GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities

Anonymous ACL submission

Abstract

Perceiving and understanding non-speech sounds and non-verbal speech is essential to making decisions that help us interact with our surroundings. In this paper, we propose GAMA, a novel **G**eneral-purpose Large **A**udio-Language **M**odel (LALM) with **A**dvanced Audio Understanding and Complex Reasoning Abilities. We build GAMA by integrating an LLM with multiple types of audio representations, including features from a custom Audio Q-Former, a multi-layer aggregator that aggregates features from multiple layers of an audio encoder. We fine-tune GAMA on a large-scale audio-language dataset, which augments it with audio understanding capabilities. Next, we propose CompA-R (Instruction-Tuning for **Comp**lex **A**udio **R**easoning), a synthetically generated instruction-tuning (IT) dataset with instructions that require the model to perform complex reasoning on the input audio. We instruction-tune GAMA with CompA-R to endow it with complex reasoning abilities, where we further add a soft prompt as input with high-level semantic evidence by leveraging event tags of the input audio. Finally, we also propose CompA-R-*test*, a human-labeled evaluation dataset for evaluating the capabilities of LALMs on open-ended audio question-answering that requires complex reasoning. Through automated and expert human evaluations, we show that GAMA outperforms all other LALMs in literature on diverse audio understanding tasks by margins of 1%-84%. Further, GAMA IT-ed on CompA-R proves to be superior in its complex reasoning and instruction following capabilities [1].

1 Introduction

Large Language Models (LLMs) possess impressive abilities to understand and reason about the world through language (Zhao et al., 2023). While spoken language understanding tasks, like automatic speech recognition, have had a long history of benefiting from language comprehension with (L)LMs (Watanabe et al., 2018; Hu et al., 2024), the ability to improve the perception and understanding of non-speech sounds and non-verbal speech through language has been less explored (from hereon we refer to these kinds of audios or sound as "audio" in the paper). Beyond visual and language perception, the ability to understand audio is unarguably important and necessary for autonomous agents to interact with the world.

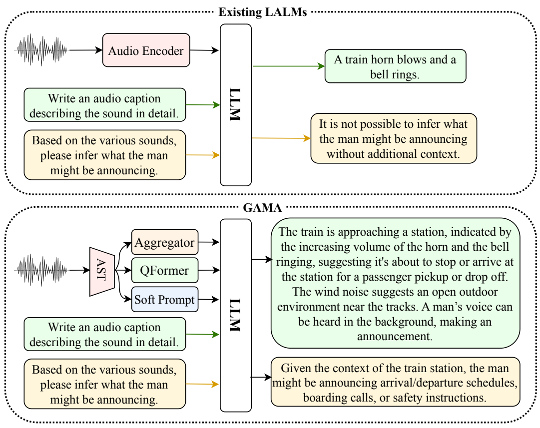


Figure 1: Comparison of existing LALMs (LTU by Gong et al. (2024) here) and GAMA. With improved audio understanding abilities (via diverse audio feature integration) and training on our proposed CompA-R, GAMA can provide more detailed captions of input audio and is also able to answer questions regarding it that demand complex reasoning.

Contrastive Language Audio Pre-training (CLAP) (Elizalde et al., 2023a) was one of the first Audio-Language Models (ALM) to improve audio understanding through a language interface. Following this, several attempts have been made to improve CLAP and its reasoning abilities (Ghosh et al., 2024b). On the other hand, Deshmukh et al. propose Pengi, a pre-trained decoder-only LLM coupled with an audio-encoder, that can solve all kinds of audio tasks by framing them as open-ended text-generation tasks. Similarly, Large Audio Language Models (LALMs) like

[1]We will open-source code and data on paper acceptance

LTU (Gong et al., 2024) and SALMONN (Tang et al., 2024) follow a similar architecture and attempt to solve audio tasks by empowering the model with instruction following capabilities (Wei et al., 2022). Specifically, all audio tasks are first framed in instruction-response pairs. The model is then fine-tuned on these pairs to learn audio reasoning and, thereby, instruction following. As an emergent ability, these models also show remarkable capabilities in open-ended question answering by reasoning over the input audio. However, two significant problems still persist: (1) All these models employ simple connection modules between the audio encoder and the language decoder to enable the latter with audio understanding capabilities. This hinders comprehensive multimodal connection and alignment, thereby increasing the risk of hallucinations and leading to suboptimal performance (Liu et al., 2023a). (2) Complex reasoning with LALMs is still under-explored. While these models excel at audio event detection (in various forms like captioning, event classification, etc.) and information-seeking questions (e.g., close-ended audio questions like *"How many birds are squawking?"*), they fail to provide a faithful response for questions involving complex reasoning like *"Identifying the context of laughter and its relationship with the automotive sounds in the recording. Draw a conclusion on the possible scenario occurring."*. We define complex reasoning for LALMs in Section 3.2 and show examples in Fig. 1 and Fig. 4.

Main Contributions. Our primary contributions are as follows:

- **A Novel LALM.** We introduce GAMA, an LALM with advanced audio understanding and complex reasoning abilities. To improve audio perception and understanding abilities, we propose integrating an LLM with multiple types of audio features that encode diverse aspects of information about the input audio. Specifically, we couple the output features from an Audio Q-Former and an Audio Spectrogram Transformer (AST) (Gong et al., 2021), where the AST is further equipped with an *aggregation module*. While the Audio Q-Former possesses impressive semantic generalization capabilities (Li et al., 2023), the AST possesses strong knowledge of surface-level audio properties. Additionally, inspired by the fact that different layers in audio models learn audio information at different scales (Singla et al., 2022), the aggregation module aggregates the features from multiple layers of AST, which helps encode diverse knowledge. Both representations are passed through MLP layers that connect these features into the word embedding space before adding them as the prefix. As a result, GAMA possesses improved audio understanding capabilities by moving away from the simple coupling of audio encoders and linear layers commonly employed as connection modules to align the audio and textual modalities, which generally suffer from comprehensive multimodal alignment (Liu et al., 2023a). GAMA is first fine-tuned on a large-scale audio-language corpus, and the resulting model outperforms all other models on standard audio and music understanding benchmarks.

- **A Novel Instruction Tuning Dataset.** To endow an LALM with complex reasoning abilities, we propose CompA-R, a dataset synthetically generated with LLMs with multi-aspect information and human-written in-context examples. Specifically, we prompt GPT to synthesize an instruction-response pair by guiding it with various metadata related to the audio.

- **A Novel Evaluation Dataset.** To evaluate an LALM's complex reasoning abilities, we develop CompA-R-*test*, a human-labeled benchmark. Specifically, CompA-R-*test* evaluates an LALM on open-ended AQA that demands complex reasoning over the audio. GAMA-IT (GAMA fine-tuned on CompA-R) shows significant improvements on CompA-R-*test* over all other baselines from literature.

2 Related Work

Large Multi-Modal and Audio-Language Models. Prior to the exploration of LLMs as efficient reasoners, encoder-based multi-modal language models, trained to learn a shared space between language and other modalities, have shown great promise. For example, CLAP, inspired by CLIP (Radford et al., 2021) in vision, showed state-of-the-art performance on audio-language tasks like retrieval, zero-shot classification, etc.

LLMs pre-trained at an incredible scale with the next token prediction objective implicitly com-

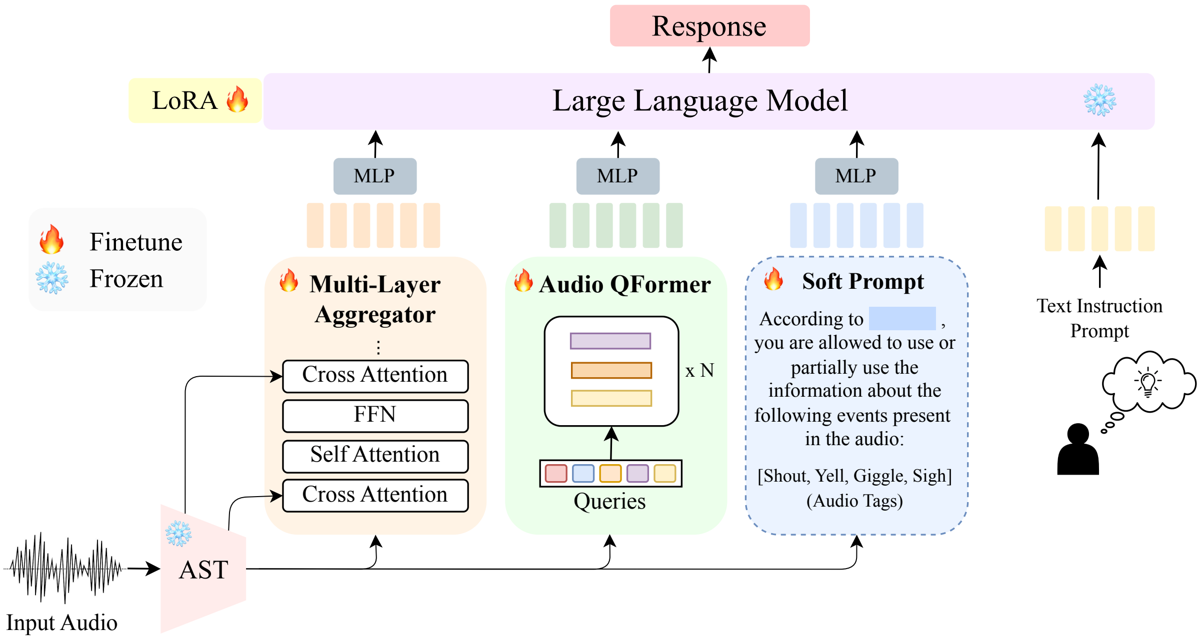


Figure 2: Illustration of GAMA. The GAMA architecture involves integrating various audio features with a text-only (pre-trained) LLM. The audio features are added as a prefix to the text instruction (by the user), and the LLM responds using a text output. We feed GAMA with 3 different types of audio features: ① The Multi-Layer Aggregator that takes as input features from the last and multiple other layers of the AST and calculates cross-attention to output a feature that encodes diverse surface features from audio. ② The pre-trained Audio Q-Former takes as input the last layer features of AST and outputs a feature that encodes the audio into a semantically rich and generalized space. ③ After fine-tuning, we instruction-tune this model on our proposed CompA-R and further feed the model with a soft prompt with audio tags from AST (with an added classification head). This additional information serves as high-level semantic knowledge to improve complex reasoning abilities.

press world knowledge in their parameters (Zhao et al., 2023). These models learn general-purpose representations, which can then be *aligned* with the desired response characteristics (Zhang et al., 2023). Instruction Tuning (IT), the process of fine-tuning an LLM with instruction-response pairs, has proved to be one of the most popular forms of alignment. Recent work shows that LLMs can also be instruction-tuned for multi-modal alignment. LLaVa (Liu et al., 2024), a pioneering work on multi-modal vision-language alignment, showed that fine-tuning an LLM on visual instruction-response pairs with additional vision features as prefix can endow the model with visual reasoning and understanding abilities. Several works following LLaVa improve aspects of LVLMs and have achieved impressive performance on several vision-language tasks (Zhang et al., 2024). On the other hand, LALMs like LTU and SALMONN showed impressive performance on several audio-language tasks by reasoning over the audio. Though these models extensively evaluate several closed- and open-ended tasks, their ability to perform complex reasoning is largely under-explored.

Instruction Tuning and Complex Reasoning. IT-based alignment has also shown significant improvements for LLMs on Natural Language Understanding tasks, unlocking impressive capabilities (Bubeck et al., 2023), suggesting that fine-tuning is key to building and improving LLM-based agents. Very recently, (Xu et al., 2024) and (Cui and Wang, 2024) show that well-curated IT data can improve various reasoning capabilities in LLMs, like logical, mathematical, complex reasoning, etc. More specifically, IT teaches LLMs better and more effective methods to reason about a problem, presented in the input instruction (like step-by-step reasoning (Kojima et al., 2022)).

3 Methodology

In the next sub-sections, we first describe the GAMA architecture and its components in detail, followed by fine-tuning GAMA on audio-language pairs, CompA-R creation, and instruction-tuning GAMA on CompA-R.

3.1 GAMA Architecture

Fig. 2 illustrates the architecture of GAMA. GAMA builds on the same base architecture proposed in prior works (Gong et al., 2024) but introduces several novel components for improving audio perception. More specifically, we feed the pre-trained LLM with features from multiple audio encoders, including a pre-trained Audio-Q-Former and a pre-trained AST that encode diverse audio knowledge. Additionally, unlike prior work, we do not just use the last layers of the AST but couple it with a multi-layer aggregator that takes features from multiple layers as input and outputs a feature that is aware of various low-level and high-level properties of the input audio. Finally, to endow the model with effective complex reasoning abilities, we employ AST again to extract high-level semantic knowledge, i.e., audio event tags, as supplementary information.

