AUDIO-DEPENDENT QUESTION ANSWERING AT THE DCASE 2026 CHALLENGE

Technical Report

Weiteng Hu[1,2], *Yin Cao*[1], *Jun Yang*[1,2,3],

[1] Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
{huweiteng, jyang}@mail.ioa.ac.cn
yin.k.cao@gmail.com
[2] University of Chinese Academy of Sciences, Beijing, China
[3] State Key Laboratory of Acoustics and Marine Information, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China

ABSTRACT

In this technical report, we describe our submission systems for the DCASE 2026 Task 5: Audio-Dependent Question Answering (ADQA). In this work, we choose Qwen2.5-Omni and MOSS-Audio-8B-Thinking as our foundation models. For Qwen2.5-Omni, we investigate two post-training strategies with Low-Rank Adaptation (LoRA). The first system employs answer-only Supervised Fine-Tuning (SFT), followed by Reinforcement Learning (RL) optimization with accuracy and format rewards. The second system introduces a structured reasoning approach. We reconstruct the structured chain-of-thought (CoT) from AudioMCQ into a unified schema, including question analysis, question type, audio evidence, reasoning, and final answer. This model is trained with structured CoT SFT and then optimized using RL with a composite reward function. For MOSS-Audio-8B-Thinking, we perform direct inference without additional fine-tuning, leveraging its strong native reasoning capability. On the development set, our best Qwen2.5-Omni-based system achieves 58.93% top-1 accuracy, and our best MOSS-Audio-based system achieves 67.70% top-1 accuracy.

***Index Terms*—** Audio-Dependent Question Answering, Large Audio-Language Model, Chain of Thought, Reinforcement Learning

1. INTRODUCTION

Large Audio-Language Models (LALMs) have demonstrated remarkable performance on complex audio understanding tasks [1–5]. These models can process diverse audio inputs, including speech, music, and environmental sounds, and generate natural language responses for various downstream tasks such as audio captioning, automatic speech recognition (ASR), and audio question answering. Specifically, Audio Question Answering (AQA) requires LALMs to not only recognize specific acoustic events, but also comprehend textual questions, identify relevant audio evidence, and perform cross-modal reasoning to derive final answer.

However, research reveals that many models rely on text prompts and internal linguistic priors rather than actual audio perception in the AQA task [6]. In conventional AQA benchmarks like MMAU [7], MMSU [8], and MMAR [9], LALMs can answer some questions correctly even when the original audio is replaced by silent input. This phenomenon suggests that high accuracy on existing AQA benchmarks does not always indicate genuine audio understanding capability. To address this issue, DCASE 2026 Task 5 introduces Audio-Dependent Question Answering (ADQA), a multiple-choice question answering task where each question is designed to be truly dependent on the audio content. ADQA employs a rigorous evaluation framework based on Audio-Dependency Filtering (ADF) to distinguish weak and strong audio-contribution samples. The official DCASE 2026 Task 5 training set, AudioMCQ-StrongAC-GeminiCoT, is derived from strong audio-contribution split of AudioMCQ and includes native Chain-of-Thought (CoT) [10] reasoning labels generated by Gemini 3.1 Pro [11]. This dataset provides a valuable foundation for training models to rely more on acoustic evidence rather than text-only shortcuts.

