Beyond Single-Audio: Advancing Multi-Audio Processing in Audio Large Language Models

Yiming Chen[†]   **Xianghu Yue**[†*]   **Xiaoxue Gao**[¶]   **Chen Zhang**[†]
Luis Fernando D'Haro[*]   **Robby T. Tan**[‡,†]   **Haizhou Li**[♮,†]

[†]National University of Singapore   [♮]The Chinese University of Hong Kong, Shenzhen
[*]Universidad Politécnica de Madrid   [‡]ASUS Intelligent Cloud Services
[¶]I2R, Agency for Science, Technology, and Research (A*STAR)
{yiming.chen,xianghu.yue,chen_zhang}@u.nus.edu
Gao_Xiaoxue@i2r.a-star.edu.sg luisfernando.dharo@upm.es
robby_tan@asus.com haizhouli@cuhk.edu.cn

Abstract

Various audio-LLMs (ALLMs) have been explored recently for tackling different audio tasks simultaneously using a single, unified model. While existing evaluations of ALLMs primarily focus on single-audio tasks, real-world applications often involve processing multiple audio streams simultaneously. To bridge this gap, we propose the first multi-audio evaluation (MAE) benchmark that consists of 20 datasets from 11 multi-audio tasks encompassing both speech and sound scenarios. Comprehensive experiments on MAE demonstrate that the existing ALLMs, while being powerful in comprehending primary audio elements in individual audio inputs, struggling to handle multi-audio scenarios. To this end, we propose a novel multi-audio-LLM (MALLM) to capture audio context among multiple similar audios using discriminative learning on our proposed synthetic data. The results demonstrate that the proposed MALLM outperforms all baselines and achieves high data efficiency using synthetic data without requiring human annotations. The proposed MALLM opens the door for ALLMs towards multi-audio processing era and brings us closer to replicating human auditory capabilities in machines. [1]

1 Introduction

Large language models (LLMs) have become remarkably powerful, driving advancements in various tasks across the field of natural language processing (NLP) (Touvron et al., 2023; Achiam et al., 2023; Team et al., 2023). Recent advancements in LLMs have also led to the development of various powerful audio large language models (ALLMs) (Chu et al., 2023; Huang et al., 2024; Rubenstein et al., 2023), which have achieved impressive results on a range of audio tasks, e.g., automatic speech recognition (Hu et al., 2024), speech
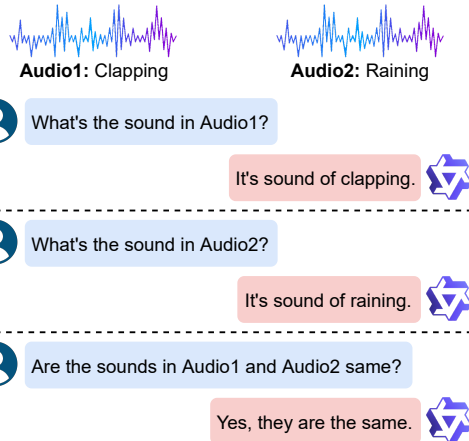

Figure 1: Motivating Example - Qwen-Audio's responses for single audio inputs and a two-audio input task.

synthesis (Gao et al., 2024c), sound event classification (Tang et al., 2024; Yue et al., 2024b).

However, a crucial limitation exists: current ALLM training and evaluation primarily focus on single audio inputs. This is a significant drawback, as various real-world applications, e.g., virtual assistants, often require processing multiple audio streams simultaneously. Additionally, multi-audio processing is essential for effectively implementing few-shot in-context learning, which is a fundamental capability for advanced LLMs. Unlike text-based LLMs, which excel at handling multiple texts (Wang et al., 2024; McKenna et al., 2023), and vision LLMs, which have established benchmarks for processing multiple images (Li et al., 2024; Huang et al., 2023b; Zhao et al., 2024; Li et al., 2023), the audio field lacks systematic evaluations and benchmarks for multi-audio tasks with ALLMs. Although several ALLMs (Chu et al., 2023; Zhan et al., 2024) claim to handle multi-audio contexts, their performance quantification remains unclear. This underscores a significant gap in the audio field, in contrast to the vision and language fields, which

---

already possess established evaluations and benchmarks for their respective tasks.

To bridge this gap, we propose the first dedicated multi-audio benchmark for ALLMs, encompassing 11 tasks for both open-ended and close-formed generation across speech and sound domains. Comprehensive experiments on 15 ALLMs reveal that existing open-source ALLMs fall short in multi-audio scenarios despite their strong single-audio processing abilities. For instance, as shown in Fig. 1, while ALLMs can perfectly identify Audio1 as clapping and Audio2 as raining respectively in a single-audio scenario, they struggle to determine whether Audio1 and Audio2 are the same when tasked with understanding the relationship between two audio inputs. This suggests that current ALLMs are not well-equipped to handle multi-audio tasks, even those that are straightforward for humans. This further underlines the need for advancements in ALLMs to handle multiple audio inputs to enhance human-computer interaction.

To this end, we propose an innovative and scalable multi-audio large language model (MALLM) that effectively captures audio contexts essential for reasoning over multiple audios. Inspired by the success of discriminative learning (Clark et al., 2020; Li et al., 2024; Jia et al., 2021), we define a challenging discriminative task that trains the model to discover the subtle differences between two similar audio samples. Furthermore, we introduce a scalable audio pairs synthesis strategy to enable multi-audio processing ability without the need for data collection and human labeling. Comprehensive experiments show that MALLM significantly outperforms existing open-source ALLMs under multi-audio scenarios while maintaining competitiveness under single-audio scenarios.

Overall, our contributions include: **(1) Novel Evaluation Benchmark:** We propose the first multi-audio benchmark (MAE) for evaluating the multi-audio processing capabilities of ALLMs, encompassing a diverse range of tasks from both the speech and sound domains. **(2) Advanced Multi-Audio LLM:** Beyond focusing on the single audio content, our proposed MALLM not only demonstrates remarkable performance across diverse multi-audio tasks but also achieves remarkable data efficiency through an innovative data synthesis strategy. **(3) Comprehensive Evaluation:** We conduct a comprehensive evaluation on 15 ALLMs across various tasks, providing a solid foundation for future research.