3.1.1 Audio Spectrogram Transformer (AST)

Audio Spectrogram Transformer (AST), was one of the first attempts to model audio signals with a pure Transformer network. We employ an AST model fine-tuned on the AudioSet dataset. AST has been employed as an audio encoder and a feature extractor in a wealth of prior works due to its high informativeness (Gong et al., 2023, 2024). To extract the last-layer features, we drop the audio classification head and employ it only for event classification for soft prompts.

3.1.2 Audio Q-Former

Motivation. Our primary goal is to integrate `GAMA` with an audio encoder that possesses strong semantic generalization capabilities for any input audio. Prior work has extensively explored CLAP-style training for learning audio-language encoders. However, other methods and architectures have rarely been explored. As a more powerful alternative, we explore the Q-Former architecture proposed by (Li et al., 2023). We show in Table 1 that Q-Former outperforms other audio-language models and achieves SOTA on audio understanding benchmarks.

Architecture. The architecture of our Audio Q-Former is based on the Querying Transformer proposed in Li et al. (2023), which is initialized from BERT (Devlin et al., 2018) and has Q querying tokens. We employ AST as the audio encoder (in place of the ViT-based vision encoder) and keep the rest of the architecture the same. Similar to the original implementation, we train the model in two stages. For the first stage, we solve three tasks, namely the Audio-Text Matching loss, the Audio-Grounded Text Generation loss, and the Audio-Text Contrastive Learning loss. For the second stage, we employ LLaMa-2 $_{7\mathrm{B}}$ as the language decoder and solve the language-modeling loss. For training, we use 2.5M+ audio-caption pairs (detailed in Section E.2). For architectural details, we refer our readers to Li et al. (2023).

Training with Caption Augmentation. Additionally, due to the lack of large-scale audio caption pairs, we adopt a *caption-augmentation* methodology to augment the existing audios with diverse additional captions. More specifically, we instruct an LLM to generate k rewrites of the original caption. We employ two different prompts that rewrite the input caption with two different objectives:

Prompts. For Prompt 1, our primary aim is that the resultant rewrite should describe each acoustic event in the caption similarly but more vividly. These augmentations help the model learn various distinctive characteristics of the audio concepts corresponding to the acoustic events. For Prompt 1, our primary aim is such that the resultant rewrite should describe each acoustic event in the caption differently from the original caption. These augmentations aid the model in understanding the diverse linguistic expressions that can describe a single audio concept. We show examples below: (more examples in Table 13):

> ***(1) Original Caption:*** Someone made a cool vocal for a dubstep track.
> ***(1) Rewritten Caption by Prompt 1:*** A captivating vocal performance ignites the dubstep track, delivering a hypnotic and enthralling sound that reverberates through the air.
> ***(1) Rewritten Caption by Prompt 2:*** The dubstep track features a slick, stylish vocal performance that adds a layer of sophistication to its heavy beats and basslines.
> ***(2) Original Caption:*** Someone eating crisps and talking.
> ***(2) Rewritten Caption by Prompt 1:*** Crunchy crisps mingle with the sound of a lively conversation, creating a cozy and intimate atmosphere.
> ***(2) Rewritten Caption by Prompt 2:*** The crunch of crisps and the rustle of papers create a cozy, intimate atmosphere, accompanied by the gentle hum of a conversation.

During training, for each audio sample, we choose the original caption with a probability $p = 0.4$ or one of the rewritten versions (with a probability $1 - p$), where each rewritten caption has an equal probability of selection. Both instrcutions are provided in Appendix B. We employ LLaMa-2-13B (Touvron et al., 2023) with human-written in-context examples. We randomly sample 5 in-context examples from a collection of 50.

3.1.3 Multi-Layer Aggregator

Motivation. To extract additional details about the input audio, we devise a multi-layer aggregator that integrates multi-level hidden features of the pre-trained AST. Although AST has a global reception field in all layers, different layers learn auditory information at different scales (Singla et al., 2022), i.e., the middle layers encode more generic features (e.g., basic sounds, textures), while deeper layers capture high-level concepts (e.g., speech intonations, complex sound patterns). By aggregating these features, the multi-layer aggregator outputs features that encode a more holistic and fine-grained understanding of the audio. Thus, our multi-layer aggregator makes fine-grained auditory knowledge more likely to be learned while training.
Architecture. Our multi-layer aggregator is a transformer-style network consisting of two transformer layers for aggregating the hidden features of the audio encoder. Given the hidden features A_j and A_k from the middle layers in the audio encoder, the aggregation module uses two blocks to sequentially integrate the former two features with the last layer feature A_i. Each block $\mathcal{B}$ is composed of self-attention, cross-attention, and Feed-forward network (FFN) arranged in a sequential manner. Finally, the output features $\bar{A}$ is generated as follows,

$$\bar{A} = \mathcal{B}_2 \left(\mathcal{B}_1 \left(A_i ; A_j \right) ; A_k \right) \quad (1)$$

$$\mathcal{B}(X;Y) = \mathrm{FFN}(\text{Cross-Attn}(\mathrm{Attn}(X), Y)). \quad (2)$$

In practice, we employ j = 4 and k = 8 from AST as our input to the multi-layer aggregator.

3.1.4 Soft Prompt

Motivation. Though models like AST and Audio Q-Former have shown much promise in audio tasks, a major problem still exists: real-world audio generally has multiple and overlapping acoustic events, and understanding all such events from model features proves to be inherently complex (Ghosh et al., 2024b). This eventually leads to sub-optimal performance for complex reasoning, where the explicit knowledge of *plausible* acoustic events in the audio can improve model responses. Thus, to improve fine-grained audio perception capabilities, we augment `GAMA` with high-level semantic understanding of the input audio. To do this, we employ an off-the-shelf audio model to extract high-level semantic knowledge, i.e., audio event tags, as supplementary information. However, as audio event classification is not a solved problem, errors in tag predictions are inevitable. Thus, to mitigate the potential adverse effects of inaccurate predictions, we are inspired by prompt tuning to introduce a soft prompting technique that enables the model to utilize the embedded tags within the instructions adaptively.
Architecture. Fig. 2 shows an example of how we design our soft prompt together with an instruction. Specifically, we construct a fixed instruction template where we add the audio event tags along with the soft prompt, where the soft prompt is a trainable vector. In contrast to standard prompt tuning, where the model activations are generally steered towards completing the task for which the prompt is optimized, in our version the direction is specified by a tailored input sentence, "According to <hint>, you are allowed to use or partially use the following tags:", and "<hint>" will be replaced by the soft prompt. This design allows us to select valuable information from tags adaptively rather than serving a specific task, as seen in standard prompt tuning methods. We only employ the soft prompt in the instruction tuning for complex reasoning step and not in the fine-tuning step. We provide a rationale in Appendix C.1.

3.1.5 Connection Module

We employ a multi-layer perceptron (MLP) to connect audio features into the word embedding space. All features are passed through separate MLP layers before being added as prefixes to word embeddings of the text instruction prompt.

3.2 `CompA-R`

Motivation. We define complex reasoning as the capability of an LALM to understand the input audio, every individual acoustic event in the audio, and reason the corresponding scene in which the audio might have occurred, such that it can infer nuanced relationships between them and its underlying contexts, thereby enabling it to draw sophisticated conclusions. We design `CompA-R` with the primary goal of endowing LALMs with complex reasoning abilities. We are motivated by the primary finding that current SOTA LALMs can only perform well in prompts that require describing the audio (e.g., *Describe the audio*) or reasoning-based prompts where identifying the acoustic events present in the audio would suffice for a faithful response (e.g., *What type of video can this audio be used for dubbing?*). However, when posed with complex reasoning questions, these models often hallucinate or fail to provide a faithful response (see Fig. 4). Inspired by a wealth of prior work that shows how IT on well-curated datasets can align model behaviors for the execution of novel skills like reasoning and complex problem solving (Xu et al., 2024), we propose a systematic multi-stage pipeline to synthesize instruction-response pairs for `CompA-R`. `CompA-R` trains a model to engage in complex reasoning by querying it with instructions that cannot be directly inferred by identifying individual audio events and would require analyzing each event and its context in relation to other scene elements and world knowledge.

Synthesis Pipeline. We employ the AudioSet-strong subset to synthesize `CompA-R`. Our data synthesis pipeline consists of 3 stages: ***i) Caption Generation.*** To generate a caption that is aware of both the audio and the visual scene, we feed GPT-4 with multiple types of information about the audio and its corresponding video. These include a caption of the middle frame of the video generated using BLIP-2 (Li et al., 2023), objects in the frame identified using Grounding DINO (Liu et al., 2023c), image labels for the frame using the ImageNet (Deng et al., 2009) ontology obtained from CLIP, environment context using PlaceCNN (Zhou et al., 2017), caption of the audio obtained using RECAP (Ghosh et al., 2024a) and audio event tags using the AudioSet ontology obtained from AST. Finally, we prompt GPT-4 to aggregate these descriptions into a comprehensive caption. ***ii) Dataset Synthesis.*** We pass the generated caption together with the ground-truth acoustic event information and their corresponding time slices to GPT-4. We prompt GPT-4 with 3 human-written exemplars (which are randomly sampled from a pool of 50 exemplars) to synthesize an instruction-response pair. The exemplars and prompt are designed such that the synthesized instructions demand complex reasoning. We synthesize a total of 25000 instruction-response pairs. ***iii) Human Verification.*** We discard instructions due to untended noise and hallucinations. We, the authors of this paper, manually verify a subset of `CompA-R` corresponding to 500 unique audios for creating the test set, i.e., `CompA-R`-*test*. The remainder of the synthesized dataset is used as the training set. We describe the process and annotation details further in Appendix G.1. This finally led to 200,234 unique pairs in training and 1,561 in testing.

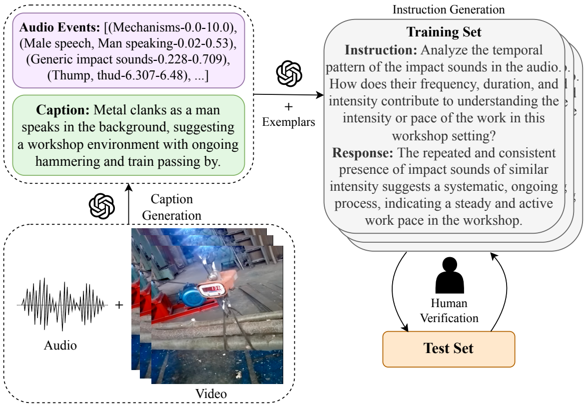


Figure 3: **Pipeline for synthesizing CompA-R.** For an audio in the AudioSet-strong dataset, we first use the audio and its corresponding video to generate a caption (described in Section 3.2). This caption is then fed into GPT-4 together with the ground-truth time slices for each event in the audio (available for AudioSet-strong). We then instruct GPT-4 to generate instruction-response pairs where the instruction is accompanied by human-written exemplars (prompt in Fig. 5). To generate the test set, we perform expert human verification for a subset of the generated dataset.

3.3 Training

Fine-tuning. We fine-tune `GAMA` on the OpenAQA training set released by Gong et al. (2024). We use a faction of all the instances due to the unavailability of the entire AudioSet and resource constraints. Dataset details are provided in Appendix H.1. Additionally, we augmented OpenAQA with 4 more datasets, including MusicCaps, MusicQA, NSynth, and Magna, to improve its music understanding capabilities. For fine-tuning, we follow the exact same 4-stage method proposed by Gong et al. (2024) where all parameters of all encoders are trainable, and we train only the LoRA modules of the LLM. We request our readers to refer to Gong et al. (2024) for more details.

Instruction Tuning on `CompA-R`. Post fine-tuning, we instruction-tune `GAMA` on `CompA-R` to endow it with complex reasoning abilities. Following common conventions (Liu et al., 2023b), we fine-tuned only the LoRA modules. We call the Instruction Tuned `GAMA` as `GAMA-IT`. Although fine-tuning on AQA also endows `GAMA` with instruction-following capabilities, `CompA-R` differs in the nature of training instances (thereby the capabilities it endows) and thus we differentiate with such a naming convention for ease of reading.