In this technical report, we explore two LALMs as foundation backbones: Qwen2.5-Omni and MOSS-Audio-8B-Thinking. For Qwen2.5-Omni [5], we investigate two parameter-efficient post-training variants with Low-Rank Adaptation (LoRA) [12]. The first variant begins with answer-only Supervised Fine-Tuning (SFT), where the model is trained to produce the final answer directly. Then, it is further optimized with Group Relative Policy Optimization (GRPO) [13] using answer accuracy and format rewards. Although this system does not utilize explicit CoT supervision during SFT, the model is prompted to think step by step during GRPO, allowing reasoning behavior to emerge from answer-level reward optimization. The second variant adopts a Structured CoT paradigm inspired by the structured CoT annotations in AudioMCQ [6]. We reconstruct the original structured CoT into a unified schema that decomposes the answer generation process into four intermediate fields before the final answer: question analysis, question type, audio evidence, and reasoning. This design encourages the model to first understand what the question asks, then identify the required acoustic evidence, and finally infer the correct answer based on the audio. The Structured CoT system is trained with Structured CoT SFT and then optimized via Group reward-Decoupled Normalization Policy Optimization (GDPO) [14] using a composite reward function, which combines final answer accuracy, format correctness, question-type correctness, semantic similarity of intermediate reasoning fields, and length regularization. For MOSS-Audio-8B-Thinking, we submit two zero-shot inference systems, leveraging its strong native reasoning capability without additional fine-tuning. Empirical results show that the Qwen2.5-Omni post-training systems improve over the corresponding baseline, while the best MOSS-Audio system achieves superior overall performance on the development set.

2. METHODOLOGY

2.1. System Overview

Our submissions consist of four systems based on two LALM backbones: Qwen2.5-Omni and MOSS-Audio-8B-Thinking. The two Qwen2.5-Omni systems are adapted with parameter-efficient post-training, whereas the two MOSS-Audio-8B-Thinking systems are used for direct zero-shot inference.

2.2. Data Preparation

For the two Qwen2.5-Omni systems, we only use the official training set, AudioMCQ-StrongAC-GeminiCoT, as our post-training data. Although the training set includes native CoT reasoning generated by Gemini 3.1 Pro, we extract and reconstruct the structured CoT annotations from AudioMCQ for the corresponding training samples. This structured schema decomposes the answer generation process into four intermediate fields before the final answer: question analysis, question type, audio evidence, and reasoning. The intermediate reasoning schema is defined as follows:

```
<question_analysis>...</question_analysis>
<question_type>...</question_type>
<audio_evidence>...</audio_evidence>
<reasoning>...</reasoning>
```

The two MOSS-Audio-8B-Thinking systems do not use any training data, as they are submitted as zero-shot inference systems without additional fine-tuning.

2.3. Qwen2.5-Omni Systems

Qwen2.5-Omni [5] is an end-to-end multimodal model designed to perceive diverse modalities, including text, images, audio, and video. In this work, we build two post-training systems based on Qwen2.5-Omni-7B: **Qwen-CoT** and **Qwen-Structured-CoT**. **Qwen-CoT** is first trained through answer-only SFT and subsequently finetuned via GRPO under a step-by-step reasoning prompt, while **Qwen-Structured-CoT** explicitly learns a structured reasoning schema and is further optimized with composite rewards.

2.3.1. Qwen-CoT

Qwen-CoT follows a two-stage post-training pipeline in order to enhance its overall performance across the ADQA task of DCASE 2026 Task 5.

Stage 1: answer-only Supervised Fine-tuning. In this stage, we perform answer-only SFT. We only prompt the model to directly generate answer choice without any intermediate reasoning. This stage aims to adapt the model to the ADQA multiple-choice instructions and provides a stable initialization for subsequent reinforcement learning.

Stage 2: Group Relative Policy Optimization. In this stage, we further optimize the model with Group Relative Policy Optimization (GRPO) [13]. Unlike the answer-only SFT stage, the GRPO prompt instructs the model to think step by step before the final answer. The reward function for Qwen-CoT consists of an accuracy reward and a format reward:

$$R_{\mathrm{CoT}} = w_{\mathrm{acc}} r_{\mathrm{acc}} + w_{\mathrm{fmt}} r_{\mathrm{fmt}}, \tag{1}$$

where $r_{\mathrm{acc}} = 1$ if the extracted answer matches the ground-truth answer and $r_{\mathrm{fmt}} = 1$ if the response follows the required output format.

