AUDIO QUESTION ANSWERING AT THE DCASE 2025 CHALLENGE

Technical Report

Haolin He[1,*], *Mingru Yang*[2,*], *Renhe Sun*[3,*,†], *Jiayi Zhou*[3], *Jian Liu*[3], *Qianhua He*[2], *Qiuqiang Kong*[1]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China
harlandzzc@link.cuhk.edu.hk, qqkong@ee.cuhk.edu.hk
[2] School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
eemryang@mail.scut.edu.cn, eeqhhe@scut.edu.cn
[3] Machine Intelligence, Ant Group, Shanghai, China
{sunrenhe.srh, zjy326112, rex.lj}@antgroup.com

[*] Equal contribution.
[†] Corresponding author.

ABSTRACT

In this technical report, we describe the submission system for DCASE2025 Task 5: Audio Question Answering. In this work, we introduce a comprehensive audio question answering dataset named **DCASE-AQA-Boost**, featuring diverse question types and carefully curated answer options to address the limitations of existing collections. Based on the DCASE-AQA-Boost, we have developed two models, Kimi-Audio-SFT-12B and Qwen2-Audio-R1-8B. **Kimi-Audio-SFT-12B** is obtained through a two-stage Supervised Fine-Tuning (SFT) process using the Pretraining and Finetuning split of DCASE-AQA-Boost. **Qwen2-Audio-R1-8B** is trained using our proposed three-stage training paradigm based on the DCASE-AQA-Boost and DCASE2025 task 5 training set, which incorporates Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO).

Experimental results demonstrate that the proposed method significantly improves the accuracy of multiple-choice audio question answering systems. Kimi-Audio-SFT-12B and Qwen2-Audio-R1-8B achieve 77.66% and 78.18% accuracy on the DCASE2025 Task 5 development set, respectively.

***Index Terms*—** DCASE2025, Audio Question Answering, Multi-stage Training, GRPO

1. INTRODUCTION

Audio Question Answering (AQA) represents a sophisticated multimodal challenge that combines audio processing and natural language understanding. This task requires systems to comprehend audio signals and generate precise responses to queries about the audio content.

The complexity of AQA lies in its multi-layered requirements: systems must first process raw audio inputs, then extract meaningful features, and finally perform complex reasoning to provide contextually appropriate answers. This process involves understanding temporal relationships, identifying acoustic events, and making logical inferences from the audio content. Building upon the foundation of automated audio captioning, AQA extends beyond simple description to enable interactive understanding of audio content.

Recent Large Audio Language Models [1], such as Qwen-Audio [2], Qwen2-Audio [3], and Audio-Flamingo [4], primarily enhance their instruction-following ability and task-specific performance through Supervised Fine-Tuning (SFT). This fine-tuning process typically involves aligning the model with human-annotated audio-text pairs to better adapt to downstream tasks like audio captioning, retrieval, and question answering. Group Relative Policy Optimization (GRPO) [5], recently proposed by DeepSeek, is a promising reinforcement learning method for fine-tuning large models. R1-AQA [6] applies GRPO to Qwen2-Audio on the AVQA [7] dataset, achieving state-of-the-art results on the MMAU benchmark [8].

In this technical report, we introduce two models: Qwen2-Audio-R1-8B and Kimi-Audio-SFT-12B. **Qwen2-Audio-R1-8B** is built upon Qwen2-Audio-Instruct-7B [3] and optimized using a customized three-stage training paradigm. This paradigm incorporates SFT and GRPO, aiming to first enhance general audio question-answering capabilities and subsequently improve performance on DCASE2025 Task 5. The final model achieves an accuracy of 77.66% on the DCASE2025 Task 5 development set. **Kimi-Audio-SFT-12B** is based on the pre-trained audio-language model Kimi-Audio [9] and is optimized through a two-stage SFT process. The first stage uses the Pretraining Split of DCASE-AQA-Boost, while the second stage uses the Finetuning Split of DCASE-AQA-Boost. The final model achieves an accuracy of 78.18% on the DCASE2025 Task 5 development set.

A custom-built audio multiple-choice question dataset, referred to as **DCASE-AQA-Boost**, is used to optimize our models. DCASE-AQA-Boost consists of a Pretraining Split and a Finetuning Split. The Pretraining Split contains general audio question-answer pairs across four categories: sound, music, speech, and a cross-cutting category termed temporal. The Finetuning Split is carefully curated to closely align with the DCASE2025 Task5 dataset, with the goal of improving model performance on the corresponding evaluation set. A more comprehensive version of the dataset with broader applicability will be released in future work.