Model	ESC50$^{\#}$ (Acc)	DCASE$^{\#}$ (Mi-F1)	VS† (Acc)	TUT† (Acc)	BJO† (Acc)	VGG (Acc)	FSD (mAP)	NS$_{\text{ins.}}$ (ACC)	NS$_{\text{src.}}$ (ACC)	GTZAN† (ACC)	MSD† (ACC)	AudioSet (mAP)	Classif. Avg.	AudioCaps (SPICE)	Clotho (SPICE)	Cap. Avg.	ClothoAQA (ACC)
Audio-Language encoder-based models. They are generalizable to unseen labels, but a pre-defined label set is required for inference.																	
AudioCLIP	69.4	-	-	-	-	- -	-	-	-	-	-	-	-	-	-	-	
CLAP (Elizalde et al., 2023a)	82.6	30.0	48.4	29.6	47.5	24.0	30.2	22.7	16.4	25.0	44.0	5.8	29.4	-	-	-	-
CLAP (Wu* et al., 2023a)	89.1	31.3	47.1	35.6	48.0	26.3	30.8	25.2	18.9	26.3	46.9	6.2	36.0	-	-	-	-
CompA-CLAP	90.1	30.6	49.5	35.8	48.2	29.5	31.5	24.9	17.0	26.1	46.2	6.2	36.3	-	-	-	-
Audio Q-Former (w/o rw) *(ours)*	91.9	31.1	49.9	38.9	50.4	33.2	34.7	27.5	22.0	30.4	48.3	8.2	38.9	-	-	-	-
CLAP (w/ rw) *(ours)*	90.7	30.9	50.7	36.2	53.4	30.7	37.1	24.1	17.4	27.6	48.1	6.1	37.7	-	-	-	-
Audio Q-Former (w/ rw) *(ours)*	92.4	32.5	50.2	39.1	51.5	35.1	35.3	29.2	22.3	31.3	47.5	8.9	39.6	-	-	-	-
Audio-Language generation-based models. They directly output label names and do not need a pre-defined label set is needed at inference.																	
Qwen-Audio-Chat	71.7	32.4	74.2	16.9	50.8	17.5	39.8	30.2	41.3	41.6	69.1	13.4	41.1	14.7	9.8	12.3	32.3
LTU	81.7	37.5	**53.3**	19.9	67.8	50.3	43.9	28.0	41.8	9.9	74.2	18.3	42.4	16.9	11.7	15.8	25.1
SALMONN	16.4†	18.0†	16.9†	7.8†	25.0†	23.3†	22.1†	16.2†	33.7†	10.1†	28.8†	13.4†	17.9	8.3	7.6	8.0	23.1†
Pengi	80.8†	29.6†	46.4†	18.4†	47.3†	16.6†	35.8	39.2	46.0	11.9	**93.0**	11.5	39.7	12.7	7.0	9.9	63.6
AudioGPT	41.3	20.9	35.8	14.9	21.6	5.6	18.8	40.9	15.6	11.9	28.5	12.7	22.4	6.9	6.2	6.6	33.4
GAMA *(ours)*	**82.6**	**38.4**	52.4	21.5	**69.5**	**52.2**	**47.8**	**63.9**	**99.5**	**13.8**	85.6	**19.2**	**53.9**	**18.5**	**13.5**	**16.0**	**71.6**
w/o AST & Aggregator	80.5	36.9	51.6	19.2	66.2	50.8	45.3	62.4	89.6	11.6	83.2	17.3	51.2	17.2	12.4	14.8	68.3
w/ Last Layer Features	81.3	37.6	50.2	20.4	68.2	51.7	45.8	62.6	92.3	11.2	81.5	18.1	51.7	17.7	12.8	15.3	69.5
w/o Audio Q-Former	79.7	37.4	51.3	20.2	68.0	51.6	46.4	60.1	90.4	11.6	79.8	18.4	51.2	16.9	11.9	14.4	61.2
w/ CLAP	81.8	38.4	52.2	**21.6**	69.1	52.0	47.5	58.8	99.5	12.4	77.9	19.0	52.5	17.2	13.1	15.1	66.4

Table 1: Comparison of GAMA with baselines on evaluation datasets described on close-ended general audio and music understanding benchmarks. GAMA outperforms most ALMs on most settings. † and $^{\#}$ indicate zero-shot and weak zero-shot, respectively. **Note:** Qwen-Audio-Chat does not provide training details. We also mark baseline values which are zero-shot.

	CompA-R-*test* (GPT-4/Human)				OpenAQA				Dense Captioning		
Models	Clarity	Correctness	Engagement	Avg.	Clarity	Correctness	Engagement	Avg.	AudioCaps	Clotho	Avg.
Qwen-Audio-Chat	3.5 / 3.4	3.3 / 3.4	3.6 / 3.7	3.5 / 3.5	3.6	3.6	3.5	3.6	3.8	3.6	3.7
LTU	3.5 / 4.0	3.2 / 3.3	3.4 / 3.5	3.4 / 3.6	3.5	3.7	3.5	3.6	3.5	3.6	3.5
SALMONN	2.6 / 2.8	2.4 / 2.3	2.0 / 2.2	2.3 / 2.4	2.4	2.5	2.7	2.5	2.8	3.1	2.9
Pengi	1.8 / 1.6	1.5 / 1.4	1.3 / 1.2	1.5 / 1.4	1.7	1.5	1.4	1.5	2.6	2.8	2.7
AudioGPT	1.3 / 1.4	1.6 / 1.5	1.4 / 1.7	1.4 / 1.5	1.6	1.5	1.5	1.5	2.7	2.9	2.8
LTU w/ CompA-R	3.5 / 4.0	3.2 / 3.3	3.4 / 3.5	3.6 /3.6	3.5	3.7	3.5	3.6	3.7	3.8	3.8
GAMA-IT *(ours)*	**4.3 / 4.5**	**3.9 / 4.1**	**3.9 / 4.3**	**4.0 / 4.3**	**4.0**	**4.2**	**3.8**	**4.0**	**4.3**	**4.1**	**4.2**
w/o Soft Prompt	4.1 / 4.2	3.7 / 3.8	3.6 / 3.4	3.8 / 3.8	3.9	3.8	3.7	3.8	4.1	3.9	4.0
w/o Aggregator	4.0 / 4.2	3.5 / 3.5	3.6 / 3.5	3.7 / 3.7	3.7	3.7	3.5	3.6	3.7	3.8	3.8
w/o Audio Q-Former	3.8 / 3.7	3.4 / 3.6	3.5 / 3.3	3.6 / 3.5	3.4	3.9	3.5	3.6	3.7	3.5	3.6

Table 2: Comparison of GAMA with other baselines on open-ended AQA on OpenAQA, complex open-ended AQA on CompA-R-*test* and Dense Captioning on 500 instances from AudioCaps and Clotho.

3.4 Experimental Setup

Hyper-parameters. For the fine-tuning stage, we follow the exact same hyper-parameter setup proposed by Gong et al. (2024). However, we scale down our batch sizes to 4, 2, 2, and 2 (due to compute constraints) with an effective batch size of 256 in all stages. For Instruction Tuning, we employ a batch size of 2, an effective batch size of 256, and a learning rate of 1e-4. For both training and evaluation, we sampled audio at 16kHz.

Baselines. We compare GAMA with ***i)*** generation-based LALMs: LTU, Qwen-Audio, SALMONN, Pengi and AudioGPT. We only employ the original checkpoints open-sourced by the authors and do not re-train the models due to compute constraints (except LTU, which we retrain on our version of OpenAQA, the same batch size as GAMA, and with LLaMa-2 as the LLM). We do not compare with Audio Flamingo (Kong et al., 2024) as the checkpoint was not available at the time of writing the paper, and we are constrained by compute for training it from scratch. ***ii)*** audio-language encoders: CLAP by Wu* et al. (2023b) and Elizalde et al. (2023b), CompA-CLAP (Ghosh et al., 2024b), AudioCLIP (Guzhov et al., 2021) and Audio Q-Former. For dense captioning and close- and open-ended AQA, we evaluate using GAMA-IT. For all other tasks, we evaluate using the only fine-tuned version of GAMA (rationale in Appendix C).

Evaluation Datasets and Metrics. Evaluation metrics used for all evaluation datasets are mentioned in Table 2 and detailed statistics about each dataset is mentioned in Section H.2. For classification, zero-shot evaluation refers to datasets GAMA that have never been seen during training; weak zero-shot evaluation refers to datasets GAMA that have not been seen in training but are sourced from the same project as part of the training data, and seen datasets refer to datasets GAMA has been trained on. Similar to Deshmukh et al. (2023); Gong et al. (2024), we first caption the audio and retrieve the most similar label using SentenceBERT. We employ either accuracy (Acc), Micro-F1 (Mi-F1), or Mean Average Precision (mAP) for classification evaluation. For captioning, we also propose dense captioning, which evaluates a model for its capability to identify every event in the audio and the context of its occurrence with respect to other events in the audio (more in Section 4). For evaluation, we randomly select a subset of 500 samples from AudioCaps and Clotho. We also employ human evaluation for OpenAQA, CompA-R-*test*, and dense captioning. For human evaluation, we ask human annotators to score the caption on a scale

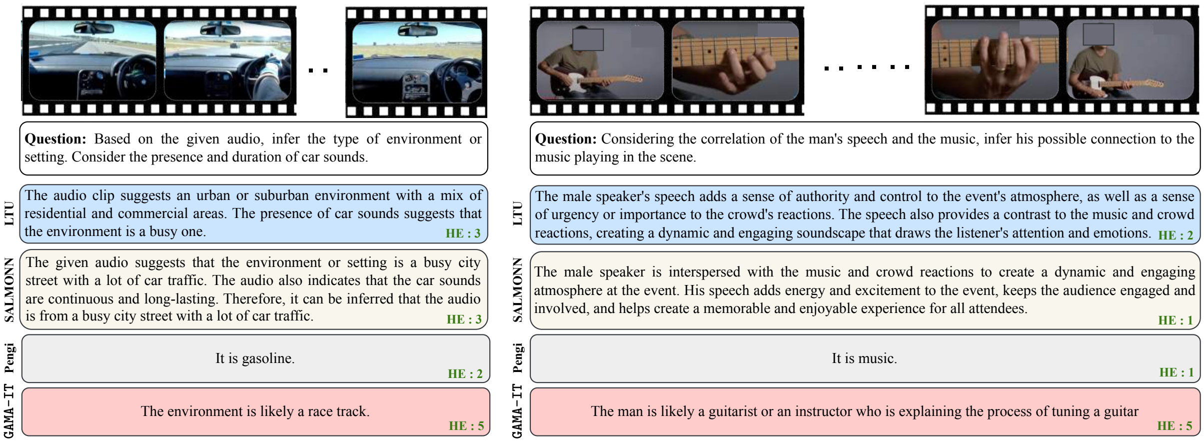


Figure 4: Qualitative comparison of `GAMA` with other baselines on instances from `CompA-R`-*test*. Both instances challenge an LALM with a question about the input audio that requires advanced understanding and complex reasoning regarding the audio and its individual events. `GAMA` is able to provide faithful and accurate responses through improved audio perception and reasoning capabilities. HE refers to scores assigned by human evaluators. Note that the video is only provided for illustration purposes and not provided as input to the LALM. More examples here: https://gamaaudio.github.io/gamaaudio/

of 1-5 and report the score averaged across the 3. More details on recruitment and background of annotators can be found in Appendix D. Finally, due to human evaluation being prohibitively expensive, we also propose an automated evaluation methodology for complex open-ended AQA on `CompA-R`-*test*. We evaluate model responses using text-only GPT-4, where we provide it with the audio caption generated in Section 3.2 and the gold-standard audio event with timestamps (prompt in Appendix B).

4 Results and Analysis

Quantitative Results. Table 1 compares `GAMA` with other baselines on classification and captioning tasks. For zero-shot classification evaluation on VocalSound (VS) (Gong et al., 2022), TUT 2017 (TUT) (Mesaros et al., 2018), Beijing Opera (BJO) (Tian et al., 2014), GTZAN (GTZ) (Park et al., 2022) and Medley-solos-DB (MDB) (Lostanlen et al., 2018), `GAMA` outperforms our baselines by 2%-67%. For weak zero-shot evaluation on ESC-50 (Piczak, 2015) and DCASE2017 Task 4 (DCASE) (Mesaros et al., 2017), `GAMA` outperforms our baselines by 1%-66%. Finally, for in-domain evaluation on VGGSound (VGG) (Chen et al., 2020), FSD50K (FSD) (Fonseca et al., 2021), AudioSet (AS) (Gemmeke et al., 2017) and NSynth (NS) (Engel et al., 2017) `GAMA` outperforms our baselines by 1%-84%. `GAMA` sees the steepest drop in performance when the AST and Aggregator are removed (i.e., only Auio Q-Former is employed).

Table 2 compares `GAMA` with other baselines on AQA (open-ended and complex open-ended) and dense captioning. `GAMA` outperforms all our baselines on all settings. `GAMA` shows absolute improvement of 4% - 50% on OpenAQA, 8% - 58% on `CompA-R`-*test* and 8% - 30% on Dense Captioning. Similar to the tasks in Table 1, performance on benchmarks suffers the most when without the Audio Q-Former (when only the AST and Aggregator are employed). Audio Q-Former proves to especially effective (over employing CLAP) in AQA.