2.3.2. Qwen-Structured-CoT

Qwen-Structured-CoT also follows a two-stage post-training pipeline, but differs from Qwen-CoT in both the supervised target and the reward design.

Stage 1: Structured CoT Supervised Fine-tuning. In this stage, we perform structured CoT SFT using the reconstructed annotations described in Section 2.2. Instead of directly predicting only the final answer, the model is trained to generate the entire structured reasoning sequence, including question analysis, question type, audio evidence, reasoning, and the final answer. This stage forces the model to adhere to the structured output schema and serves as a stable warm-up for subsequent reinforcement learning.

Stage 2: Group Reward-Decoupled Normalization Policy Optimization. In this stage, we design a composite reward that evaluates both the final answer and intermediate reasoning fields. The total reward is defined as:

$$\begin{aligned} R_{\mathrm{struct}} =& w_{\mathrm{acc}} r_{\mathrm{acc}} + w_{\mathrm{fmt}} r_{\mathrm{fmt}} + w_{\mathrm{qtype}} r_{\mathrm{qtype}} \\ &+ w_{\mathrm{qa}} r_{\mathrm{sim}}^{\mathrm{qa}} + w_{\mathrm{ae}} r_{\mathrm{sim}}^{\mathrm{ae}} + w_{\mathrm{reason}} r_{\mathrm{sim}}^{\mathrm{reason}} \\ &+ w_{\mathrm{len}} r_{\mathrm{len}}. \end{aligned} \tag{2}$$

Here, r_{acc} and r_{fmt} denote the final answer accuracy reward and format reward respectively. The question-type reward r_{qtype} is 1 if the predicted question type matches the reference question type. The similarity rewards $r_{\mathrm{sim}}^{\mathrm{qa}}$, $r_{\mathrm{sim}}^{\mathrm{ae}}$, and $r_{\mathrm{sim}}^{\mathrm{reason}}$ denote the cosine similarity rewards calculated by Qwen3-Embedding-0.6B [15] between the model generated field and the reference annotations for question analysis, audio evidence, and reasoning, respectively. r_{len} serves as a length regularization reward to discourage overly short or overly long responses. Specially, the question-analysis similarity is gated by r_{qtype}, while the audio-evidence similarity, reasoning similarity, and length regularization reward are gated by r_{acc}. This gated design prevents the model from receiving high rewards for intermediate reasoning fields when the predicted question type or final answer is incorrect. To avoid reward collapse in standard GRPO, we adopt Group Reward-Decoupled Normalization Policy Optimization (GDPO) [14] in this stage, which normalizes each reward separately before aggregation.

2.4. MOSS-Audio-8B-Thinking Systems

MOSS-Audio [1] is a unified audio-language model designed for speech, environmental sound, and music understanding. It supports a broad range of audio-centered tasks, including audio captioning, time-aware question answering, timestamped transcription, and audio-grounded reasoning. In this work, we build two direct zero-shot inference systems based on MOSS-Audio-8B-Thinking for the ADQA task.

In our preliminary experiments, we also explored SFT for MOSS-Audio-8B-Thinking using the official training data. However, the fine-tuned model did not yield improvements on the development set. Therefore, we preserve MOSS-Audio-8B-Thinking as a zero-shot inference backbone in our final submissions.

We design two zero-shot inference systems for MOSS-Audio-8B-Thinking: **MOSS-Thinking-Full** and **MOSS-Thinking-Label**.

MOSS-Thinking-Full asks the model to output the complete answer text after reasoning, while **MOSS-Thinking-Label** constrains the model to output only the option label, e.g., A, B, C, D, or E, in its final answer.

2.5. Post-Processing

All submitted systems share the same post-processing pipeline. First, we extract the final answer from the model response and then map it to an option using exact or containment string matching. If no answer is extracted or no valid option is matched, we employ Qwen3-Embedding-0.6B [15] to select the candidate choice that is the most semantically similar to the model response. Furthermore, to mitigate option-order bias and enhance robustness, inference is performed across both the original choice order and four randomly shuffled choice orders for each system. The final answer is determined by majority voting across these five runs.

