking" for all speech-only audio), and Wrong (factually incorrect references randomly selected from other samples). As shown in following table, our analysis confirms that LLM-as-Judge method serves as a reliable evaluator. It successfully distinguishes between the quality tiers, with mean scores consistently following the expected Right > Safe > Wrong order for both captioning and QA tasks. Furthermore, its inter-rater reliability, measured by Fleiss' Kappa (κ), is substantial for QA ($\kappa = 0.73$) and moderate for captioning ($\kappa = 0.43$). However, the significant practical limitations of LLM-as-Judge method—including high computational cost, slow speed, and sensitivity to prompt engineering—motivate our development of the DATE metric as an efficient and scalable alternative.

Type	Mean		Fleiss' Kappa (κ)	
	Caption	QA	Caption	QA
Right	0.78	0.97	0.68	0.74
Safe	0.24	-	0.17	-
Wrong	0.13	0.12	0.45	0.72
Overall	-	-	0.43	0.73

F. Task-specific prompts for LALM in MECAT tasks

This section introduces the prompts used by LALM in the caption task during MECAT evaluation. For Audio-Flamingo2 and Kimi-Audio models, their prompts can be found in Table C.1. For Qwen-Omni models, their prompts are shown in Table C.2, where the system prompt has been modified to "You are a helpful assistant". For the QA task, the prompt for each sample is simply the corresponding question.

Table F.1: Prompts for Audio-Flamingo2 and Kimi-Audio models in caption task

Category	Subcategory	Prompt
Systematic	Short	Provide a caption for this audio within 15 words
	Long	Provide a caption for this audio within 1-2 sentences
Content-Specific	Speech	Provide a caption for the speech content in this audio
	Music	Provide a caption for the music content in this audio
	Sound	Provide a caption for general sound excluding speech and music
Content-Unrelated	Environment	Provide a caption for quality or acoustic environment for this audio

Table F.2: Prompts for Qwen-Omni models in caption task

Category	Subcategory	Prompt
Systematic	Short	Listen to the audio and provide a caption for this audio within 15 words
	Long	Listen to this audio and provide a caption for this audio within 1-2 sentences
Content-Specific	Speech	Listen to the audio and provide a caption describing the speech content in this audio
	Music	Listen to the audio and provide a caption for the music content in this audio
	Sound	Listen to the audio and provide a general sound excluding speech and music
Content-Unrelated	Environment	Listen to this audio and provide a caption for quality or acoustic environment for this audio