Qualitative Results. Fig. 4 compares `GAMA-IT` against other LALMs from literature with instances from `CompA-R`-*test*. All models compared by default possess audio chat or open-ended AQA capabilities. `GAMA-IT` is able to provide more faithful responses that are both correct and preferred more by humans. We provide additional comparisons in Figs. 8, 9, 10, 11, 12, and our demo page: (where we also show comparisons of dense captioning).

5 Conclusion

In this paper, we propose `GAMA`, an LALM with improved audio perception abilities. We integrate an LLM with multiple types of audio representations, which are responsible for providing diverse knowledge about the input audio. `GAMA` fine-tuned on a mixture of open-source datasets outperforms prior audio-language models by significant margins on 16 datasets spanning 4 tasks. Next, we propose `CompA-R`, an instruction-tuning dataset that we synthesize using a robust pipeline for endowing an LALM with complex reasoning abilities. `GAMA` IT-ed on `CompA-R` outperforms baselines on complex open-ended AQA and dense captioning.

Limitations and Future Work

GAMA and our experimental setup have several limitations, including:

- For the scope of our experiments, we do not evaluate and compare music understanding extensively. We do not do this as we do not train GAMA on diverse and large-scale music datasets. We also acknowledge that it is possible to employ the GAMA architecture for comprehensive music understanding if trained on large-scale music understanding datasets. As part of future work, we plan to release a music-only version of GAMA, similar to Gardner et al. (2024).
- We do not employ larger LLMs, for example, the 13B versions of the LLaMA family, similar to Tang et al. (2024) and Gong et al. (2024), due to compute constraints.
- The audio-encoder(s) in GAMA have more parameters than in our baselines. However, we also acknowledge that this adds to only a fraction of the total parameter count of the LALM.

References

2018. A dump of BBC's sound effects library. This dump was created using the script found at https://github.com/FThompson/BBCSoundDownloader. Identifier: BBCSoundEffectsComplete.

2023. SoundBible - Free Sound Clips, Sound Bites, and Sound Effects. Accessed: 25 September 2023.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset.

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*.

Wanyun Cui and Qianle Wang. 2024. Ada-instruct: Adapting instruction generators for complex reasoning.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks.

Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. 2022. Audio retrieval with wavtext5k and clap training. *arXiv preprint arXiv:2209.14275*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023a. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023b. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. Fsd50k: An open dataset of human-labeled sound events.

Joshua P Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. 2024. LLark: A multimodal foundation model for music.

Jort F Gemmeke, Daniel PW 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. IEEE.

Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Ramani Duraiswami, and Dinesh Manocha. 2024a. Recap: Retrieval-augmented audio captioning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1161–1165.

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

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

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*.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. 2023. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155. IEEE.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022*.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. Audioclip: Extending clip to image, text and audio.

Yuchen Hu, CHEN CHEN, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and Ensiong Chng. 2024. Large language models are efficient learners of noise-robust speech recognition. In *The Twelfth International Conference on Learning Representations*.

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.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Vincent Lostanlen, Carmine-Emanuele Cella, Rachel Bittner, and Slim Essid. 2018. Medley-solos-db: a crosscollection dataset for musical instrument recognition. *Zenodo*.

Vincent Lostanlen, Carmine-Emanuele Cella, Rachel Bittner, and Slim Essid. 2019. Medley-solos-DB: a cross-collection dataset for musical instrument recognition.

Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2017. Dcase 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2018. A multi-device dataset for urban acoustic scene classification. *arXiv preprint arXiv:1807.09840*.

Irene Martin Morato and Annamaria Mesaros. 2021. Macs - multi-annotator captioned soundscapes.

Junwoo Park, Youngwoo Cho, Gyuhyeon Sim, Hojoon Lee, and Jaegul Choo. 2022. Enemy spotted: in-game gun sound dataset for gunshot classification and localization. In *2022 IEEE Conference on Games (CoG)*, pages 56–63. IEEE.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Yaman Kumar Singla, Jui Shah, Changyou Chen, and Rajiv Ratn Shah. 2022. What do audio transformers hear? probing their representations for language delivery & structure. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 910–925. IEEE.

Sonniss Limited. 2022. Sonniss Game Audio. Registered in England, UK. Company number: 09377364. Accessed: 25 September 2023.

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*.

Mi Tian, Ajay Srinivasamurthy, Mark Sandler, and Xavier Serra. 2014. A study of instrument-wise onset detection in beijing opera percussion ensembles. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2159–2163. IEEE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

George Tzanetakis, Georg Essl, and Perry Cook. 2001. Automatic musical genre classification of audio signals.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023a. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023b. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

A Additional Results

B Prompts employed for LLMs

Fig. 5 illustrates the prompt employed for synthesizing `CompA-R`. Fig. 6 illustrates the prompt employed for evaluating model responses on `CompA-R`. For dense captioning, we just prompt the model: *Write an audio caption describing the sound in detail*.

C `GAMA-IT` vs `GAMA` and Evaluation Choices.

`GAMA` is first fine-tuned on OpenAQA and then instruction-tuned on `CompA-R` for complex reasoning. We call the instruction-tuned version `GAMA-IT`. We do not evaluate `GAMA-IT` on general tasks like classification and vanilla captioning [2]. `GAMA-IT` is aligned to generate detailed descriptions as part of the complex reasoning stage, and we found a lack of metrics and methods that can faithfully evaluate such descriptions for classification or captioning. For example, the retrieval-based classification evaluation method, employed extensively in prior work, including ours, uses a Sentence-BERT to retrieve the label closest to the description for classification evaluation. During our preliminary analysis, we found that Sentence-BERT, which just performs retrieval using semantic matching, is unable to faithfully retrieve the correct label despite the caption mentioning the label as an audio event. We further investigated CLAP as our retrieval model for evaluation and found that it suffers from the same limitations. We attribute this to the detailed and dense nature of the descriptions and the fact that these models only focus on high-level semantic meaning for retrieval. Our initial experiments show that LLM prompting serves as a feasible alternative for automatic evaluation (beyond human evaluation) using such dense descriptions, but due to the lack of resources and a formal framework, we leave this as part of future research.

[2]**Note:** Both depend on the description of the input audio generated by the model

C.1 Soft Prompts

We employ the soft prompt only in the instruction tuning stage for learning complex reasoning and not in the fine-tuning step. We do this for 2 reasons: (i) Fine-tuned `GAMA` is only expected to solve generic audio tasks like classification, captioning, etc. Thus, we hypothesize that such high-level semantic cues are not necessary for effective and optimal performance. (ii) Since fine-tuning is done on a large-scale dataset and acoustic event classification is far from accurate, our soft prompt method might add unwanted noise to the training process, thereby leading to sub-optimal performance. On the contrary, our instruction-tuning stage, which is done on relatively low-resource data and is only responsible for aligning a model for complex reasoning, is robust to inaccurate audio tags due to our soft-prompting methodology.

D Additional Details: Human Study

Note. Our institution's Institutional Review Board (IRB) has granted approval for both human studies presented in the paper.

Background and Recruitment for Dense Captioning and `CompA-R`-*test* Evaluation. We recruit 3 professionals for human evaluation of dense captioning and `CompA-R`-*test* evaluation. All these 3 professionals come with at least a Ph.D. in Engineering or Sciences and were asked to use headphones to first analyze the audio and then judge the response quality. The authors of this paper gave these annotators 5 examples of responses and the corresponding judgments. The work was done voluntarily and not paid. We refrain from recruiting crowd raters as prior research has noticed discrepancies in evaluation by them (Gudibande et al., 2023). More precisely, they have been shown to possess a tendency to rate an answer with a high score only by visualizing the style of answering and not the exact factual information making up the response.

All 3 human annotators score the response between 1-5 and we report score averaged across the 3. Prior to evaluation all annotators were given at least 10 examples from the authors of the paper of generations and their corresponding scores. For evaluation, only the audio was provided to them with a software that could play the audio and has fields to input the scores.

Background and Recruitment for OpenAQA. Since the size of OpenAQA is relatively larger than `CompA-R`-*test*, we perform evaluation on Amazon Mechanical Turk similar to Gong et al. (2024). Evaluation was done with a total of 267 unique human evaluators and each generation was scored by 2 evaluators. The same software was used for evaluation as `CompA-R`-*test*.

E Additional Details: Audio Q-Former

E.1 Audio Q-Former Training Details

Pre-training Hyper-parameter. For Stage 1 of training, we employ a training batch size of 192, an initial learning rate of 1e-4, a minimum learning rate of 1e-5, and a warm-up learning rate of 1e-6. We do cosine decay as the learning rate scheduling technique. We do warmup for 5000 steps. Stage 1 was pre-trained on 8 A6000 GPUs for 100 epochs. For Stage 2 of training, we keep the exact same settings as Stage 1 but change the batch size to 128.

Fine-tuning. For zero-shot audio classification evaluation, we fine-tune the Audio Q-Former after Stage 1 pre-training on the same corpus presented in Table 3 and using the same Stage 1 objective. The only difference in the fine-tuning step is that we train the AST model, which is otherwise kept frozen in the pre-training stage.

Fine-tuning Hyper-parameter. For fine-tuning, we again use the same hyper-parameter setting as Stage 1 pre-training but use a batch size of 64.

E.2 Training Dataset Details

Table 3 provides dataset statistics of all individual datasets used for training Audio Q-Former. We employ ≈2.2M audio-caption pairs for training with no speech-transcription pairs.

Dataset	#Audio-Caption Pairs
Audio Set (Gemmeke et al., 2017) [3]	1591364
Free Sound (Fonseca et al., 2022) [4]	259020
VGGSound (Chen et al., 2020) [5]	185161
AudioSet Strong (CompA Version) (Ghosh et al., 2024b) [6]	108311
MACS (Morato and Mesaros, 2021) [7]	14400
BBC (BBC, 2018) [8]	31201
AudioCaps (Kim et al., 2019) [9]	48649
Clotho (Drossos et al., 2020) [10]	18735
SONISS (Sonniss Limited, 2022) [11]	1602
Musical Instrument (Agostinelli et al., 2023) [12]	7990
SoundBible (sou, 2023) [13]	1232
WavText5K (Deshmukh et al., 2022) [14]	4347
MusicCaps (Agostinelli et al., 2023) [15]	2645
GTZAN (Tzanetakis et al., 2001) [16]	6014
Medley-solos (Lostanlen et al., 2019) [17]	732

Table 3: List of open-source datasets used for collating our final dataset for training ReCLAP with ≈2.2M audio-caption pairs. All datasets are free to use for research purposes.

E.3 Augmentation Examples

Table 11 illustrates prompt augmentations for two categories from each dataset. Table 12 illustrates caption augmentations for training Audio Q-Former.

E.4 Hyper-parameter Tuning

E.4.1 Number of the custom prompts N

In this subsection, we show the effect of the number of custom prompts N on the final zero-shot audio classification performance. Table 4 compares performance across N={1,2,3,4,5}. As we see, the optimal performance is achieved at N=2, and model performance decreases with an increase in N. This decline is hypothesized to be due to the introduction of more noise into the process with each additional caption.

N	1	2	3	4	5
Score	38.1	**39.6**	39.0	39.4	36.5

Table 4: Impact of N on ZSAC with ReCLAP.

E.4.2 Probability of choosing rewritten captions

In this subsection, we show the effect of probability p on the final zero-shot audio classification performance.

p	0.2	0.4	0.6	0.8
Score	35.1	**39.6**	38.4	37.0

Table 5: Impact of p on ZSAC with ReCLAP.

[1] https://research.google.com/audioset/download.html
[2] https://huggingface.co/datasets/cvssp/WavCaps
[3] https://www.robots.ox.ac.uk/ vgg/data/vggsound/
[4] https://research.google.com/audioset/download.html
[5] https://zenodo.org/records/5114771
[6] https://sound-effects.bbcrewind.co.uk/
[7] https://research.google.com/audioset/download.html
[8] https://zenodo.org/records/4783391
[9] https://labs.freesound.org/datasets/
[10] https://www.kaggle.com/datasets/soumendraprasad/musical-instruments-sound-dataset
[11] https://soundbible.com/
[12] https://github.com/microsoft/WavText5K
[13] https://github.com/seungheondoh/music_caps_dl
[14] https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification
[15] https://zenodo.org/records/1344103
[16] https://zenodo.org/records/1344103

F Baseline Details

AudioCLIP. (Guzhov et al., 2022) AudioCLIP is an extension of the CLIP model that can handle audio in addition to text and images by incorporating the ESResNeXt audio model in the CLIP framework. It was trained on the AudioSet dataset, which contains millions of audio clips with corresponding labels.

