PARAMETER-EFFICIENT TUNING OF LARGE AUDIO-LANGUAGE MODELS FOR DCASE 2025 CHALLENGE TASK5

Technical Report

Pengfei Cai[†]*, Yanfeng Shi*[†]*, Qing Gu*[†]*, Nan Jiang*[†]*, Yan Song*

National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China, China

ABSTRACT

In this technical report, we describe our systems developed for DCASE 2025 Challenge Task5. Our system is mainly based on parameter-efficient tuning of large audio-language models, *e.g.,* `Qwen2-Audio` and `Kimi-Audio`. The training process is conducted using Low-Rank Adaptation (LoRA) and divided into two stages: Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL). In addition, we reformatted the annotations of the AudioSet-Strong and MMAU datasets into a question-answer format to augment the official task dataset. Our final system achieves an accuracy of 80.0% on the development set.

Index Terms— Audio Question Answering, Large Audio-Language Models, Low-Rank Adaptation, Reinforcement Learning.

1. INTRODUCTION

Recent advancements in large audio-language models [1, 2, 3, 4] have significantly advanced the field of audio understanding and analysis. These models are typically pre-trained on large-scale audio-text paired corpora covering diverse tasks such as automatic speech recognition, speech emotion recognition, and audio event classification. Through multi-task pre-training, they acquire general-purpose audio-language representations that can be efficiently adapted to downstream applications.

To specifically evaluate the audio understanding and reasoning capabilities of these models, the DCASE 2025 Challenge Task 5 introduces the Audio Question Answering (AQA) benchmark. This task requires models to listen to an audio clip, interpret its content, and answer a natural language question grounded in the acoustic information. Unlike conventional classification tasks, AQA demands not only perceptual understanding of sound events but also contextual reasoning and temporal inference across diverse domains. The challenge is divided into three sub-tracks, each designed to test different facets of audio comprehension:

- **Bio-acoustics QA** – This subset focuses on recognizing and understanding marine mammal vocalizations. Models are tasked with identifying species and specific vocalization types from audio clips sourced from the Watkins Marine Mammal Sound Database. Success in BQA requires both knowledge of biological sound-emitting behaviors and auditory perception.
- **Temporal Soundscapes QA** – Here, the challenge centers on temporal reasoning. Given environmental recordings containing multiple overlapping or sequential events, models must answer questions about event order, onset/offset times, or duration. This sub-task tests a model's ability to detect temporal boundaries and relate sound events in time.
- **Complex QA** – This subset focuses on complex question answering grounded in audio understanding. Models must perform high-level reasoning—*e.g.*, identifying overlapping sound events, interpreting sequences of auditory phenomena, or discerning abstract relationships implied by the soundscape.

In this challenge, we adopt a parameter-efficient tuning strategy to adapt large audio-language models to the three sub-tasks of AQA. Our solution leverages `Qwen2-Audio` and `Kimi-Audio` as foundation models, with Low-Rank Adaptation (LoRA) [5] applied for efficient fine-tuning. The training process follows a standard two-stage post-training paradigm: Supervised Fine-Tuning (SFT) with constructed prompts, followed by Group Relative Policy Optimization (GRPO) based reinforcement learning (RL) to further improve accuracy and consistency. To enhance temporal reasoning, we reformatted a subset of AudioSet-Strong into QA pairs targeting event counting, sequencing, and duration. Our best model achieves 80.0% accuracy on the development set, demonstrating strong generalization across the three AQA sub-tasks.

[†] These authors contributed equally to this work.

2. METHODOLOGY

2.1. Foundation Models

2.1.1. Qwen2-Audio

`Qwen2-Audio` [3] is a large-scale audio-language model developed by Alibaba Group, consisting of a Whisper-large-v3-based audio encoder and a Qwen-7B language decoder. In this work, we adopt the instruction-tuned variant `Qwen2-Audio-7B-Instruct` as our base model. `Qwen2-Audio` is trained via a three-stage pipeline: multi-task pre-training with natural language prompts, SFT, and GRPO-based RL. It supports both audio analysis and voice chat modes, enabling seamless multi-modal interaction without explicit switching. The model achieves state-of-the-art performance on benchmarks such as AIR-Bench and VocalSound, and its strong generalization and instruction-following capabilities make it well-suited for DCASE 2025 Task 5, particularly for temporal reasoning and complex audio understanding.