2 Related Works

Recently, numerous studies have integrated audio encoders with pre-trained LLMs to develop multimodal ALLMs, serving as general-purpose task solvers for various audio-input tasks across speech and sound domains. Most ALLMs focus on specific audio types, like speech (Zhang et al., 2023; Zhan et al., 2024; Das et al., 2024), or sound (Panagopoulou et al., 2023; Kong et al., 2024; Moon et al., 2023; Han et al., 2023). A few recent models handle multiple types of audio, showing strong capabilities in universal audio understanding (Gong et al., 2023; Chu et al., 2023; Tang et al., 2024). Meanwhile, to effectively benchmark the advancements in ALLMs, new evaluation benchmarks such as AIR-Bench (Yang et al., 2024) and Dynamic-SUPERB (Huang et al., 2023a) have been introduced. However, these benchmarks and existing ALLMs primarily focus on tasks involving single audio inputs, largely overlooking scenarios with multiple audio streams. In contrast, our work introduces the MAE benchmark, specifically designed to assess multi-audio processing capabilities. Notably, MAE includes a variety of tasks across diverse scenarios, encompassing both audio and speech, making it suitable for benchmarking a broad spectrum of ALLMs. Additionally, we develop the MALLM, the first ALLM specifically tailored for multi-audio tasks, demonstrating significant improvements in processing multiple audio streams while maintaining competitive performance on single-audio tasks.

3 MAE Benchmark

We propose the first multi-audio benchmark (MAE) for evaluating ALLMs, as illustrated in Fig. 2. The MAE comprehensively benchmarks the multi-audio processing capabilities of ALLMs by including a wide variety of generation tasks from different fields and scenarios. It comprises six speech tasks and five sound tasks, covering both open-ended and close-form generation tasks. Open-ended questions allow models to produce free-form responses without predefined constraints. Conversely, closed-form questions restrict models to a predetermined set of possible outcomes. The MAE is automatically constructed using an advanced text-only LLM from various existing single-audio datasets without the need for further human annotation. Each sample in the MAE includes a combination of two audio contexts and a task instruction.
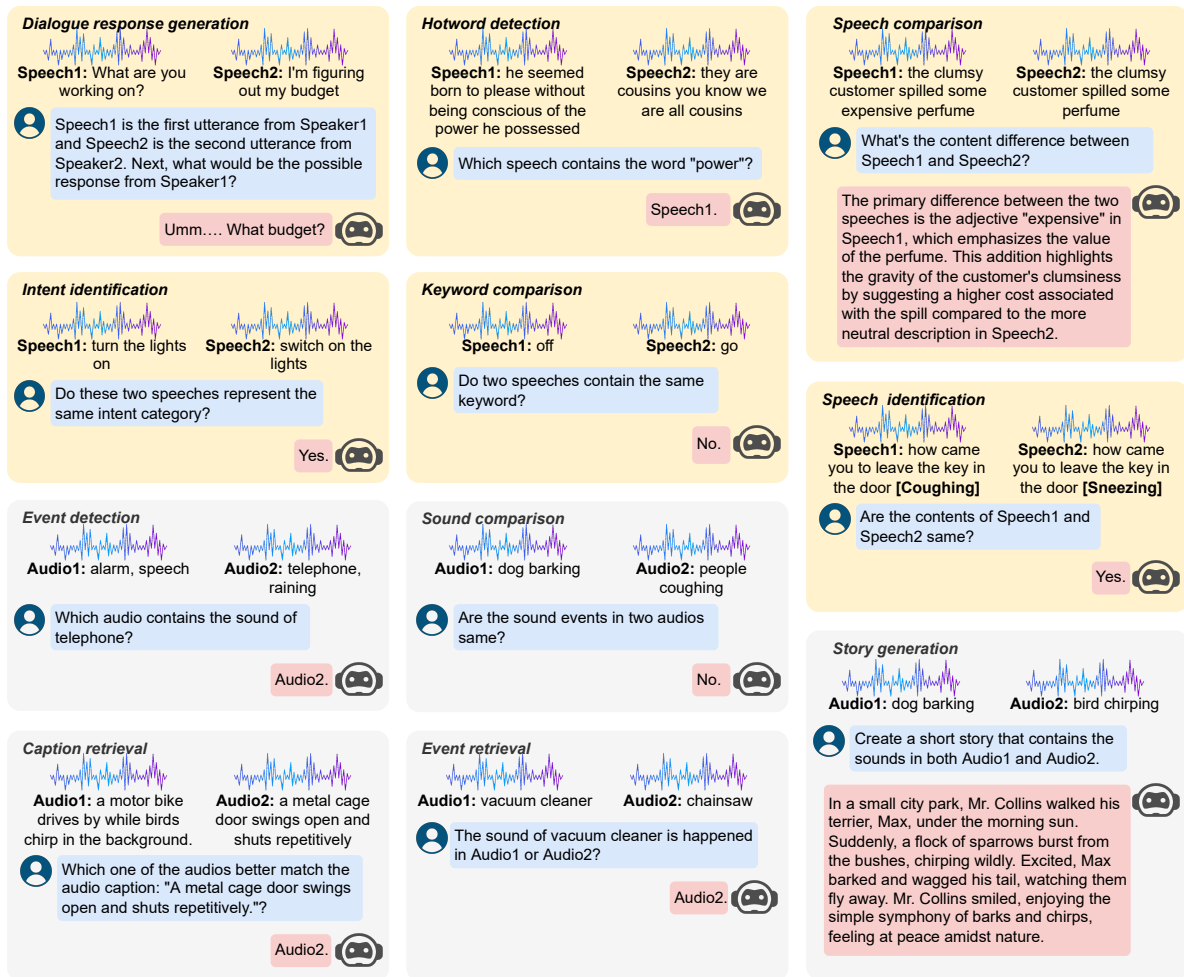
Figure 2: Overview of proposed MAE benchmark. Yellow blocks show the speech tasks, while grey blocks show the sound tasks. Speech transcriptions, audio labels, input instructions, and expected output responses are given.