CLAP. (Elizalde et al., 2023a) CLAP (Contrastive Language-Audio Pre-training), similar to CLIP, is an audio-language model trained with contrastive learning between audio data and their corresponding natural language descriptions. Representations are obtained from audio encoders and text encoders. Wu* et al. (2023b) further extend this using a feature fusion mechanism and keyword-to-caption augmentation into the model design to further enable the model to process audio inputs of variable lengths and enhance performance.

CompA-CLAP. (Elizalde et al., 2023a) CompA-CLAP, an extension to CLAP, is trained on completely open-sourced datasets and further fine-tuned using specific algorithms and datasets to improve compositional reasoning.

Pengi. (Deshmukh et al., 2023) Pengi was one of the first efforts to achieve general-purpose audio understanding through free-form language generation with transfer learning. Precisely, Pengi integrates an audio encoder with a decoder-only pre-trained language model (LM) where the audio features serve as a prefixes for the LM during response generation. Following this, similar to our evaluation strategy, they prompt the model to caption the input audio and calculate the similarity between the caption and the ground-truth audio label for zero-shot classification.

LTU. (Gong et al., 2024) As a concurrent work to Pengi, took a step forward and showed that substituting the pre-trained language model with an LLM can induce an LALM with reasoning capabilities. Precisely, they achieved this by integrating an audio encoder to LLaMA (Touvron et al., 2023) and fine-tuning the model on close-ended and open-ended instruction-tuning datasets. Finally, beyond just close-ended tasks, they also evaluate their models on open-ended reasoning tasks and show superior performance compared to baselines.

AudioGPT. (Huang et al., 2024) Different from Pengi and LTU, AudioGPT differs in how the audio models and LLMs are integrated for completing audio tasks. More specifically, different from end-to-end training and alignment, they integrate a closed-source model (ChatGPT) with a pre-trained audio model, already capable of completing the required task, using a modality-transfer transformer τ. The integration or interaction between the two models is accomplished using the prompts. Additionally, AudioGPT is capable of solving more tasks, which include human verbal speech, beyond just non-verbal speech like Pengi and LTU.

SALMONN. (Tang et al., 2024) SALMONN follows a similar architecture to LTU and Pengi and does prefix conditioning with an LLM. However, in addition to an audio encoder, they also integrate a speech encoder for speech or verbal audio understanding. Precisely, the audio and speech features are concatenated before feeding them as prefixes to the LLM. SALMONN shows unique reasoning capabilities over speech inputs overlayed with non-verbal audio.

Qwen-Audio. (Chu et al., 2023) Qwen follows a similar architecture to LTU, Pengi, and SALMONN, i.e., adding audio features as prefix to the model, and additionally employs a novel multi-task learning formulation for pre-training. More specifically, they append specific tags to specific parts of the instruction-response text pairs and train the model on diverse speech, non-speech, and music tasks. Post-pre-training, similar to `GAMA`, employs an instruction-tuning stage for alignment. The resultant model, Qwen-Audio-Chat, is able to respond to respond to diverse queries about the input speech and audio.

G Additional Details: `CompA-R`

G.1 Annotation and Annotator Details

As mentioned earlier, `CompA-R` was cleaned and `CompA-R`-*test* was verified by the paper authors themselves. To preserve anonymity, we briefly provide some details about the authors. All authors of the paper are either enrolled in or have graduated from a graduate degree (MS and/or Ph.D.). All authors have at least 2 years of professional research experience at a academic or industry lab. Their research experience spans across speech, audio and language processing. This provides them with adequate knowledge to faithfully complete the process.

For `CompA-R`-*test* verification, after at least 3 authors verified the test set, with proper rationales (which they were also asked to provide) the lead author cross-verified all instances. The verification was done manually on local laptops and no kind of application was used which was made specifically for this. More details will be provided on camera-ready.

H Additional Details: General

H.1 GAMA Training Dataset Details

Table 6 shows statistics of all datasets used for fine-tuning and instruction-tuning GAMA. Table 7 shows statistics of CompA-R, which is sourced entirely from the AudioSet-Strong dataset.

Dataset	# Audio Samples	# QA Pairs
AudioSet-Strong	102K	636K
AudioSet	500K	441K
VGGSound	184K	336K
FSD50K	41K	82K
AudioCaps	46K	90K
FreeSound	91K	91K
Clotho	5K	32K
Sound Bible	1.2K	12K
NSynth(Instrument+Source)	301K	602K
Clotho AQA	1.5K	4.2K
MusicCaps	5.5K	2.8K
MusicQA	13.1K	118k
Magna	51.7K	51.7K
Sum (Closed-Ended)	**1,217K**	**2,555K**
AudioSet-Strong (Open-Ended)	91K	901K
AudioSet-20K	19K	184K
VGGSound (Open-Ended)	184K	907K
FSD50K (Open-Ended)	41K	403K
AudioCaps (Open-Ended)	46K	478K
Freesound (Open-Ended)	91K	791K
Clotho (Open-Ended)	5K	89K
Sound Bible (Open-Ended)	1.2K	10K
Sum (Open-Ended)	**453K**	**3,764K**
Total	**1,670K**	**6,319K**

Table 6: The statistics of the OpenAQA dataset.

Dataset	# Audio Samples	# QA Pairs
AudioSet-Strong	62613	200234
Total	**62613**	**200234**

Table 7: The statistics of the CompA-R dataset.

H.2 GAMA Evaluation Dataset Details

Table 8 shows statistics of all datasets used for evaluating GAMA. Table 10 shows statistics of CompA-R-*test*, which is sourced entirely from the AudioSet-Strong dataset.

[1]https://www.kaggle.com/datasets/modaresimr/sound-event-detection-audioset-strong

[2]https://zenodo.org/records/4060432

[3]https://www.tensorflow.org/datasets/catalog/nsynth

[4]https://zenodo.org/records/6473207

Dataset	# Instances
AudioSet-Strong [18]	102K
AudioSet	500K
VGGSound	184K
FSD50K [19]	41K
AudioCaps	46K
FreeSound	91K
Clotho	5K
Sound Bible	1.2K
NSynth$_{instrument}$ [20]	4K
NSynth$_{source}$ [21]	4K
Clotho AQA [22]	1.3K
GTZAN	3K
Medley-solos-DB	12.2K

Table 8: The statistics of the datasets used for evaluation of GAMA.

Dataset	Evaluation Metric
Classification (zero-shot)	
VocalSound (VS) (Gong et al., 2022)	Acc.
TUT 2017 (TUT) (Mesaros et al., 2018)	Acc.
Beijing Opera (BJO) (Tian et al., 2014)	Acc.
GTZAN (GTZ) (Park et al., 2022)	Acc.
Medley-solos-DB (MDB) (Lostanlen et al., 2018)	Acc.
Classification (weak zero-shot)	
DCASE2017 Task 4 (DCASE) (Mesaros et al., 2017)	Mi-F1
ESC-50 (Piczak, 2015)	Acc.
Classification (seen)	
VGGSound (VGG) (Chen et al., 2020)	Acc.
FSD50K (FSD) (Fonseca et al., 2021)	mAP
AudioSet (AS) (Gemmeke et al., 2017)	mAP
NSynth (NS) (Engel et al., 2017)	Acc.
Captioning (vanilla & dense)	
AudioCaps (Kim et al., 2019)	SPICE & Human
Clotho (Drossos et al., 2020)	SPICE & Human
AQA (close-ended)	
Clotho AQA (Lipping et al., 2022)	Acc.
AQA (open-ended)	
OpenAQA (Gong et al., 2024)	Human
AQA (complex open-ended)	
CompA-R-*test (ours)*	GPT-4 & Human

Table 9: List of evaluation datasets and their corresponding evaluation metrics for GAMA.

H.3 Other Details

Model Parameters: GAMA has a total of ≈7B parameters. Out of this, LLaMA-2-7B has 32 transformer-encoder layers and ≈6.7B parameters, the Audio Q-Former has ≈280M parameters, and our LoRA modules introduce 4.2 M learnable parameters for fine-tuning. The AST used in our experiments (audio-encoder of CAV-MAE (Gong et al., 2023)) has ≈ 85M parameters with 12 transformer-encoder layers, 768-hidden-state, and 12 attention-heads.

Compute Infrastructure: All our experiments are conducted on four NVIDIA A6000 GPUs. Training GAMA required four days of continuous training. Training GAMA-IT requires 4 hours of training. Pre-training Audio Q-Former requires 7 days each for stages 1 and 2.

Dataset	# Audio Samples	# QA Pairs
`CompA-R`-*test*	500	1561
Total	**500**	**1561**

Table 10: The statistics of the `CompA-R`-*test* dataset.

Implementation Software and Packages: We implement all our models in PyTorch [23] and use the HuggingFace [24] implementations of T5$_{\text{large}}$ and the original implementation of HTSAT$_{\text{tiny}}$[25].

For our baselines, we use the original GitHub repository provided by the authors: LAION-CLAP [26], CompA-CLAP [27], CLAP [28], Wav2CLIP [29], AudioCLIP [30], MMT [31], ML-ACT [32], Pengi [33], LTU [34], AudioGPT [35], SALMONN [36], Qwen-Audio [37].

Potential Risks. `GAMA` might encode biases from the pre-trained LLM or during its fine-tuning stage. Additionally, Audio Q-Former used as a backbone for audio-to-text/music generation, might generate synthetic audio that is misused.