2.1.2. Kimi-Audio

`Kimi-Audio` [6] is an open-source audio-language foundation model developed by Moonshot AI, featuring a hybrid architecture that integrates a Whisper-large-v3-based audio tokenizer with a 7B-scale language decoder. In this study, the instruction-tuned variant `Kimi-Audio-7B-Instruct` is adopted as the base model. The model is pre-trained on over 13 million hours of diverse audio data using a combination of uni-modal tasks, audio-text alignment tasks, and interleaved multi-modal tasks. To bridge the modality gap between audio and text, a dual-token input format—comprising discrete semantic tokens and continuous acoustic features—is employed and processed at a resolution of 12.5 Hz. Although only the pre-trained model is utilized without further SFT, `Kimi-Audio` exhibits strong audio representation and generalization capabilities, making it well-suited for DCASE 2025 Task 5, especially in addressing complex audio question answering tasks across Bio-acoustics QA, Temporal Soundscapes QA, and MMAU [7].

2.2. Data Preparation and Augmentation

To supplement the limited training data available for this task, we extended our training set by reformatting the annotations of the AudioSet-Strong [8] dataset into question answer pairs while using the MMAU dataset [7].

For the MMAU dataset, we utilized the `test-mini` subset containing 1,000 audio QA samples. Each sample includes:

$$\text{Sample} = \{\text{Audio}, Q, A_{\text{correct}}, \{D_1, D_2, D_3\}\}, \quad (1)$$

where D_i represents distractors. We standardized these samples by adding letter codes to match the task format.

The original AudioSet Strong annotations follow an *onset offset event-label* format. Our restructuring pipeline involved:

- Removing ambiguous event tags
- Temporal relationship extraction, including event occurrence sequence and duration
- Generate temporal audio QA samples

Based on task dataset analysis, we categorized questions into four distinct types:

Category	Example
Counting	How many different sounds are in the audio clip?
Sequencing	What is the sequence of the first two sounds in the audio?
Duration	What is the duration of the event-label sound?
Frequency	How many times does the event-label sound occur?

Table 1: Temporal audio question taxonomy

For each type, we prepared a fixed description of several questions. Then, for each audio sample in AudioSet Strong, we randomly generated a type of question and gave the correct answer. Finally, we called the GLM-4 API to generate three distractors to form an audio QA sample.

During the training process, for the audio part of the samples, we use the SpecAugment in frequency dimension. For the text part, we shuffle the order of options during data sampling to prevent the model from over-fitting to positional biases present in the training data.

2.3. Supervised Fine-Tuning

Our SFT approach focuses on adapting the foundation models to the specific requirements of AQA. We construct detailed prompts that include both the audio content (represented as embeddings) and the question text, formatted to guide the model towards generating accurate and informative answers.

The SFT process focuses on optimizing the **answer generation** while preserving the model's ability to effectively process audio inputs. Given a sample $(\mathbf{q}, \mathbf{x})$, where $\mathbf{q}$ denotes the question, and $\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$ denotes the answer sequence of length T, the objective is to maximize the probability of the next answer token conditioned on the previous tokens and the question:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log P(x_t \mid \{x_{<t}\}, \mathbf{q}) \quad (2)$$

2.4. GRPO-based Reinforcement Learning

To further enhance the performance of large audio-language models on the AQA task, we explore RL following SFT to

Table 2: Performance on development set

Model	Accuracy				
	Part1	Part2	Part3	Arithmetic Mean	Weighted Mean
`Qwen2-Audio` (zero-shot)	30.0%	39.2%	49.6%	39.6%	45.0%
`Kimi-Audio` (zero-shot)	43.3%	42.5%	60.3%	48.7%	53.8%
`Qwen2-Audio` (SFT)	82.4%	59.3%	80.0%	73.9%	75.2%
`Qwen2-Audio` (SFT+RL)	83.0%	**62.6%**	80.1%	75.2%	76.1%
`Kimi-Audio` (SFT)	87.5%	56.8%	85.3%	76.5%	78.5%
Both (SFT+Ensemble)	**88.0%**	60.1%	**86.3%**	**78.1%**	**80.0%**

strengthen its generalization capabilities. Our exploration is based on the `Qwen2-Audio-7B-Instruct model` [3], and we leverage the GRPO [9, 10] algorithm, which eases the burden of training an additional value function approximation model in proximal policy optimization (PPO) [11].

Specifically, given an input question q, a group of G outputs $\mathbf{o} = \{o_1, o_2, \cdots, o_G\}$ is first sampled from the old policy $\pi_{\theta_{\text{old}}}$. The corresponding rewards $\mathbf{r} = \{r_1, r_2, \cdots, r_G\}$ are then computed using a rule-based reward function that evaluates correctness: if a response provides the correct answer, it receives a reward of $+1$; otherwise, it is assigned a reward of 0. We employ the average reward as the baseline and the advantage is computed as:

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}. \tag{3}$$

Subsequently, the policy model π_θ is optimized by maximizing the following objective:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}[q, \mathbf{o} \sim \pi_{\theta_{\text{old}}}(O|q)] \\ \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}\hat{A}_{i,t},\right.\right. \\ \left.\text{clip}\left(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] \\ \left.\left. - \beta\mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]\right\}\right],\end{aligned} \tag{4}$$

where π_{ref} denotes the reference policy, adopting the initial SFT model in this work, $\mathbb{D}_{\text{KL}}$ serves to regularize the KL divergence between the trained policy and the reference, ϵ is the clipping threshold introduced in PPO [11] to stabilize training, and β is the scaling coefficient for the KL penalty.