To correctly answer the questions from MAE, ALLMs are required to effectively combine the information from both context audios. Notably, MAE is designed to examine ALLMs' multi-audio processing ability at different levels. For closed-form generation tasks, we directly derive the ground truth labels from existing single-audio labels to compute metric scores. Motivated by the recent success of LLM-based text evaluation (Fu et al., 2023), we utilize a text-only LLM as an automatic evaluator to assign binary good/bad labels to responses generated by ALLMs for open-ended generation tasks (see APPX. A). We summarize the data statistics of the MAE in Tab. 1. Next, we introduce the construction process for each task.

3.1 Speech Tasks

For speech, we design two open-ended generation tasks from sentence and dialogue levels and four closed-form tasks from word and sentence levels.

Speech comparison: This sentence-level, open-ended task requires ALLMs to identify content differences between pairs of speeches. Speech pairs are sourced from ASR datasets with timestamp-level transcription. We segment the speech into spoken words with timestamps and instruct an LLM to reconstruct a subset of these words into a new speech without semantic errors, such as omitting the adjective "expensive" in Fig. 2. Then, we combine the original and reconstructed speech or two reconstructed speeches as an evaluation pair. Note that ground truth transcriptions of two speeches can be obtained from the original labels, allowing us to utilize the LLM as an evaluator to label the ALLMs' responses based on the derived transcriptions.

Dialogue response generation: This dialogue-level, open-ended task involves generating a subsequent utterance from the examined ALLMs in a dialogue based on the first two utterances. Similarly, the LLM evaluator assesses the overall quality of the generated dialogue response with respect to

Type	Task	Category	Source	# Samples	Metric
Speech	Speech comparison	Open-ended	Librispeech (Panayotov et al., 2015) TIMIT (Lamel et al., 1989)	1000 1000	Acc
	Dialogue generation	Open-ended	DailyTalk (Lee et al., 2023) MELD (Poria et al., 2019)	1000 304	Acc
	Hotword detection	Closed-form	Librispeech (Panayotov et al., 2015) TIMIT (Lamel et al., 1989)	500 / 500 500 / 500	Acc
	Speech identification	Closed-form	Librispeech (Panayotov et al., 2015) WSJ (Paul and Baker, 1992)	500 / 500 500 / 500	F1/Acc
	Keyword comparison	Closed-form	AudioKeyword (Mazumder et al., 2021) SpeechCommands (Warden, 2018)	513 / 513 500 / 500	F1/Acc
	Intent identification	Closed-form	FLUENT (Lugosch et al., 2019)	504 / 507	F1/Acc
Sound	Story generation	Open-ended	ESC-50 (Piczak, 2015) UrbanSound (Salamon et al., 2014)	1000 1000	Acc
	Sound comparison	Closed-form	ESC-50 (Piczak, 2015) UrbanSound (Salamon et al., 2014)	500 / 500 500 / 500	F1/Acc
	Caption retrieval	Closed-form	Clotho (Drossos et al., 2020) AudioCaps (Kim et al., 2019)	506 / 494 528 / 472	Acc
	Event retrieval	Closed-form	ESC-50 (Piczak, 2015) UrbanSound (Salamon et al., 2014)	503 / 497 503 / 497	Acc
	Event detection	Closed-form	AudioSet (Gemmeke et al., 2017)	500 / 500	Acc

Table 1: Data statistics of MAE benchmark. For closed-form tasks, class-wise number of samples are given.

the transcriptions of the preceding utterances.

Hotword detection: Hotword detection is defined as a word-level, close-ended task derived from existing ASR datasets. Given a pair of randomly sampled speeches, the ALLMs are asked to detect which speech contains a specific noun that appears only in one of them. This task evaluates the models' precision in recognizing and differentiating specific lexical items within speech contexts.

Intent identification: This task is defined as a sentence-level, close-ended task. We construct context speech pairs by randomly sampling speeches that either share the same intent or represent different intents from existing intent classification datasets. The ALLMs are tasked with determining whether the paired speeches belong to the same intent category, evaluating their ability to discern and categorize the underlying communicative purposes in speech data.

Keyword comparison: This task is a word-level, close-ended challenge derived from existing keyword-spotting datasets. We sample pairs of speeches that either share the same keywords or contain different ones. The ALLMs are asked to determine whether the keywords presented in the two speeches are identical.

Speech identification: Speech identification is a sentence-level, close-formed task involving context-positive speech pairs and negative pairs with distinct transcriptions. The ALLMs are required to discern whether speech pairs have the same content. Positive pairs are formed by adding different background noises (e.g., coughing and sneezing in Fig. 2) to the same speech, yielding identical content but different backgrounds. Negative pairs consist of randomly selected speeches.

3.2 Sound Tasks

For sound, we design one open-ended and four closed-form tasks to challenge the ALLM comprehension of global and local level audio intricacies.