[23]https://pytorch.org/
[24]https://huggingface.co/
[25]https://github.com/RetroCirce/HTS-Audio-Transformer
[26]https://github.com/LAION-AI/CLAP/tree/main
[27]https://github.com/Sreyan88/CompA
[28]https://github.com/microsoft/CLAP
[29]https://github.com/descriptinc/lyrebird-wav2clip
[30]https://github.com/AndreyGuzhov/AudioCLIP
[31]https://github.com/akoepke/audio-retrieval-benchmark
[32]https://github.com/akoepke/audio-retrieval-benchmark
[33]https://github.com/microsoft/pengi
[34]https://github.com/YuanGongND/ltu
[35]https://github.com/aigc-audio/audiogpt
[36]https://github.com/bytedance/salmonn
[37]https://github.com/QwenLM/Qwen-Audio

```
# Prompt 1

I will provide you with 2 different types of information about a 10-second audio clip:

1. A list where each comma-separated element indicates the individual events occurring in the audio at various time segments. For example, '(Speech-
0.0-0.64)' would mean human speech between 0.0 second to 0.64 second.
2. A caption of the audio describing in a brief and abstract manner the scene in which the audio takes place.

I want you to act as a Prompt Generator. According to the event information and the caption, design some instructions and corresponding responses. The
instruction should be designed in a way such that it can be answered only from the audio without the caption and any other detail provided. The
instruction should involve one or more hops of complex knowledge and complex reasoning based on the scene created by the audio and the correspnding
caption. Ensure that the knowledge and reasoning chains in the instructions are precise and sufficiently challenging, to the extent that only well-
educated people and experts in the respective field can provide adequate responses.

The instructions must meet the following conditions:
1. Do NOT use phrases like 'according to the caption' in both the questions and answers; you should ask and answer as if you were observing the image
by yourself.
2. The questions and answers should be as diverse as possible.
3. Please don't ask some simple questions about the intensity of the audio or the gender speaking the utterance; your questions must involve some
knowledge.
4. Your instructions should not be answered directly based on the image and your instructions. Instead, it requires the test-taker to carefully observe
the image and have a deep knowledge of the content within the image in order to answer correctly.
5. If a question cannot be answered, please do not ask.

Come up with 3 diverse instructions for the knowledge topics above with different language styles and accurate answers. The instructions should contain
interrogative sentences and declarative sentences. The answers should be less than 30 words.

Output format, which is a list of jsons:

[{'Instruction': instruction example, 'Answer': answer example, 'Knowledge topic': The specific knowledge topic}, {'Instruction': instruction example,
'Answer': answer example, 'Knowledge topic': The specific knowledge topic}, …]
Here are some examples of inputs and outputs:

Input list of audio events: ['(Speech-0.0-0.64)', '(Mechanisms-0.0-10.0)', '(Dog-0.221-0.547)', '(Dog-0.803-0.966)', '(Generic impact sounds-0.885-
1.129)', '(Tick-0.99-1.083)', '(Dog-1.432-1.665)', '(Speech-1.537-4.901)', '(Dog-1.921-2.119)', '(Dog-2.456-3.202)', '(Dog-3.434-3.597)', '(Dog-4.016-
4.121)', '(Dog-4.936-5.39)', '(Generic impact sounds-5.204-5.611)', '(Dog-5.774-5.972)', '(Speech-5.984-6.787)', '(Tick-6.508-6.636)', '(Dog-6.717-
8.266)', '(Generic impact sounds-7.649-8.277)', '(Laughter-8.347-9.488)', '(Dog-9.767-10.0)']
Caption: A baby cries while a woman laughs, creating a joyful and lively atmosphere in a domestic setting.

Output list of jsons: [{'Instruction': 'Analyze the sounds in the audio and determine the most likely cause of the laughter heard towards the end of
the recording. Consider the potential interactions between the different sound sources and their temporal overlaps.' , 'Answer': 'The laughter likely
results from the playful interaction between the dogs and the baby, as indicated by the overlapping sounds of dogs and the baby's presence.',
'Knowledge topic': 'Human and Animal Behavior Interpretation'}, {'Instruction': 'From the given audio, infer the type of domestic setting depicted in
the scene. Base your inference on the variety and sequence of sounds, particularly focusing on the interaction between the human speaking, the dog
barking, and other background noises that may be there.', 'Answer': 'The setting is likely a home with an active family environment, evidenced by the
continuous presence of dogs, speech, and everyday household sounds.', 'Knowledge topic': 'Environmental Acoustics and Domestic Soundscapes'},
{'Instruction': 'Considering the duration and placement of speech and laughter in the audio, infer the possible emotional dynamics between the
speakers. How do these elements interact to shape the scene's atmosphere?', 'Answer': 'The scene likely transitions from a more chaotic or lively mood
and finally to a more joyful and relaxed atmosphere.'}]

Input list of audio events: ['(Insect-0.0-0.724)', '(Mechanisms-0.0-9.777)', '(Female speech, woman speaking-0.737-1.434)', '(Bird vocalization, bird
call, bird song-1.243-1.775)', '(Insect-2.376-3.182)', '(Female speech, woman speaking-3.386-3.509)', '(Insect-4.397-5.23)', '(Dog-7.906-8.78)',
'(Surface contact-8.603-9.654)']
Caption: 'Birds chirp in the distance as a dog barks, creating a lively atmosphere in a peaceful outdoor setting.'
Output list of jsons: [{'Instruction': 'What time of day this scene is likely set in?.' , 'Answer': 'The concurrent presence of insect and bird sounds
suggests a natural, outdoor environment, possibly during early morning or evening when such wildlife is typically active.', 'Knowledge topic':
'Environmental Sound Analysis and Wildlife Behavior'}, {'Instruction': 'Analyze the presence and timing of the dog's barking in the latter part of the
audio. Considering the preceding sounds and infer the dog's possible reaction or behavior in this context.', 'Answer': 'The dog's barking following the
peaceful nature sounds and speech could indicate a response to a new stimulus, possibly a visitor or an animal in the area.', 'Knowledge topic':
'Animal Behavior Analysis in Diverse Sound Environments'}, {'Instruction': 'Deduce the woman's likely activity or purpose in this setting.', 'Answer':
'The woman might be engaging in an outdoor activity like gardening or bird-watching.','Knowledge topic': 'Human activity recognition through scene
analysis' }]

Input list of audio events: ['(Music-0.0-10.0)', '(Male singing-0.0-10.0)', '(Male speech, man speaking-0.354-1.364)', '(Male speech, man speaking-
7.674-10.0)', '(Crowd-7.681-10.0)']
Caption: 'A basketball bounces while music plays, and a man speaks in an indoor stage environment.'
Output list of jsons:  [{'Instruction': 'Considering the presence of crowd sounds towards the end of the audio, deduce the nature of the event taking
place. How do the elements of music, singing, and speech suggest the type of event and audience involvement?', 'Answer': 'The event seems to be a live
performance or concert, with the crowd's reaction indicating an engaged and responsive audience, typical in such settings.', 'Knowledge topic': 'Event
Atmosphere Analysis'}, {'Instruction': 'Given the continuous presence of music and male singing throughout the audio, analyze the role of the man's
speech in shaping the atmosphere of the scene. How does his speech, interspersed with music and singing, contribute to the overall environment?',
'Answer': 'The man's speech likely serves as commentary or narration, adding a personal or interactive element to the musical performance, enhancing
the audience's engagement.', 'Knowledge topic': 'Music and Speech Dynamics'}, {'Instruction': 'Identify the genre of music being played and explain how
it complements the atmosphere of the indoor stage environment.', 'Answer': 'The genre is likely upbeat or energetic, enhancing the lively ambiance of a
sports or performance event in an indoor setting.','Knowledge topic': 'Music Genre Detection and Scene Analysis'}]

Input list of audio events: {timestamp events}
Caption: {caption}
Output list of jsons:
```

Figure 5: Prompts/Instructions used for caption augmentation with LLaMa-7B. Prompts are indexed according to the description in Section 3.2.

Complex AQA Evaluation Prompt

```
Please act as an impartial judge and evaluate the quality of the response provided with respect to the details provided. You will rate the quality of the response on multiple aspects, such as Helpfulness, Clarity, Factuality, Depth and Engagement. The response has been provided by an AI agent for a query related to an input audio, which the agent can perceive. You will be provided with 4 kinds of information for evaluating the response:
1. A list where each comma-separated element indicates the individual events occurring in the audio at various time segments. For example, '(Speech-0.0-0.64)' would mean human speech between 0.0 second to 0.64 second.
2. A caption of the audio describing in a brief and abstract manner the scene in which the audio takes place.
3. The question asked to the AI agent related to the audio.
4.An answer provided by an expert judge which you can consider as a reference.
5. An answer by the AI agent.

##Query: {query}


## Evaluate
### Aspects
- Helpfulness: Rate the response based on how well it addresses the users query about the audio and provides a relevant answer. A score of 5 indicates the answer fully aids the user, while a 1 suggests it offers little to no help.
- Clarity: Rate the response based on how well-structured it is, with ideas presented in a clear and coherent manner. A high score of 5 means the answer is clear and logically structured, while a 1 suggests a disjointed or confusing reply.
 - Correctness: Evaluate the correctness or accuracy of the response provided with respect to the information provided to you. A perfect 5 indicates the response is entirely correct and accurate, while a 1 suggests it has significant errors or has not provided an answer to the question asked at all.
 - Depth: Determine the level of detail and thoroughness in the response. A score of 5 means the answer delves deeply into the aspects of the input image for answering the question, while a 1 indicates it barely scratches the surface.

### Format
### Given the query and the extra information about the audio provided (the caption and comma-separated list of individual individual events), please rate the quality of the output by scoring it from 1 to 5, individually on **each aspect**. You are allowed to use all 3 information provided to you about the audio, in any way you want, to judge the response.

Now, please output your scores in the following json format by filling in the placeholders in [].
  { 'helpfulness': { 'reason': '[your rationale]', 'score': '[score from 1 to 5]' }, 'clarity': { 'reason': '[your rationale]', 'score': '[score from 1 to 5]' }, 'correctness': { 'reason': '[your rationale]', 'score': '[score from 1 to 5]' }, 'depth': { 'reason': '[your rationale]', 'score': '[score from 1 to 5]' }, 'engagement': { 'reason': '[your rationale]', 'score': '[score from 1 to 5]' } }

Only return the json and nothing else."
```

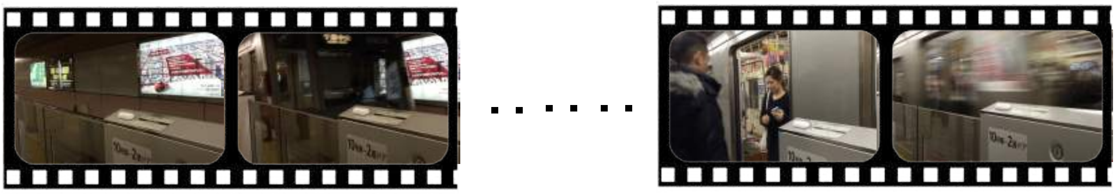
Figure 6: Prompts/Instructions used for evaluating the answers of models on CompA-R benchmark with GPT-4 as a judge.

Caption Evaluation Prompt

```
I want you to act as a Caption Evaluator. I will provide you with an audio caption generated by an AI agent. The agent was asked to generate a dense and detailed caption of the audio. To evaluate the caption, I will provide you with 2 different types of information about the 10-second audio clip:

1. A list where each comma-separated element indicates the individual events occurring in the audio at various time segments. For example, '(Speech-0.0-0.64)' would mean human speech between 0.0 second to 0.64 second.
2. A scene caption of the audio describing in a brief and abstract manner the scene in which the audio takes place. Using these two pieces of information, assign a score of 1-10 to the caption, where 1 is the lowest score and 10 is the highest score. Your evaluation should be based on the detailedness, correctness, and bluntness of the caption. Return a JSON with a single key 'score', where the value of the key is the score. Here are the details: {query}
```

Figure 7: Prompts/Instructions used for evaluating the captions generated by models Clotho and AudioCaps with GPT-4 as a judge.



Question: Based on the combination of subway sounds and male speech, suggest the actions the man might be performing at this time.

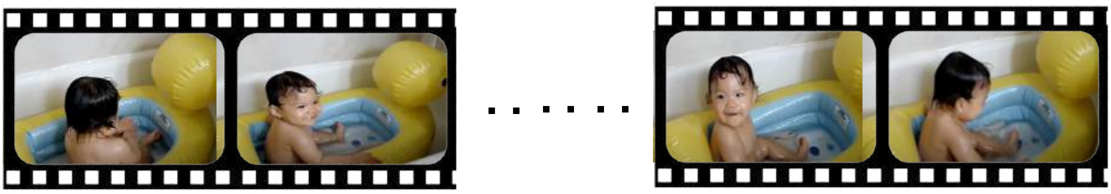
LTU: The man may be giving a lecture or presentation in the subway car while people are moving around him, possibly taking notes. **HE : 3**

SALMONN: It is difficult to determine the exact actions the man might be performing based solely on the combination of subway sounds and male speech. However, it is possible that the man is giving a public announcement or warning to the passengers on the subway. **HE : 1**

Pengi: It is a man. **HE : 1**

GAMA-IT: The man is likely giving instructions to passengers or announcing a station stop. He may also be explaining the train's route or destination. **HE : 5**

Figure 8: Prompts/Instructions used for evaluating the captions generated by models Clotho and AudioCaps with GPT-4 as a judge.



Question: Given the variety of sounds including mechanisms, generic impact sounds, and water splashes, determine the most probable activity occurring during the audio clip. Consider the presence of baby laughter in your analysis.

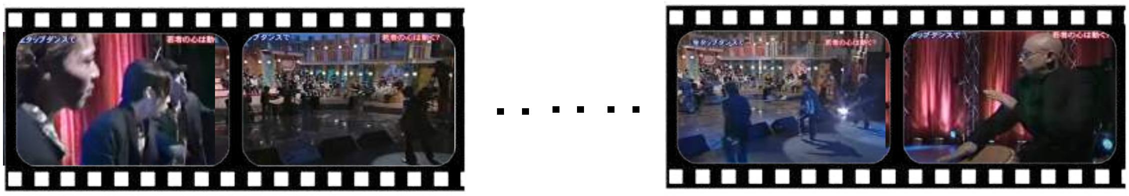
LTU: The most probable activity occurring during the audio clip is a family or group of people having fun together, possibly playing with babies or toddlers. The mechanical sounds could be from a playground or other outdoor equipment. **HE : 2**

SALMONN: Based on the variety of sounds including mechanisms, generic impact sounds, and water splashes, as well as the presence of baby laughter, the most probable activity occurring during the audio clip is likely a bath or a water-related activity. **HE : 5**

Pengi: It is a combo. **HE : 1**

GAMA-IT: The woman is likely bathing her child, as indicated by the sound of a bathtub filling up with water. **HE : 5**

Figure 9: Prompts/Instructions used for evaluating the captions generated by models Clotho and AudioCaps with GPT-4 as a judge.



Question: Inferring from the nature and arrangement of sounds, what can be said about the skill level of the tap dancer?

LTU

The skill level of the tap dancer cannot be inferred from the audio clip.

HE : 1

SALMONN

The skill level of the tap dancer can be inferred from the nature and arrangement of sounds. If the sounds are complex and intricate, it suggests that the dancer is highly skilled. On the other hand, if the sounds are simple and repetitive, it suggests that the dancer may not be as skilled.

HE : 3

Pengi

It is a big crowd.

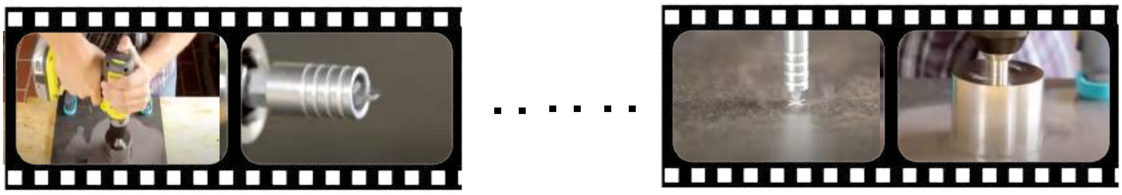
HE : 1

GAMA-IT

The tap dancing is likely performed by a professional or an expert. The rhythm and tempo are consistent throughout the entire duration of the clip.