2. DATA SOURCES

2.1. Pretraining Split

The following sources are utilized to construct the pretraining-split of the dataset:

- **Clotho [10]:** A human-annotated audio captioning dataset featuring multiple captions per audio sample, designed to provide rich descriptive content for general audio understanding tasks.
- **AudioCaps 2.0 [11]:** A comprehensive dataset pairing audio clips with human-written textual descriptions, providing diverse acoustic content across various environmental and everyday sound scenarios.
- **MusicCaps [12]:** A specialized music dataset with structured annotations authored by professional musicians, ensuring domain-expert quality in musical descriptions and aspect analysis.
- **LP-MusicCaps-MTT [13]:** A music dataset with synthetically generated captions using large language models, employing tag-to-caption generation from MagnaTagATune dataset tags.
- **CompA-R [14]:** An audio question-answering dataset synthesized through a three-stage pipeline involving multimodal caption generation, instruction-response synthesis, and human verification for quality assurance.
- **SpeechCraft (LibriTTS-R split) [15]:** A bilingual expressive speech dataset featuring automatic natural language descriptions generated through expert classifiers and fine-tuned language models for speech-language learning.
- **TACOS [16]:** A temporal audio captioning dataset providing precise temporal localization of sound events with comprehensive timing annotations including onset, offset, and duration information.

2.2. Finetuning Split

A small portion of **VocalSound [17]** and **VGGSound [18]** datasets is also incorporated, along with **MMAU-test-mini [8]** and the **DCASE [19]** training set, to construct high-quality multiple-choice questions for fine-tuning.

- **VocalSound:** A crowdsourced dataset containing recordings of human vocal sounds including laughter, sighs, coughs, throat clearing, sneezes, and sniffs.
- **VGGSound:** A large-scale audio-visual dataset with clips from YouTube videos, spanning 310+ classes across challenging acoustic environments.
- **MMAU-test-mini:** A subset of the MMAU benchmark containing curated audio clips with expert-level questions across 27 diverse tasks requiring advanced audio understanding.
- **DCASE Training Set:** Audio question-answering data from DCASE 2025 challenge, including Bioacoustics QA, Temporal Soundscapes QA, and Complex QA subsets.

3. DATASET CONSTRUCTION

DCASE-AQA-Boost is divided into two parts: Pretraining Split and Finetuning Split. The Pretraining Split consists of general audio question-answer pairs falling into four categories: *sound*, *music*, *speech*, and a cross-cutting type called *temporal*. This split is designed to enhance the model's general audio question answering capabilities. The Finetuning Split is carefully constructed to be highly co-distributed with the DCASE2025 Task 5 dataset, focusing on improving the model's performance specifically on the DCASE2025 Task5 evaluation set.

3.1. Pretraining Split

To construct the Pretraining Split of DCASE-AQA-Boost, we design a fully automated pipeline based on Qwen3-235B [20]. This pipeline converts datasets from various audio tasks into a unified question answering format. It consists of three key steps: Basic QA Formation, Multiple-Choice Question Construction, and Automated Quality Gating.

Basic QA Formation. This stage converts datasets from various audio tasks into a basic QA formation. Among the data sources used to construct the Pretraining Split (refer to 2.1), only CompA-R contains native question-answer pairs. Other datasets, including Clotho, AudioCaps 2.0, MusicCaps, LP-MusicCaps-MTT, SpeechCraft, and TACOS, require conversion into the basic QA format.

Multiple-Choice Question Construction. This stage constructs multiple-choice questions based on the basic question-answer pairs. The multiple-choice questions are represented as structured items, each consisting of: (i) a context-based question, (ii) three incorrect distractors, (iii) one correct answer, and (iv) question types. The question types fall into four categories: Sound-based Questions, Music-based Questions, and Speech-based Questions, along with a cross-cutting type called Temporal Questions. TACOS is exclusively used to generate Temporal Questions, as it focuses on temporal sound event localization. For the remaining datasets, flexible question generation is allowed across the three primary types: *sound*, *music*, and *speech*. This design maintains generative flexibility while ensuring diverse and balanced question construction.

Automated Quality Gating. This stage implements automated quality control. Qwen3-235B is used as an evaluator to assess each item across three dimensions, using a five-point rating scale: Answer Consistency, Incorrect Options Quality, and Language Fluency. Questions scoring below four are automatically filtered, ensuring only high-quality sound questions remain in the final dataset.

3.2. Finetuning Split

The Finetuning Split is used to fine-tune the model in order to enhance its performance specifically on the DCASE2025 Task 5 evaluation set. A natural question arises: What data should be selected to construct the Finetuning Split? To address this, the DCASE2025 Task 5 training set is first selected for fine-tuning. The performance of the fine-tuned model is then evaluated on the DCASE2025 Task 5 development set, which consists of three parts: Part 1: *Bioacoustics QA*, Part 2: *Temporal Soundscapes QA*, and Part 3: *Complex QA*. The development results showed that the model performed best on Part 1, while only achieving suboptimal results on Parts 2 and 3.