Story generation: This open-ended task tests the ALLMs' global understanding of audio by challenging them to integrate diverse sound events from multiple audios into a coherent narrative. We randomly select two audio samples, each with a distinct event, and task the ALLMs to create a brief narrative incorporating both sounds. The correctness of the story is contingent upon the inclusion of both specified sounds; any omission marks it as incorrect. For instance, as shown in Fig. 2, a story is correct if it includes both the dog barking and
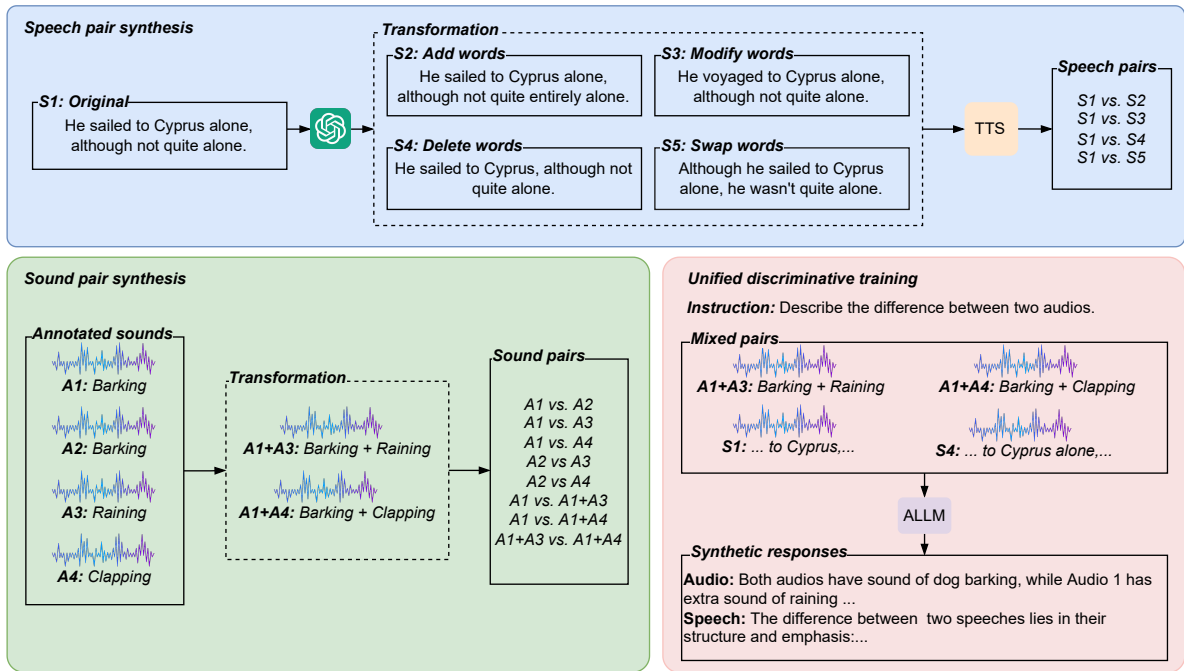
Figure 3: MALLM training framework: speech/sound pair synthesis, and unified synthetic discriminative training.

bird chirping sounds. We utilize the LLM evaluator to assess the completeness of the sound integration.

Event detection: Using multi-label sound event classification datasets, we create a task where ALLMs are asked to detect a specified sound in given audios containing multiple sounds, such as identifying if telephone ringing is present in Audio1 or Audio2 (Fig. 2). This task demands precise understanding and interpretation of local audio details to identify specific sound events accurately.

Sound comparison: We randomly select two audio samples, which may either share the same sound event labels or differ. ALLMs are then instructed to compare two audios to determine if these two audios belong to the same sound event category. This requires the models to not only recognize the sound event in a single audio but also compare them for similarities and differences.

Caption retrieval: In this task, the ALLMs are presented with two audio samples and a corresponding audio caption. The models are required to determine which of the two audios aligns more closely with the given caption. This task assesses the ability of the ALLMs to understand audio content globally and analyze their similarities with descriptive text.

Event retrieval: Similar to caption retrieval but simpler, ALLMs are asked to retrieve the audio sample based on one single specific sound event label each time, testing their ability to match labels with corresponding audio samples.

4 MALLM

In this section, we introduce the specifics of our proposed MALLM. Fig. 3 depicts our discriminative ALLM fine-tuning strategy. Specifically, MALLM is trained to describe the subtle distinctions between two similar audios, enhancing its capability to handle multi-audio scenarios. To effectively tackle the challenging discriminative problem, the MALLM is expected to possess a deep comprehension of all input audios and infer both intra-audio and inter-audio relationships. In this way, MALLM can be effectively augmented with reasoning abilities across multiple audios. Furthermore, to improve the data efficiency and scalability, we propose the automatic construction of a synthetic training dataset without additional human intervention. This dataset includes pairs of both synthetic speech and sound pairs with subtle differences, which are used to fine-tune MALLM, expanding its multi-audio processing capabilities.

Speech pair synthesis: To automate the construction of speech pairs, we start by randomly selecting a sentence from a text corpus. An advanced text-based LLM is then utilized to generate four variations of the sentence, each containing slight modifications. This is achieved by instructing the LLM with different prompts that direct it to add, delete, or modify a limited number of words in the original sentence and to alter its structure. Detailed

prompts for this process are provided in APPX. C. Following these text modifications, a robust pre-trained text-to-speech (TTS) model converts both the original and modified sentences into speech. This controlled synthesis process ensures that accurate ground-truth transcriptions are available for all generated speech samples. Ultimately, each original sentence is paired with its four modified counterparts, yielding four synthetic speech pairs with subtle differences per input sentence.

Audio pair synthesis: To generate audio pairs with synthetic data, we start from individual audio recordings labeled with distinct sound events. For two individual audios containing different sound events, we directly mix the audios with balanced signal-to-noise ratios (SNR), resulting in a new audio file containing both sound events. For instance, mixing an audio of a dog barking with an audio of rain produces a new synthetic audio featuring both barking and raining. This mixing process ensures the ground truth sound event labels for the synthetic data are accurately maintained. Ultimately, various combinations of these mixed audios are generated to form audio pairs, as illustrated in Fig. 3.

Discriminative learning: Following the acquisition of synthetic data pairs, we fine-tune the MALLM through inter-audio discriminative training tasks encompassing a mixture of speech and audio pairs with the instruction tuning loss. As previously stated, we have access to ground truth labels for all synthetic data, facilitating the straightforward transformation of these labels into natural responses for training purposes. For speech pairs, we utilize an advanced text-only LLM to generate differences between two speech transcriptions as the ground-truth responses. Regarding audio pairs, we apply predefined rules to map annotated sound event labels to their corresponding responses (see APPX. B). To stabilize the training and prevent catastrophic forgetting, we also sample single audios from various tasks for joint training. Details regarding the MALLM training dataset are provided in APPX. B.