HE : 5

Figure 10: Prompts/Instructions used for evaluating the captions generated by models Clotho and AudioCaps with GPT-4 as a judge.



Question: Determine the type of power tool being used in this woodworking process based on the available sounds. Consider the mechanisms sounds.

LTU

Based on the available sounds, it is difficult to determine the exact type of power tool being used. However, the sharp and metallic sound suggests that a circular saw or a hand drill might be present in the scene.

HE : 1

SALMONN

Based on the available sounds, it is likely that a circular saw is being used in this woodworking process. The circular saw is a power tool that uses a spinning circular blade to cut wood, and the sound of the blade cutting through the wood can be heard in the audio. Additionally, the sound of the motor and other mechanisms of the circular saw can also be heard in the audio.

HE : 2

Pengi

It is a cameraman.

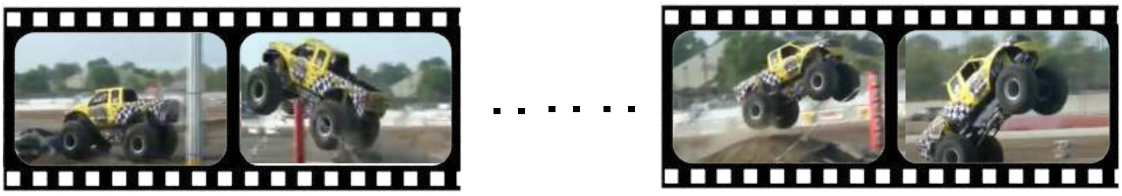
HE : 1

GAMA-IT

The sound is a drill, which is usually used for woodwork.

HE : 5

Figure 11: Prompts/Instructions used for evaluating the captions generated by models Clotho and AudioCaps with GPT-4 as a judge.



Question: Analyze the audio and infer the likely engine type based on the sound of acceleration and revving.

LTU

Based on the sound of acceleration and revving, it is likely that the engine type is a high-performance or sports car engine.

HE : 2

SALMONN

The audio contains the sound of an engine revving and accelerating. Based on this, it is likely that the engine type is a high-performance gasoline engine, such as a V8 or a turbocharged four-cylinder. The sound of the engine revving and accelerating is a common characteristic of high-performance engines, which are designed to produce a lot of power and torque. The sound of the engine can vary depending on the specific make and model of the vehicle, but the general characteristics of a high-performance engine can be identified by the sound of the engine revving and accelerating.

HE : 2

Pengi

It is diesel.

HE : 4

GAMA-IT

The truck is probably a diesel, as indicated by the heavy engine sounds and the absence of a typical gasoline engine's characteristic high-pitched whine at idle.

HE : 5

Figure 12: Prompts/Instructions used for evaluating the captions generated by models Clotho and AudioCaps with GPT-4 as a judge.

Dataset	Category	Acoustic Property	Captions
AudioSet	Baby cry, infant cry	1: "*a high-pitched, piercing wail*" 2: "*a persistent, rhythmic sobbing*"	1. A tiny voice pierced the stillness of the night, demanding attention with its *high-pitched piercing wail*. 1. In the midst of the bustling market, *a high-pitched wail* rose above the chatter, signaling a tiny dissenter among us. 2. The *persistent, rhythmic sobbing* of a newborn punctuates the night's silence. 2. Amidst the rustling leaves, a baby's *rhythmic sobs* weave an unexpected lullaby.
	Stomach rumble	1: "*a low, gurgling growl*" 2: "*a grumbling undercurrent*"	1. As she leaned in for the kiss, *a low, gurgling growl* betrayed her pre-date fasting. 1. In the midst of the exam, *a low, gurgling growl* echoed from her stomach, punctuating the tension. 2. As the classroom fell into a hushed anticipation of the next lecture, *a grumbling undercurrent* betrayed the student's skipped breakfast. 2. Amidst the solemnity of the library, *a grumbling undercurrent* served as a comical reminder of lunchtime's approach.
ESCC50	siren	1: "*Continuous, shrill alarm*" 2: "*Ear-splitting, cyclic blare*"	1. The night was pierced by the *continuous, shrill alarm* of the siren, signaling an urgency that couldn't be ignored. 1. The relentless, *shrill alarm* of the siren wove through the corridors of the hospital, prompting swift movements from the staff. 2. The *ear-splitting, cyclic blare* cut through the night as the ambulance raced down the street. 2. The *ear-splitting, cyclic blare* of the air raid siren cast a shadow of dread over the city.
	sheep	1: "*bleating*", 2: "baaing"	1. In the tranquil meadow, a chorus of woolly creatures serenaded the dawn with their gentle *bleating*. 1. The shepherd smiled as the flock's *bleating* echoed through the valley, signaling a return to the fold. 2. A chorus of baaing accompanied the farmer as he made his morning rounds in the misty fields. 2. Under the starry sky, the gentle baaing of the flock blended with the whispers of the night.
NSynth	flute	1: "*Melodious and silvery, carrying a light, airy tune that seems to float on the breeze.*", 2: "*Clear and resonant, with a pure tone that sings above the orchestra like a bird in the morning sky*."	1. The flute's melody weaved through the garden, mingling with the rustle of leaves. 1. Amidst the hustle of the market, the silvery flute tune danced above the clamor, a ribbon of tranquility. 2. The flute's crystalline melody soared above the orchestra, a lark greeting the dawn. 2. Amidst the rustle of the forest, the flute's song danced through the leaves, pure and high.
	bass	1: "*Thumping, providing a rhythmic pulse that can drive the beat of music.*", 2: "*Booming, with a powerful, pervasive quality that can reverberate through a space.*"	1. The bass pulsated through the dance floor, a heartbeat synchronizing every dancer's move. 1. Amidst the serene silence of the night, the bass from the distant festival throbbed like a gentle earthquake. 2. The bass pulsated like a second heartbeat, filling the room with its unyielding presence. 2. As the bassline dropped, it seemed to command the very air, a force unseen yet unforgotten.
FSD50K	Slam	1: "an abrupt, resonant boom that startles anyone nearby" 2: "a sharp, impactful smack as two hard surfaces collide with force"	1. The mailbox lid clapped shut, a resonant signal marking the departure of the day's correspondence. 1. The oven door's heavy thud resonated in the kitchen, a prelude to the aroma of freshly baked bread. 2. The kitchen was filled with the aroma of spices and the sharp smack of dough being forcefully thrown onto the countertop. 2. In the crisp morning air, the sharp smack of the newspaper hitting the doorstep announced the arrival of daily news.
	Dishes, pots, and pans	1: "Clanging and clattering" 2: "Metallic clinking and clunking"	1. A symphony of clanging and clattering announces the busy bustle of a restaurant kitchen in full swing. 1. The rhythmic clanging and clattering of pots and pans punctuate the air as grandma orchestrates her holiday feast. 2. The metallic clinking and clunking heralded the start of the dinner rush in the bustling restaurant kitchen. 2. A symphony of metallic clinking and clunking rose from the sink as grandma washed up after the family feast.
TUT Urban	bus	1: "*a deep, rumbling engine*", "2": "*the low, steady hum of the diesel motor*"	1.The city pulse beats with *a deep, rumbling engine*, heralding the arrival of the morning commute. 1. A gentle giant purrs in the stillness of dawn, its *deep, rumbling engine* announcing the start of a journey. 2. Market stalls buzz with life, their vibrant colors and smells underscored by the bus's *diesel hum* rolling down the avenue. 2. Leaves rustle in the autumn breeze, a natural chorus to the bus's *diesel motor humming* along the cobblestone path.
	residential area	1: "The symphony of children's laughter and chatter fills the air, punctuated by the occasional bark of a dog and the hum of lawn mowers in the distance." 2: "A serene hush blankets the neighborhood, broken occasionally by the soft whoosh of passing cars and the rustle of leaves stirred by a gentle breeze.",	1. The neighborhood comes alive with the melody of playful banter and the sporadic chorus of canines. 1. Amidst the gentle drone of distant lawn mowers, the air vibrates with juvenile mirth and convivial exchanges. 2. The neighborhood rests under a tranquil silence, punctuated now and then by the whisper of tires on asphalt and the soft dance of leaves in the wind. 2. Calmness envelops the streets, save for the faint hum of vehicles gliding by and the tender shuffling of foliage in the zephyr's caress.
UrbanSound 8K	air conditioner	1: "a steady humming" 2: "a low, monotonous droning"	1. The room filled with the steady humming of the air conditioner as they focused intently on their chess match. 1. A steady humming enveloped the library, where pages turned almost in rhythm with the air conditioning's constant song. 2. The air conditioner's low, monotonous droning became the unlikely lullaby for a midsummer's nap. 2. Amid the quiet study hall, the air conditioner's low, monotonous droning was a steady companion to the students' focused brows.
	gun shot	1: "A loud, sharp crack that echoes through the air. 2: "A thunderous boom that startles and reverberates."	1. The night's silence shattered with a loud, sharp crack echoing through the air. 1. A burst of sudden, sharp noise split the tranquil afternoon, reverberating off the canyon walls. 2. A thunderous boom startles a flock of birds into the sky, their wings flapping frantically against the silence that had just been. 2. The night's silence was shattered by a boom, reverberating through the alleyways and causing stray cats to scurry.
VGG Sound	mouse squeaking	1: "a high-pitched, sharp chirp" 2: "a soft, repetitive squeal"	1. In the moonlit barn, a tiny silhouette pauses to release its high-pitched, sharp chirp, disturbing the stillness of the hay-strewn loft. 2. Amidst the rustling leaves, a diminutive creature contributes its sharp chirp to the dusk chorus, a minuscule soloist in nature's vast orchestra. 3. A soft, repetitive squeal punctuated the silence of the old attic. 4. The cheese plate on the kitchen counter became the stage for a soft, repetitive squeal.
	typing on typewriter	1: "a rhythmic series of sharp clicks" 2: "a steady clatter of keys striking paper"	1. Fingers dance across keys, a rhythmic series of sharp clicks punctuating the silence of the library. 1. In the attic, a story unfolds to the staccato beat of a rhythmic series of sharp clicks. 2. Each steady clatter of keys striking paper weaves a tapestry of words, painting stories on the blank canvas. 2. In the dimly lit corner of the library, the rhythmic dance of metallic hammers against the page composes a silent symphony.

Table 11: Examples of prompt augmentations.