To specifically enhance performance on Part 2, a custom-built Temporal Augmentation Set is created. Its construction involves three steps: (i) audio event selection, (ii) audio concatenation, and (iii) multiple-choice question generation.

Audio Event Selection. Eight audio events are selected from VocalSound and VGGSound, all of which also appear in Part 2 of

Table 1: Question template samples for the two categories in the Temporal Augmentation Set: Sound Detection Questions and Sound Sequence Questions.

Categories	Question Template Samples
Sound Detection	What is the first occurring sound in the audio?
	What is the second sound in the audio clip?
	What is the third sound in the audio clip?
	What is the last occurring sound?
	What is the longest sound?
Sound Sequence	What is the sequence of sounds?
	What is the order of the sounds?
	In what order do the sounds occur?
	Which sound occurs before the [sound event]?
	Which sound occurs after the [sound event]?

the DCASE2025 Task5 dataset. These events are used to build the Temporal Augmentation Set.

Audio Concatenation. Two to four audio events are randomly selected and concatenated along the temporal dimension to form a composite audio clip. The total duration is kept under 30 seconds, with 0 to 1 second of silence randomly inserted before and after each event.

Multiple-Choice Question Generation. Based on the generated composite audio and the corresponding event timestamps, multiple-choice questions are generated using Qwen3-32B-Instruct [20]. These questions are divided into two categories: Sound Detection Questions and Sound Sequence Questions, with 500 multiple-choice questions generated for each category. The corresponding question templates for each category are provided in Table 1.

To specifically enhance performance on Part 3, the MMAU-test-mini dataset, which contains 1,000 audio-related complex multiple-choice questions, is selected for joint fine-tuning. Therefore, the Finetuning Split consists of three subsets: the DCASE2025 Task5 training set (8,221 items), the Temporal Augmentation Set (1,000 items), and MMAU-test-mini (1,000 items), resulting in a total of 10,221 multiple-choice questions.

4. TRAINING

4.1. Qwen2-Audio-R1-8B

Qwen2-Audio-R1-8B is developed based on the pretrained Qwen2-Audio-Instruct-7B. We introduce a three-stage training paradigm to optimize the model, aiming to enhance its overall performance across the three evaluation tasks of DCASE2025 Task 5: Bioacoustics QA, Temporal Soundscapes QA, and Complex QA.

Stage 1: Supervised Fine-Tuning. In this stage, Qwen2-Audio-R1-8B is trained using a broad range of audio question-answering datasets to enable unified learning across diverse audio QA tasks. The goal is to develop a model capable of handling four categories of questions: sound-based, music-based, speech-based, and temporal questions. To this end, we perform supervised fine-tuning using the Pretraining Split of the DCASE-AQA-Boost dataset within the SWIFT [21] framework.

Stage 2: Group Relative Policy Optimization. In this stage, Qwen2-Audio-R1-8B is optimized to improve its performance across the three components of DCASE2025 Task 5. To enhance training stability and task adaptability, GRPO [5, 6] is applied using the DCASE2025 Task 5 training set.

Stage 3: Group Relative Policy Optimization. This stage further refines Qwen2-Audio-R1-8B by reinforcing its performance on underperforming tasks while continuing to optimize across all three components of DCASE2025 Task 5. The training is conducted using the carefully curated Finetuning Split of the DCASE-AQA-Boost dataset, with GRPO employed as the training strategy.

4.2. Kimi-Audio-SFT-12B

For Kimi-Audio, a streamlined two-stage fine-tuning approach is adopted, tailored to the model's architectural characteristics. The training process is initiated with SFT using the pretraining dataset, establishing fundamental audio-language understanding capabilities across diverse audio domains represented in the dataset.

Subsequently, an additional SFT phase is conducted using the Finetuning Split. This stage specifically targets temporal audio understanding and sound event localization tasks, leveraging the structured nature of the DCASE benchmark to enhance the model's precision in temporal audio analysis and event recognition capabilities.

Table 2: Performance comparison on DCASE development set across training stages. Results show accuracy percentages for each model configuration.

Model Configuration	Accuracy (%)
Qwen2-Audio-R1-8B	
Baseline (no fine-tuning)	48.74
+ Pretrain-split SFT	54.80
+ DCASE train-split GRPO	77.17
+ Finetune-split GRPO	77.66
Kimi-Audio-SFT-12B	
Baseline (no fine-tuning)	54.83
+ Pretrain-split SFT	62.08
+ Finetune-split SFT	78.18

5. EXPERIMENTAL RESULTS

The fine-tuned models are evaluated on the DCASE development set to assess the effectiveness of the multi-stage training approach. Table 2 presents the performance progression of both models across different training stages, demonstrating significant improvements through the proposed fine-tuning methodology.