5 Experiment

5.1 Experiment Setup

Throughout the experiments, we use GPT-4[2] for data synthesis and downstream performance evaluation. For the TTS module in speech pair synthesis,

we utilize MMS-TTS (Pratap et al., 2024). We implement the MALLM based on Qwen-Audio (Chu et al., 2023), incorporating an audio encoder, Whisper-large-v2 (Radford et al., 2023), and a large language model, Qwen-7B (Bai et al., 2023). We use LoRA (Hu et al., 2022) to fine-tune all modules of Qwen-Audio (Chu et al., 2023) except the audio encoder with a learning rate 5e-5 and batch size 16 for 5 epochs. The training takes 48 GPU hours on Nvidia A100.

5.2 Examined Models

Our evaluation encompasses a diverse range of ALLMs on the MAE benchmark. For speech-related tasks, we assess LTU-AS (Gong et al., 2023), SALMONN (Tang et al., 2024), Qwen-Audio (Chu et al., 2023), SpeechGPT (Zhang et al., 2023) and AnyGPT (Zhan et al., 2024). In the domain of sound processing, we evaluate LTU (Gong et al., 2024), LTU-AS (Gong et al., 2023), SALMONN (Tang et al., 2024), Qwen-Audio (Chu et al., 2023), NextGPT (Wu et al., 2023), PandaGPT (Su et al., 2023), X-InstructBLIP (Panagopoulou et al., 2023) and Pengi (Deshmukh et al., 2023). In addition, we include advanced proprietary models, Gemini-Flash[3] and Gemini-Pro[4], as the performance upper bound of speech tasks, since they cannot handle the sound tasks. Note that the examined ALLMs sometimes fail to follow the instructions, leading to the absence of predicted answers for closed-form tasks. In that case, we default the predicted answer to "No" or "Audio1", depending on the task context. Since MAE is class-balanced, this strategy effectively equates to a random guessing strategy.

5.3 MAE Results

MAE-Speech: The evaluation results for various ALLMs on the MAE-Speech benchmark (Tab. 2) show that all open-source ALLMs struggle with multi-audio scenarios across all tasks. The highest-performing model, Qwen-Audio, merely achieves an average accuracy of 39.6%. In closed-form tasks, these models tend to collapse to consistently give the same answer for all queries, leading to approximately 50% accuracy. For example, Qwen-Audio always answers "Yes" for intent identification. Furthermore, models like LTU-AS frequently fail to adhere to the provided instructions, resulting in extremely poor performance, particularly in

Model	Speech comparison (Acc)		Dialogue response generation (Acc)		Speech identification (F1/Acc)	
	Librispeech	TIMIT	DailyTalk	MELD	Librispeech	WSJ
LTU-AS	0.6	0.9	14.7	11.5	66.7 / 50.0	66.7 / 50.0
SALMONN-7B	17.7	17.5	1.6	3.6	66.7 / 50.0	66.7 / 50.0
SALMONN-13B	22.9	27.9	2.2	3.6	68.5 / 54.1	67.3 / 51.5
SpeechGPT	1.2	0.6	1.9	3.3	2.0 / 50.5	1.6 / 50.2
AnyGPT	3.7	2.9	2.1	3.0	27.6 / 44.5	32.5 / 48.8
Qwen-Audio	11.6	13.1	20.4	15.5	1.58 / 50.2	0.00 / 50.0
Gemini-Flash	66.4	55.8	89.7	56.6	99.4 / 99.4	99.6 / 99.6
Gemini-Pro	60.5	52.5	79.0	44.4	99.3 / 99.3	99.2 / 99.2
MALLM	**70.3**	**59.6**	**50.7**	**40.5**	**96.5 / 96.6**	**87.6 / 89.0**

Model	Hotword detection (Acc)		Keyword comparison (F1/Acc)		Intent identification (F1/Acc)	Avg. (Acc)
	Librispeech	TIMIT	SpeechCommand	AudioKeyword	FLUENT	
LTU-AS	49.8	49.9	52.6 / 49.2	64.0 / 48.3	66.6 / 50.3	34.1
SALMONN-7B	61.4	55.7	64.5 / 50.3	65.2 / 48.8	66.8 / 50.2	37
SALMONN-13B	51.9	51.2	0.4 / 50.0	6.0 / 50.9	1.2 / 50.2	37.9
SpeechGPT	50.0	50.1	10.1 / 48.0	11.2 / 48.7	20.8 / 50.3	32.3
AnyGPT	49.0	49.8	2.7 / 49.1	3.7 / 49.4	14.8 / 49.9	32.0
Qwen-Audio	60.9	63.9	0.0 / 50.0	0.0 / 50.0	66.8 / 50.2	39.6
Gemini-Flash	99.9	100.0	91.7 / 92.3	79.4 / 82.8	80.2 / 75.5	83.5
Gemini-Pro	99.7	99.8	94.9 / 95.0	89.36 / 90.16	83.0 / 79.6	81.7
MALLM	**80.0**	**86.8**	**88.2 / 88.7**	**78.03 / 80.90**	**75.7 / 68.6**	**73.8**

Table 2: Comparison of various ALLMs on MAE-Speech. The first block shows the open-source ALLMs, while the second block shows the proprietary ALLMs. The best performance among open-source ALLMs are bold.

open-ended tasks. This issue is hypothesized to stem from insufficient supervised fine-tuning data.

In contrast, the proprietary Gemini series models demonstrate exceptional performance across the MAE benchmark, reaching nearly 100% accuracy in simple tasks like hotword detection. This underscores the significant performance disparity between current open-source ALLMs and their proprietary counterparts. Notably, the Gemini technical reports do suggest the multi-audio processing abilities of Gemini models and examine a multi-audio in-context learning scenario (Reid et al., 2024). Therefore, it's likely that Gemini models are specially optimized with multi-audio data.

Our newly developed MALLM significantly surpasses all open-source models, posting an average accuracy of 73.8%. It is important to mention that the MALLM's tuning data does not encompass dialogue or intent-related data. Nevertheless, we observe substantial improvements in these areas, suggesting that our discriminative training strategy effectively enhances the general multi-audio processing capabilities of ALLMs. Overall, our MALLM substantially narrows the gap between open-source and proprietary models, even surpass-

ing proprietary models in tasks like speech comparison.