Original Caption	Augmented caption
A man speaks followed by the sound of shuffling cards in a small room.	1. A deep, resonant voice fills the small room, accompanied by the soft shuffle of cards as they change hands, creating an intimate and deliberate atmosphere. 2. The sound of a man's voice echoes through the small space, punctuated by the subtle rustle of cards as they are shuffled and arranged, invoking a sense of purposeful deliberation. 3. A deep voice speaks, followed by the subtle shuffle of cards, creating an intimate and anticipatory atmosphere in the small room. 4. The gentle rustle of cards breaks the silence, punctuated by a man's voice, evoking a sense of anticipation and private reflection in the cozy space.
A person strums an acoustic guitar, creating melodic music with the sound of a bell ringing in the background.	1. Soothing melodies flow from the acoustic guitar, harmonizing with the soft chime of a distant bell, crafting a peaceful ambiance. 2. The acoustic guitar's strings vibrate with grace, weaving a melodic tapestry that intertwines with the gentle ring of a bell, transporting the listener to a serene realm. 3. The gentle strumming of an acoustic guitar weaves a melodic tapestry, intertwined with the soft chime of a background bell, creating a soothing and harmonious atmosphere. 4. The rhythmic plucking of an acoustic guitar crafts a lively and uplifting melody, complemented by the delicate ringing of a background bell, transporting the listener to a serene and joyful realm.
Dogs bark while people talk in the background, creating a lively atmosphere in a field.	1. Lively chatter and joyful barks fill the air, capturing the playful spirit of a sunny day in a field. 2. The rhythmic sounds of dogs barking and people talking blend together, creating a vibrant and lively ambiance in the open field. 3. The chatter of people and the joyful barks of dogs fill the air, creating a vibrant and lively atmosphere in the field. 4. The sound of playful dogs and lively conversation fills the field, evoking a sense of happiness and energy.
A man's voice is heard speaking over a radio as a vehicle passes by in the background.	1. A clear, crisp voice pierces the airwaves, intertwining with the distant hum of a vehicle, creating an engaging audio experience. 2. The man's voice on the radio blends seamlessly with the subtle rumble of a passing vehicle, forming a captivating auditory tapestry. 3. A voiceover speaks over a radio, complemented by the distant hum of a vehicle passing by, creating a dynamic and engaging audio experience. 4. A man's voice broadcasts over the radio, intertwining with the subtle rumble of a vehicle in the background, forming a captivating audio landscape.
A woman speaks while a bird chirps in the background, creating a tranquil atmosphere in a natural setting.	1. A gentle voice echoes through the forest, harmonizing with the chirping of birds, creating a soothing ambiance. 2. The sound of a gentle voice blends seamlessly with the melodic chirping of birds, transporting the listener to a serene natural setting. 3. The woman's gentle voice blends with the soothing chirps of a bird, creating a serene ambiance reminiscent of a peaceful afternoon in nature. 4. The woman's words are accompanied by the melodic chirping of a bird, transporting the listener to a calming and picturesque outdoor setting.
Water rushes as people talk in the background near a hot spring, creating a serene ambiance.	1. Soothing waters create a peaceful ambiance, punctuated by the gentle chatter of people nearby, as if they are harmonizing with the soothing sounds of the hot spring. 2. The calm trickle of water creates an intimate atmosphere, with the soft murmur of voices in the background adding a sense of connection and tranquility to the space. 3. A soothing, babbling sound fills the air as people converse near a steaming hot spring, creating a tranquil atmosphere. 4. The gentle gurgling of water intertwines with the chatter of people in the background, crafting a peaceful and relaxing ambiance.
Soft music plays in the background as a speech is heard faintly, creating a calm and peaceful atmosphere.	1. A soothing melody floats in the background, complementing the faint speech, creating a tranquil ambiance. 2. The soft strains of music blend with the subtle speech, fostering a sense of serenity and calmness in the atmosphere. 3. Soothing tunes fill the air, complemented by a gentle speech, creating an atmosphere of tranquility and serenity. 4. Mellow music and soft speech blend together, crafting a calming environment that soothes the senses.'
A car engine revs up and then slows down, creating a vroom sound, as the vehicle accelerates in the audio.	1. The car's engine purrs and then decelerates, emitting a smooth and powerful vroom sound as it shifts gears, creating a dynamic and energizing atmosphere. 2. The vehicle's engine roars to life, producing a bold and intense vroom sound as it speeds up, then gradually slows down, immersing the listener in a thrilling and exhilarating experience. 3. The car's engine purrs powerfully, then decelerates, creating a smooth and steady vroom sound as the vehicle gains speed. 4. The car's engine roars to life, building momentum with a series of sharp vroom sounds before shifting gears and slowing down.
Background music plays softly as the theme music gradually fades in, creating a melodic ambiance in an arena/performance setting.	1. The arena comes alive with a subtle, soothing melody that gradually builds in intensity, creating an electrifying ambiance. 2. The soft strains of background music fill the air, setting the tone for an exhilarating performance in a vibrant arena setting. 3. Soft, melodic strains fill the air as the theme music subtly builds, establishing a harmonious ambiance in the arena. 4. The arena comes alive with a gentle, orchestral tune that gradually gains momentum, creating an uplifting and energetic atmosphere.

Table 12: Examples of caption augmentations.

Instruction-Response Pairs	AudioSet ID	Caption	Timestamp Events
Instruction: Analyze the audio to understand the potential emotional state or mood of the man. How does the progression from typing to speech to chewing reflect his transition through different phases of work or activity? **Output:**The man initially seems engaged and focused during the typing and speaking portion, which might then transition into relaxation during the break, suggested by the chewing sound.	YCecEf0abd4Y	A man speaks while typing on a keyboard in a small room, followed by the sound of chewing.	'(Generic impact sounds-0.0-1.037)', '(Background noise-0.0-10.0)', '(Generic impact sounds-1.191-1.421)', '(Generic impact sounds-2.01-2.202)', '(Generic impact sounds-2.343-2.574)', '(Male speech, man speaking-2.727-3.393)', '(Generic impact sounds-3.163-3.406)', '(Generic impact sounds-3.585-3.905)', '(Generic impact sounds-4.136-4.379)', '(Breathing-4.405-4.917)', '(Generic impact sounds-4.93-5.288)', '(Generic impact sounds-5.442-5.608)', '(Generic impact sounds-5.736-6.12)', '(Generic impact sounds-6.274-6.569)', '(Breathing-6.825-7.26)', '(Generic impact sounds-6.863-7.042)', '(Male speech, man speaking-7.81-8.873)', '(Generic impact sounds-8.041-8.348)', '(Breathing-9.001-9.36)', '(Human sounds-9.014-9.181)', '(Generic impact sounds-9.309-9.565)', '(Scrape-9.449-10.0)'
Instruction: Considering the diverse array of sounds present in the audio, from insects to birds to a dog, infer the type of ecosystem this outdoor setting might represent. What does the combination of these sounds tell us about the biodiversity and potential human impact in this area? **Output:**The ecosystem is likely a suburban or rural area with a mix of wildlife and human habitation, indicated by the variety of animal sounds and intermittent female speech.	YcQiEI7HLGJg	Birds chirp in the distance as a dog barks, creating a lively atmosphere in a peaceful outdoor setting.	'(Insect-0.0-0.724)', '(Mechanisms-0.0-9.777)', '(Female speech, woman speaking-0.737-1.434)', '(Bird vocalization, bird call, bird song-1.243-1.775)', '(Insect-2.376-3.182)', '(Female speech, woman speaking-3.386-3.509)', '(Insect-4.397-5.23)', '(Dog-7.906-8.78)', '(Surface contact-8.603-9.654)'
Instruction: Given the presence of mechanisms throughout the audio and the interspersed generic impact sounds, infer the type of machinery that is likely operating in the background and its commonality in an office environment. **Output:**The continuous mechanism sound suggests a printer or copier, which are common in office settings.	YXQ2XAXx7mKs	A printer hums while people converse in the background, creating a typical office ambiance.	'(Generic impact sounds-0.0-0.622)', '(Mechanisms-0.0-10.0)', '(Generic impact sounds-0.815-1.227)', '(Generic impact sounds-1.632-2.134)', '(Child speech, kid speaking-3.591-6.684)', '(Squeal-7.385-7.612)', '(Child speech, kid speaking-8.437-10.0)'

Table 13: Examples of `CompA-R`

Instruction-Response Pairs	AudioSet ID	Caption	Timestamp Events
Instruction: From the sequencing and overlapping of different sound events, infer the likely cause and process of the dripping sounds heard intermittently throughout the audio. **Output:**The dripping sounds may be a result of water overflowing from a filled sink or bath, supported by the earlier sounds of splashing and liquid gurgling.	YCU9A5xL3TVc	Water splashes and gurgles as it drips inside a small room, creating a soothing ambiance reminiscent of a tranquil bathroom.	'(Background noise-0.0-10.0)', '(Generic impact sounds-0.083-0.331)', '(Splash, splatter-0.67-1.174)', '(Liquid-1.385-1.956)', '(Splash, splatter-2.325-3.138)', '(Liquid-3.085-4.131)', '(Liquid-4.372-5.5)', '(Drip-4.949-5.047)', '(Drip-5.279-5.458)', '(Generic impact sounds-8.819-9.142)', '(Drip-9.511-9.649)'
Instruction: Assess the style and elements of the rapping and music in this audio clip. Based on this, what genre of music might the DJ be playing on stage? **Output:** Given the presence of rapping and electronic music, the DJ is likely playing Hip Hop or Electronic Dance Music (EDM).	YRjUZjMPP-nA	Electronic music plays as a whoosh sound follows, creating a lively atmosphere for the DJ performing on stage.	'(Rapping-0.0-0.376)', '(Music-0.0-10.0)', '(Rapping-0.685-1.663)', '(Rapping-2.295-2.837)', '(Sound effect-2.423-5.222)', '(Sound effect-7.427-10.0)'

Table 14: Examples of `CompA-R`

Instruction-Response Pairs	AudioSet ID	Caption	Timestamp Events
Instruction: Identify the role of the crowd's continuous conversation during the woman's speech. How does it contribute to the atmosphere of the scene? **Output:** The crowd's constant chatter indicates an informal or relaxed environment, suggesting that while the woman's speech is central, other conversations are also ongoing in the background.	Y6fRYeClf5U4	A woman delivers a speech while a crowd of people engage in conversation in an urban setting.	'(Crowd-0.0-10.0)', '(Wind-0.008-10.0)', '(Female speech, woman speaking-0.074-1.65)', '(Female speech, woman speaking-2.879-5.427)', '(Female speech, woman speaking-5.604-6.083)', '(Female speech, woman speaking-6.9-10.0)'
Instruction: Considering the presence of bird sounds and chainsaw noise, identify the probable location where this sound sequence is occurring. **Output:**The location is likely an outdoor area, possibly in a woodland or suburban setting where birds can be heard alongside human activity and power tools.	YbkG4M4TiXZg	A man speaks while a power tool revs up, indicating a speech event followed by the sound of a chainsaw.	'(Male speech, man speaking-0.0-0.268)', '(Chainsaw-0.0-10.0)', '(Male speech, man speaking-1.772-4.425)', '(Male speech, man speaking-5.008-8.118)', '(Bird vocalization, bird call, bird song-5.362-7.512)', '(Bird vocalization, bird call, bird song-8.244-8.709)', '(Bird vocalization, bird call, bird song-8.937-9.283)', '(Male speech, man speaking-9.661-10.0)'

Table 15: Examples of `CompA-R`-**test**

Instruction-Response Pairs	AudioSet ID	Caption	Timestamp Events
Instruction: Analyze the frequency and consistency of the ticking sound. What is the likely source of this sound, and what implication might it have on the depicted setting? **Output:**The steady ticking likely comes from a clock, which coupled with the soft music, suggests a relaxed, cozy domestic setting, perhaps aimed at unwinding or relaxation.	YCoBAR5Mbjys	The clock ticks steadily as soft music plays in the background, creating a calming atmosphere in a cozy living room.	'(Mechanisms-0.0-10.0)', '(Alarm clock-0.008-10.0)', '(Tick-0.386-0.583)', '(Tick-1.071-1.22)', '(Tick-1.764-1.906)', '(Tick-2.465-2.638)', '(Tick-3.197-3.331)', '(Tick-3.772-3.976)', '(Tick-4.346-4.48)', '(Tick-4.646-4.787)', '(Tick-5.087-5.22)', '(Tick-5.669-5.795)', '(Tick-6.031-6.15)', '(Tick-6.37-6.528)', '(Tick-6.724-6.795)', '(Tick-6.969-7.118)', '(Tick-7.386-7.614)', '(Tick-8.134-8.354)', '(Tick-8.882-9.094)', '(Tick-9.315-9.425)', '(Tick-9.575-9.685)'
Instruction: Identify the type of vocal music that is being depicted in the audio based on the presence of singing and beatboxing. **Output:**This audio resembles A Capella, where voices impersonate the sounds of instruments, including rhythms often mimicked through beatboxing.	Y6SvDRiIG2NY	A group of people sing and harmonize, creating vocal music with occasional beatboxing, in a room with a piano.	'(Male singing-0.0-6.594)', '(Music-0.0-10.0)', '(Mechanisms-0.0-10.0)', '(Breathing-7.064-8.314)', '(Breathing-8.911-10.0)', '(Male singing-9.713-10.0)'
Instruction: Based on the audio, ascertain the possible relationship between the gunfire sounds, artillery fire, and music. How does the sequencing and manner of these sounds contribute to the atmosphere of the scene? **Output:** The gunfire and artillery sounds likely serve as a ceremonial display, with the music adding to the grandeur and solemnity of a military parade.	YbJvOp4gmHBg	Gunshots and artillery fire echo through the air as music plays during a military parade at a raceway.	'(Music-0.0-10.0)', '(Generic impact sounds-0.166-0.307)', '(Artillery fire-0.32-0.704)', '(Generic impact sounds-0.781-0.948)', '(Generic impact sounds-1.063-1.165)', '(Generic impact sounds-1.524-1.677)', '(Generic impact sounds-2.625-2.881)', '(Artillery fire-3.035-3.521)', '(Generic impact sounds-3.611-3.777)', '(Generic impact sounds-4.213-4.43)', '(Generic impact sounds-5.096-5.262)', '(Artillery fire-5.288-5.762)', '(Generic impact sounds-5.89-6.095)', '(Generic impact sounds-6.479-6.812)', '(Generic impact sounds-6.94-7.106)', '(Artillery fire-7.222-7.606)', '(Generic impact sounds-8.207-8.425)', '(Artillery fire-8.476-8.988)', '(Generic impact sounds-9.206-9.385)', '(Generic impact sounds-9.654-9.795)'

Table 16: Examples of `CompA-R`-**test**