To provide a more comprehensive evaluation, Table 3 presents the breakdown of final model performance across different task categories in the DCASE development set.

The results demonstrate substantial performance gains for both models through our fine-tuning approach. For Qwen2-Audio-Instruct, the baseline accuracy of 48.74% improves to 77.66% after the complete three-stage training pipeline, representing a 28.92 percentage point improvement. The most significant gain occurs during the GRPO phase with DCASE training set, where accuracy jumps from 54.80% to 77.17%, highlighting the effectiveness of reinforcement learning-based optimization for audio understanding tasks.

Kimi-Audio achieves 78.18% accuracy after two-stage SFT training, starting from a baseline of 54.83%. This represents a 23.35 percentage point improvement. Notably, Kimi-Audio achieves slightly higher final performance (78.18% vs 77.66%) despite using a simpler training approach without GRPO, suggesting strong compatibility between the model architecture and our dataset construction methodology.

Table 3: Accuracy (%) of Kimi-Audio-SFT-12B and Qwen2-Audio-R1-8B on the DCASE2025 Task5 development set. Results are presented for the three sub-tasks: Bioacoustics QA, Temporal Soundscapes QA, and Complex QA.

Model	Bio	Temporal	Complex
Kimi-Audio-SFT-12B	90.62	59.93	83.28
Qwen2-Audio-R1-8B	83.48	60.43	83.28
Overall Accuracy			
Kimi-Audio-SFT-12B		78.18	
Qwen2-Audio-R1-8B		77.66	

The detailed breakdown in Table 3 reveals distinct performance patterns across different task categories. Both models excel in bioacoustics understanding, with Kimi-Audio-SFT-12B achieving 90.62% accuracy and Qwen2-Audio-R1-8B reaching 83.48% on the 224 bioacoustics samples. Performance on complex audio scenarios is equally strong for both models, with identical accuracy of 83.28% across 1,633 complex samples.

However, temporal reasoning remains the most challenging task category, where both models show relatively modest performance (Kimi: 59.93%, Qwen: 60.43% over 609 temporal samples). The gap between temporal and other task categories indicates significant room for improvement in sequential audio event reasoning capabilities.

Both models demonstrate that the initial SFT phase with DCASE-AQA-Boost provides meaningful improvements (6.06 and 7.25 percentage points respectively), establishing the foundation for subsequent training stages.

6. REFERENCES

[1] C.-K. Yang, N. S. Ho, and H.-y. Lee, "Towards holistic evaluation of large audio-language models: A comprehensive survey," *arXiv preprint arXiv:2505.15957*, 2025.

[2] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[3] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[4] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, "Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities," *arXiv preprint arXiv:2503.03983*, 2025.

[5] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[6] G. Li, J. Liu, H. Dinkel, Y. Niu, J. Zhang, and J. Luan, "Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering," *arXiv preprint arXiv:2503.11197*, 2025.

[7] P. Yang, X. Wang, X. Duan, H. Chen, R. Hou, C. Jin, and W. Zhu, "Avqa: A dataset for audio-visual question answering on videos," in *Proceedings of the 30th ACM international conference on multimedia*, 2022.

[8] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "Mmau: A massive multi-task audio understanding and reasoning benchmark," *arXiv preprint arXiv:2410.19168*, 2024.

[9] 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.

[10] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[11] C. D. Kim, B. Kim, H. Lee, and G. Kim, "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)*, 2019.

[12] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[13] S. Doh, K. Choi, J. Lee, and J. Nam, "Lp-musiccaps: Llm-based pseudo music captioning," *arXiv preprint arXiv:2307.16372*, 2023.

[14] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities," *arXiv preprint arXiv:2406.11768*, 2024.

[15] Z. Jin, J. Jia, Q. Wang, K. Li, S. Zhou, S. Zhou, X. Qin, and Z. Wu, "Speechcraft: A fine-grained expressive speech dataset with natural language description," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

[16] P. Primus, F. Schmid, and G. Widmer, "Tacos: Temporally-aligned audio captions for language-audio pretraining," *arXiv preprint arXiv:2505.07609*, 2025.

[17] Y. Gong, J. Yu, and J. Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[18] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[19] C.-H. H. Yang, S. Ghosh, Q. Wang, J. Kim, H. Hong, S. Kumar, G. Zhong, Z. Kong, S. Sakshi, V. Lokegaonkar, *et al.*, "Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge," *arXiv preprint arXiv:2505.07365*, 2025.

[20] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[21] Y. Zhao, J. Huang, J. Hu, X. Wang, Y. Mao, D. Zhang, Z. Jiang, Z. Wu, B. Ai, A. Wang, *et al.*, "Swift: a scalable lightweight infrastructure for fine-tuning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 28, 2025, pp. 29 733–29 735.