MAE-Sound: The evaluation results for various ALLMs on the MAE-Sound benchmark, shown in Tab. 3, reveal that most ALLMs struggle with multi-audio contexts similar to the findings from MAE-Speech. Among all ALLMs, Qwen-Audio achieves the best performance, particularly in event retrieval and sound detection tasks. Despite this, its overall performance is still deemed unsatisfactory. Our proposed MALLM significantly outperforms Qwen-Audio and all other models, achieving an average accuracy of 74.3% across all tasks. This superior performance again underscores the efficacy of our discriminative learning approach tailored for multi-audio contexts in large audio language modeling. Overall, MALLM demonstrates a marked improvement over existing models in multi-audio processing across both speech and sound domains, which suggests promising directions for further research and development.

5.4 Discussion

Single-audio performance: Alongside multi-audio processing, we also evaluate the single-audio

Model	Sound comparison (F1/Acc)		Caption retrieval (Acc)		Event retrieval (Acc)		Sound detection (Acc)	Story generation (Acc)		Avg. (Acc)
	ESC-50	UrbanSound	Clotho	AudioCaps	ESC-50	UrbanSound	AudioSet	ESC-50	UrbanSound	
LTU	19.2 / 51.3	42.0 / 50.0	51.6	51.4	49.7	50.3	50.0	25.7	27.5	45.3
LTU-AS	43.3 / 50.5	6.4 / 50.5	50.5	52.8	49.4	50.0	50.0	23.3	22.6	44.4
Pengi	3.1 / 50.0	2.0 / 49.8	50.3	52.7	50.3	50.3	50.0	0.9	0.3	39.4
SALMONN-7B	66.7 / 50.0	66.7 / 50.0	53.3	55.6	50.8	50.8	50.7	6.2	13.4	42.3
SALMONN-13B	3.9 / 50.5	10.3 / 51.0	51.3	50.4	54.8	53.1	53.2	25.3	47.5	48.6
NextGPT-7B	4.2 / 50.3	2.4 / 50.1	50.4	52.2	50.4	50.3	50.0	1.0	0.8	39.5
PandaGPT-7B	66.7 / 50.0	66.7 / 50.0	51.1	54.9	51.3	52.0	52.1	4.3	3.5	41.0
PandaGPT-13B	45.5 / 48.7	44.6 / 52.7	50.8	49.1	51.1	51.0	52.8	2.6	1.7	40.1
XInstructBLIP-7B	47.3 / 51.0	44.6 / 51.7	50.8	52.8	49.7	50.3	50.0	3.4	3.9	40.4
XInstructBLIP-13B	2.7 / 50.1	9.7 / 49.5	50.8	53.2	50.0	49.5	49.7	7.0	4.5	40.5
Qwen-Audio	0.4 / 50.1	17.7 / 50.0	58.8	55.2	83.1	74.3	68.5	30.6	42.0	57.0
MALLM	**80.7 / 76.6**	**69.7 / 65.7**	**81.3**	**65.0**	**95.5**	**80.4**	**72.0**	**59.2**	**70.8**	**74.1**

Table 3: The performance of various ALLMs on MAE-Sound. The best performance among ALLMs are bold.

processing ability of the MALLM compared to the backbone model, Qwen-Audio, with results for ASR and event classification in Tab. 4. Our MALLM not only demonstrates superior multi-audio understanding but also maintains competitive, sometimes superior, single-audio performance. These findings underscore the MALLM's broad applicability and validate the effectiveness of our synthetic training schema, which equips the model to handle a wide range of scenarios.

Model	Librispeech ASR (WER)		Event Classification (Acc)	
	dev-clean	test-clean	ESC-50	UrbanSound
Qwen-Audio	**2.0**	2.3	70.0	**74.9**
MALLM	**2.0**	**2.2**	**77.2**	73.5

Table 4: Comparison of single-audio performance between Qwen-Audio and our MALLM.

Case study: In addition to quantitative results, we show a case study in Tab. 5 to illustrate the quality of generated responses from different ALLMs. We leave more case studies in APPX. D. In the example, the models need to generate a story with sounds of both keyboard typing and water dropping. While both generate coherent narratives, Qwen-Audio focuses only on keyboard typing, neglecting the water sound. On the contrary, MALLM incorporates both sounds and demonstrates a more comprehensive understanding of the task requirements, demonstrating its superior capabilities in handling complex multi-audio inputs compared to other models.

Human validation of automatic metrics: For open-ended tasks in MAE, we employ an LLM evaluator to automatically assess the correctness of responses from ALLMs. To validate the reliability of this evaluator, we manually annotate responses

Model	Response
Qwen-Audio	In a dimly lit room, a person sat at a desk typing furiously on a **keyboard**. The sound of the keys clacking against the keys filled the room. Suddenly, the person stopped typing and listened to the sound of their own typing. It was a soothing, calming sound that brought them a sense of peace. They closed their eyes and took a deep breath, letting the sound wash over them. They felt a sense of accomplishment and satisfaction, knowing that they had just typed out a masterpiece.
MALLM	In a quiet, dimly lit room, there was a constant sound of typing on a **keyboard**. As the sound filled the air, drops of **water** began to fall from a faucet. The rhythmic clicking of the **keyboard** mixed with the rhythmic splashing of the **water**, creating a unique and soothing atmosphere. The combination of sounds brought a sense of calm and focus to the environment. The **keyboard** clicking was steady and consistent, much like the dripping **water**, and both sounds seemed to blend together, creating a harmonious and tranquil scene.

Table 5: Case study: ALLMs are instructed to generate a story containing the sound of water and keyboard. The occurrence of 'keyboard' and 'water' in the model response is in bold.

from Qwen-Audio and MALLM, which exhibit the best performance among open-source models. We randomly selected 50 responses from each model across six datasets within three tasks: speech comparison, dialogue response generation, and story generation. For each sample, we ask three annotators to annotate the binary label of the ALLM response.