3. EXPERIMENTS

3.1. Experimental Setup

All experiments were conducted on 4 NVIDIA GeForce RTX 4090 GPUs using bf16 mixed-precision training. For the Qwen2.5-Omni systems, we applied LoRA with a LoRA rank of 8 and a LoRA alpha of 32, and all stages were implemented within the ms-swift [16] framework. The SFT stage was trained for one epoch with a learning rate of 1×10^{-4}, per-device batch size of 1, and gradient accumulation steps of 8. The RL stage was also trained for one epoch with a learning rate of 1×10^{-5}, KL regularization coefficient of $\beta = 0.001$, number of generations of 8, generation batch size of 32, and maximum completion length of 512. For MOSS-Audio-8B-Thinking systems, we conducted zero-shot inference with a temperature of 1.0, top-p of 1.0, and top-k of 50.

3.2. Results

All the models are evaluated on the DCASE development set. Table 1 reports the top-1 accuracy of our submitted systems and the intermediate training stages.

For the Qwen2.5-Omni systems, both post-training strategies improve baseline performance. For **Qwen-CoT** system, the answer-only SFT stage slightly improves the baseline from 55.88% to 56.50%, and the subsequent GRPO stage further increases the accuracy to 58.93%. This indicates that answer-level reinforcement learning with accuracy and format rewards is effective for adapting the model to the ADQA task setting.

Table 1: Top-1 accuracy of different systems and training stages on the development set.

System	Top-1 Accuracy (%)
Qwen-CoT	
Baseline (no fine-tuning)	55.88
+ Answer-Only SFT	56.50
+ GRPO	58.93
Qwen-Structured-CoT	
Baseline (no fine-tuning)	55.88
+ Structured CoT SFT	57.93
+ GDPO	58.93
MOSS-Thinking-Full	66.02
MOSS-Thinking-Label	67.70

Qwen-Structured-CoT system exhibits a larger gain after the structured CoT SFT stage, improving the accuracy from 55.88% to 57.93%. This suggests that structured reasoning supervision helps the model better adapt to the task format and encourages more organized audio-dependent reasoning. After GDPO optimization with the structured reward function, the final accuracy reaches 58.93%, which is the same as the final **Qwen-CoT** system. Although the structured reward design does not outperform the **Qwen-CoT** system, it provides a more interpretable and controllable reasoning format by explicitly constraining intermediate fields such as question type, audio evidence, and reasoning.

For the MOSS-Audio-8B-Thinking systems, direct zero-shot inference achieves substantially higher performance than the fine-tuned Qwen2.5-Omni systems. **MOSS-Thinking-Full** system obtains 66.02% top-1 accuracy, while **MOSS-Thinking-Label** system further improves the accuracy to 67.70%. The improvement of the option label output setting suggests that directly generating the option label can reduce answer-mapping ambiguity during post-processing. In our preliminary experiments, we attempted SFT for MOSS-Audio-8B-Thinking using the official training data. However, the fine-tuned model failed to surpass the zero-shot inference baseline on the development set, with its top-1 accuracy dropping sharply to 58.99% on the development set. This degradation suggests that the robust native reasoning and instruction-following capabilities of MOSS-Audio-8B-Thinking can be easily disrupted by task-specific SFT on limited data. In particular, fine-tuning on a relatively small task-specific dataset may cause catastrophic forgetting and overfitting to the training format, thereby compromising the model's original general audio reasoning capability. Therefore, we use MOSS-Audio-8B-Thinking only as a direct zero-shot inference backbone for our final submissions.

Overall, two main observations can be drawn from the empirical results. First, post-training with CoT consistently improves Qwen2.5-Omni over the baseline, demonstrating the effectiveness of CoT in the ADQA task. Second, the superior zero-shot inference performance of MOSS-Audio-8B-Thinking indicates that the native audio reasoning capability of the foundation model remains a dominant factor for the ADQA task.