3. EXPERIMENTS

3.1. Experiment Setting

All experiments were conducted on 8 NVIDIA A800 GPUs using `bf16` mixed-precision training and optimized with DeepSpeed ZeRO Stage 2. For the `Kimi-Audio` model, training was performed for 8 epochs with a learning rate of 5e-5, batch size of 32, and a LoRA rank of 16. For the `Qwen2-Audio` model, the supervised fine-tuning stage was conducted for 8 epochs with a learning rate of 5e-5, batch size of 32, and LoRA rank of 8. The subsequent GRPO-based RL stage was trained for 3 additional epochs with a reduced learning rate of 5e-6, batch size of 8, gradient accumulation steps of 5 (resulting in an effective batch size of 40), a sampling group size of $G = 6$, a clipping threshold of $\epsilon = 0.2$, a scaling coefficient for the KL penalty of $\beta = 0.1$.

On the development set, in addition to single-model inference, we adopted an ensemble strategy by integrating the predictions of multiple fine-tuned models (denoted as **Ensemble**). Specifically, we introduce a lightweight neural module, which takes as input the prediction logits from multiple fine-tuned models and outputs a fused probability distribution. Each model's logits are first adjusted using a learnable temperature parameter to calibrate confidence via temperature scaling. Subsequently, the softmax-normalized probabilities from each model are aggregated using learnable ensemble weights, which are also optimized during training. Both the weights and temperature parameters are updated via backpropagation to minimize the cross-entropy loss against ground-truth labels. The training process is performed using the Adam optimizer [12] over 500 steps. This approach allows the ensemble to dynamically learn both the relative reliability and calibration of each individual model, leading to improved overall prediction accuracy.

3.2. Results

As shown in Table 2, GRPO-based RL significantly improved the performance of `Qwen2-Audio` on Task Part 2. In terms of model comparison, `Kimi-Audio` outperformed `Qwen2-Audio` on Task Parts 1 and 3, but under-performed on Task Part 2. By integrating the predictions of both models, the ensemble achieved the best overall performance. Among the three sub-tasks, Task Part 2 proved to be the most challenging, indicating a potential direction for future improvement.

4. CONCLUSION

In this technical report, we present our solution for DCASE 2025 Challenge Task 5, focusing on parameter-efficient tuning of large audio-language models for the AQA task. By leveraging `Qwen2-Audio` and `Kimi-Audio` as foundation models, we adopt a two-stage training framework consisting of SFT and GRPO-based RL. To enhance temporal reasoning capabilities, we further introduce an augmented QA dataset reformulated from AudioSet-Strong. Experimental results on the development set demonstrate that both `Qwen2-Audio` and `Kimi-Audio` achieve strong performance across the three AQA sub-tasks, with the ensemble of both models yielding the highest accuracy of 80.0%. Our findings highlight the effectiveness of lightweight tuning strategies for adapting large pre-trained models to complex audio reasoning tasks, and point to future opportunities in improving temporal understanding and generalization in open-domain AQA.

5. REFERENCES

[1] S. Ghosh, Z. Kong, S. Kumar, *et al.*, "Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities," in *International Conference on Machine Learning (ICML)*, 2025.

[2] KimiTeam, D. Ding, Z. Ju, *et al.*, "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025.

[3] Y. Chu, J. Xu, Q. Yang, *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[4] A. Huang, B. Wu, B. Wang, *et al.*, "Step-audio: Unified understanding and generation in intelligent speech interaction," *arXiv preprint arXiv:2502.11946*, 2025.

[5] E. J. Hu, yelong shen, P. Wallis, *et al.*, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[6] D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang, *et al.*, "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025.

[7] S. Sakshi, U. Tyagi, S. Kumar, *et al.*, "MMAU: A massive multi-task audio understanding and reasoning benchmark," in *International Conference on Learning Representations (ICLR)*, 2025.

[8] S. Hershey, D. P. W. Ellis, E. Fonseca, *et al.*, "The benefit of temporally-strong labels in audio event classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 366–370.

[9] Z. Shao, P. Wang, Q. Zhu, *et al.*, "DeepseekMath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[10] G. Li, J. Liu, H. Dinkel, *et al.*, "Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering," *arXiv preprint arXiv:2503.11197*, 2024.

[11] J. Schulman, F. Wolski, P. Dhariwal, *et al.*, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.