We observe a high inter-annotator agreement of 0.81 for the story generation task, and a moderate inter-annotator agreement of 0.39 and 0.38 for

dialogue response generation and speech comparison respectively, which validates the reliability of our human study. The accuracy computed from LLM annotated labels and human annotated ones are summarized in Tab. 6. Across all tasks, particularly in story generation, the LLM evaluator shows a high correlation with human judgments, affirming its reliability as a tool in our MAE benchmark. Note that the evaluation of speech comparison and dialogue response are subjective and challenging. Even different human annotators exhibit different perceptions, as highlighted in the inter-annotator agreement. Therefore, despite the slightly lower accuracy, the LLM evaluator still serves as a robust tool for providing reliable assessments.

Speech comparison	Dialogue generation	Story generation
71.5	78.5	88.5

Table 6: The accuracy of LLM evaluator against human annotators.

6 Conclusion

In this paper, we introduce the first multi-audio evaluation benchmark, MAE, to examine the multi-audio processing ability of ALLMs. Extensive evaluations across 15 ALLMs reveal generally unsatisfactory performance among current open-source models in handling multi-audio scenarios. To this end, we develop a simple and effective discriminative training framework that leverages synthetic data. The resulting model, MALLM, significantly surpasses all existing open-source ALLMs without incurring additional costs for human annotation. This work lays a robust foundation for future research aimed at enhancing ALLM capabilities in multi-audio processing. Future work includes introducing more complex multi-audio tasks that involve more than two audios and expanding the MALLM training dataset to cover more scenarios.

Limitations

In this paper, we propose the first benchmark to evaluate the multi-audio analysis ability of ALLMs. We also develop a novel MALLM, which outperforms existing ALLMs on multi-audio processing. Yet, there are several limitations to this work. Firstly, considering the large performance gap between open-source and proprietary models, we deliberately design the tasks in MAE with a simple nature to obtain meaningful results from open-source

ALLMs. Therefore, while MAE poses large challenges to various open-source ALLMs, some tasks are too simple for proprietary ALLMs like Gemini. In future work, we plan to explore more complex tasks, such as speech recognition (Gao et al., 2024b; Yue et al., 2024a), compositional reasoning (Ghosh et al., 2024), speech synthesis (Gao et al., 2024a) and lyrics transcription (Gao et al., 2023b). Secondly, the proposed MAE benchmark currently covers English data. Extending the MAE to the multilingual scenario is also an important future direction to ensure that it's comprehensive and applicable across diverse linguistic contexts. Thirdly, the proposed MALLM is currently trained on a relatively small scale of synthetic data due to computational resource constraints. We aim to further enhance the MALLM with large-scale, diverse training data, enabling its application to more challenging domains like singing (Gupta et al., 2019) and music (Gao et al., 2023a) in future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, et al. 2024. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Xiaoxue Gao, Yiming Chen, Xianghu Yue, Yu Tsao, and Nancy F Chen. 2024a. Ttslow: Slow down text-to-speech with efficiency robustness evaluations. *arXiv preprint arXiv:2407.01927*.

Xiaoxue Gao, Chitralekha Gupta, and Haizhou Li. 2023a. Polyscriber: Integrated fine-tuning of extractor and lyrics transcriber for polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1968–1981.

Xiaoxue Gao, Zexin Li, Yiming Chen, Cong Liu, and Haizhou Li. 2024b. Transferable adversarial attacks against asr. *IEEE Signal Processing Letters*.

Xiaoxue Gao, Xianghu Yue, and Haizhou Li. 2023b. Self-transriber: Few-shot lyrics transcription with self-training. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 2024c. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. *arXiv preprint arXiv:2409.10157*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, Ramaneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Compa: Addressing the gap in compositional reasoning in audio-language models. In *The Twelfth International Conference on Learning Representations*.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*.

Chitralekha Gupta, Karthika Vijayan, Bidisha Sharma, Xiaoxue Gao, and Haizhou Li. 2019. Nus speak-to-sing: A web platform for personalized speech-to-singing conversion. In *INTERSPEECH*, pages 2376–2377.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. 2023. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, et al. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

Chien-Yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung yi Lee. 2023a. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. *arXiv arXiv:2309.09510*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. 2023b. Sparkles: Unlocking chats across multiple images for multi-modal instruction-following models. *arXiv preprint arXiv:2308.16463*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: 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)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Forty-first International Conference on Machine Learning*.

Lori F Lamel, Robert H Kassel, and Stephanie Seneff. 1989. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Speech Input/Output Assessment and Speech Databases*.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. 2023. Textbind: Multi-turn interleaved multimodal instruction-following. *arXiv preprint arXiv:2309.08637*.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2024. Fine-tuning multimodal LLMs to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech Model Pre-Training for End-to-End Spoken Language Understanding. In *Proc. Interspeech 2019*, pages 814–818.

Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, et al. 2021. Multilingual spoken words corpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2023. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Douglas B. Paul and Janet M. Baker. 1992. The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste

Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530.*

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925.*

J. Salamon, C. Jacoby, and J. P. Bello. 2014. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. PandaGPT: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 11–23, Prague, Czech Republic. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations.*

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805.*

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288.*

Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Rethinking STS and NLI in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 965–982, St. Julian's, Malta. Association for Computational Linguistics.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209.*

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519.*

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv arXiv:2402.07729.*

Xianghu Yue, Xiaoxue Gao, Xinyuan Qian, and Haizhou Li. 2024a. Adapting pre-trained self-supervised learning model for speech recognition with light-weight adapters. *Electronics*, 13(1):190.