4. CONCLUSION

In this technical report, we present our submissions for DCASE 2026 Task 5: Audio-Dependent Question Answering. We explore two types of systems: parameter-efficient post-training based on Qwen2.5-Omni and direct zero-shot inference based on MOSS-Audio-8B-Thinking. For Qwen2.5-Omni, both post-training strategies improve over the baseline. Specifically, **Qwen-CoT** conducts answer-only SFT followed by CoT prompted GRPO, while **Qwen-Structured-CoT** performs Structured CoT SFT followed by GDPO with structured rewards. This result shows that task-specific SFT and reinforcement learning are highly effective in adapting foundational models to complex ADQA scenarios. The Structured CoT system further provides a more interpretable and controllable reasoning format by explicitly modeling question analysis, question type, audio evidence, and reasoning before the final answer. Meanwhile, MOSS-Audio-8B-Thinking achieves superior zero-shot inference performance, highlighting the importance of the foundation model's native audio reasoning capability. Among our submitted systems, the best Qwen2.5-Omni-based system achieves 58.93%

top-1 accuracy on the development set, while the best MOSS-Audio-based system achieves 67.70%.

5. REFERENCES

[1] C. Yang, C. Yu, H. Chen, J. Zhu, J. Chen, K. Chen, W. Wang, Y. Wang, Y. Jiang, Y. Jiang, *et al.*, "Moss-audio technical report," *arXiv preprint arXiv:2606.01802*, 2026.

[2] S. Ghosh, A. Goel, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. Yang, R. Duraiswami, D. Manocha, R. Valle, *et al.*, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," *Advances in Neural Information Processing Systems*, vol. 38, pp. 41 819–41 886, 2026.

[3] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, *et al.*, "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025.

[4] D. Zhang, G. Wang, J. Xue, K. Fang, L. Zhao, R. Ma, S. Ren, S. Liu, T. Guo, W. Zhuang, *et al.*, "Mimo-audio: Audio language models are few-shot learners," *arXiv preprint arXiv:2512.23808*, 2025.

[5] X. Jin, G. Zhifang, H. Jinzheng, H. Hangrui, H. Ting, B. Shuai, C. Keqin, W. Jialin, F. Yang, D. Kai, Z. Bin, W. Xiong, C. Yunfei, and L. Junyang, "Qwen2.5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[6] H. He, X. Du, R. Sun, Z. Dai, Y. Xiao, M. Yang, J. Zhou, X. Li, Z. Liu, Z. Liang, *et al.*, "Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models," in *International Conference on Learning Representations*, 2026.

[7] 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," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 84 929–84 964.

[8] D. Wang, J. Li, J. Wu, D. Yang, X. Chen, T. Zhang, and H. Meng, "Mmsu: A massive multi-task spoken language understanding and reasoning benchmark," *arXiv preprint arXiv:2506.04779*, 2025.

[9] Z. Ma, Y. Ma, Y. Zhu, C. Yang, Y.-W. Chao, R. Xu, W. Chen, Y. Chen, Z. Chen, J. Cong, *et al.*, "Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix," *Advances in Neural Information Processing Systems*, vol. 38, 2026.

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[11] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." *Iclr*, vol. 1, no. 2, p. 3, 2022.

[13] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[14] S.-Y. Liu, X. Dong, X. Lu, S. Diao, P. Belcak, M. Liu, M.-H. Chen, H. Yin, Y.-C. F. Wang, K.-T. Cheng, *et al.*, "Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization," *arXiv preprint arXiv:2601.05242*, 2026.

[15] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, *et al.*, "Qwen3 embedding: Advancing text embedding and reranking through foundation models," *arXiv preprint arXiv:2506.05176*, 2025.

[16] 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.