Xianghu Yue, Xiaohai Tian, Malu Zhang, Zhizheng Wu, and Haizhou Li. 2024b. Coavt: A cognition-inspired unified audio-visual-text pre-training model for multimodal processing. *arXiv preprint arXiv:2401.12264.*

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226.*

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2024. MMICL: Empowering vision-language model with multimodal in-context learning. In *The Twelfth International Conference on Learning Representations.*

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV).*

A Evaluation Prompt

We utilize GPT-4 to automatically give binary quality labels for open-ended generation tasks. To get more stable and consistent predictions from GPT-4, we produce 3 predictions from GPT-4 through repeatedly sampling and take the majority vote as the final prediction. The GPT-4 automatic evaluation prompts for different tasks are listed:

- **Speech comparison:**

```
Given the transcriptions of two
speeches.
Speech 1: [[Speech 1]]
Speech 2: [[Speech 2]]
Determine whether the below response
correctly captures the difference
between speech 1 and speech 2.
[[Response]]
Your answer should be a single "Yes"
or "No". Do not output anything else.
```

- **Dialogue response generation:**

```
Given the first two utterances from
a conversation between speaker 1 and
speaker 2.
Speaker 1: [[Speech 1]]
Speaker 2: [[Speech 2]]
For the question:
What would be the possible utterance
3 from speaker 1 as a response to
utterance 2 from speaker 2?
Determine whether the below response
correctly answers the question.
[[Response]]
Your answer should be a single "Yes"
or "No". Do not output anything else.
```

- **Story generation:**

```
Given the story:
[[Response]]
Determine whether the below two sounds
are presented in the story.
Sound1: [[Speech 1]]
Sound2: [[Speech 2]]
Answer "Yes" if both sounds are
presented in the story.  Otherwise,
answer "No".
Your answer should be a single "Yes"
or "No". Do not output anything else.
```

For all tasks, "No" indicates that the ALLM's response is inappropriate, while "Yes" indicates that the response is suitable.

B   MALLM Data

In Tab. 7, we provide the example responses of possible input audio pair combinations. In the first four cases, the ALLM is asked to describe the differences between two audios. For the last cases, where there's a common sound event occur in both audios, we also ask the ALLM to answer the common sound event in two audios.

The data statistics of the final MALLM training data is listed in Tab. 8. We use the test split of data source for all tasks in MAE benchmark, and use the train split of data source to construct our training data. For human validation of the LLM evaluator, we provide the human annotators with the same instructions given to the LLM evaluator. In addition, both raw audios and ground truth labels are provided to the human annotators.

C   Data Synthesis Prompt

We utilize LLM to obtain sentence pairs with the below four different prompts:

- **Add word:**

Audio Pair	Example Response
A1 vs. A2	Both audios have the same sound event of [barking].
A1 vs. A3	Audio1 has sound of [barking], while Audio2 has sound of [raining].
A1 vs. A1+A3	Both audios have sound of [barking], while Audio2 has additional sound of [raining].
A1+A3 vs. A1+A4	Both audios have sound of [barking], while Audio1 has additional sound of [raining] and Audio2 has additional sound of [clapping].
A1+A3 vs. A1+A4	The common sound event in two audios is [barking].

Table 7: Example response output for audio discriminative training. The A1,A2,A3,A4 refer to audio clips in Fig. 3. The [x] is the ground truth sound event label.

Data	Source	# Samples
Speech pairs	BookCorpus (Zhu et al., 2015)	24K
Sound pairs	VGGSound (Chen et al., 2020)	26.9K
Speech recognition	Librispeech (Panayotov et al., 2015)	28.5K
Sound classification	VGGSound (Chen et al., 2020)	20.7K
Audio caption	AudioCaps (Kim et al., 2019)	5K
Total		105.1K

Table 8: MALLM training data statistics. The first block shows the synthetic multi-audio data, and the second block shows the single-audio data.

```
Given the below sentence:
[[Sentence]]

You need to modify this sentence to a
new one by adding several words. The
number of added words should be less
than 3.
Note that adding is the only operation
you can do.
The  new  sentence  needs  to  be
semantically  correct  without  error.
Try to be creative.
Your output should be the new sentence
only without anything else.
If you cannot generate a new sentence
meeting the above criteria (less
than 3 added words and semantically
correctness),  output  a  single  word
"None".
```

- **Delete word:**

```
Given the below sentence:
[[Sentence]]

You need to modify this sentence to
a new one by deleting several words.
The number of deleted words should be
less than 3.
Note that deletion is the only
operation you can do.
The new sentence needs to be
semantically correct without error.
Try to be creative.
Your output should be the new sentence
only without anything else.
If you cannot generate a new sentence
meeting the above criteria (less
than 3 deleted words and semantically
correctness), output a single word
"None".
```

• **Modify word:**

```
Given the below sentence:
[[Sentence]]

You need to modify this sentence to
a new one. The edit distance between
the generated sentences and the old
one needs to be less than 5.
The new sentence needs to be
semantically correct without error.
Try to be creative.
Your output should be the new sentence
only without anything else.
If it's impossible to have a new
sentence meeting the above criteria
(edit distance and semantically
correctness), output "None".
```

• **Change structure:**

```
Given the below sentence:
[[Sentence]]

You need to derive a new sentence by
altering the structure of the given
sentence.
Remember to keep the wording usage
unchanged.
The difference between two sentences
need to be less than 5 words.
The new sentence needs to be
semantically correct without error.
Try to be creative.
Your output should be the new sentence
only without anything else.
If it's impossible to have a new
sentence meeting the above criteria
(less than 5 words difference and
semantically correctness), output
"None".
```

D Case Study

We compare the generated responses from Qwen-Audio and MALLM on dialogue response generation and speech comparison in Tab. 9. Similar to our findings from Tab. 5, the Qwen-Audio fails to combine the context from two input speeches to infer the correct response. On the contrary, the proposed MALLM can handle the multi-audio scenarios.

DailyTalk Response Generation	
Speech1	I want to buy some jewellery.
Speech2	What kind of jewellery do you like to have?
Qwen-Audio	I want to buy some jewelry
MALLM	I like to have a diamond ring.
LibriSpeech Speech Comparison	
Speech1	creatures attacked each other
Speech2	huge creatures attacked each other
Qwen-Audio	The content difference between speech 1 and speech 2 is that in speech 1, "creature" is mentioned, in speech 2, "huge" is mentioned.
MALLM	The difference between Speech 1 and Speech 2 lies in the adjective used to describe the creatures. In Speech 1, the creatures are simply mentioned without any specific description, implying they could be of any size or type. In Speech 2, the adjective "huge" is used to specifically describe the creatures, suggesting that they are very large. This additional detail in Speech 2 emphasizes the size of the creatures, potentially altering the reader's perception of the scene's intensity or the creatures' capabilities.

Table 9: Case study on dialogue response generation and speech comparison.