STRUCTURED AUDIO REASONING AND ROBUST MULTI-SAMPLE INFERENCE FOR DCASE 2026 AUDIO-DEPENDENT QUESTION ANSWERING CHALLENGE

Technical Report

Yucong Zhang[1,2], *Juan Liu*[3,1], *Ming Li*[2,3]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China
[3] School of Artificial Intelligence, Wuhan University, Wuhan, China
yucong.zhang@whu.edu.cn

ABSTRACT

We present a structured reasoning framework for the DCASE 2026 Audio-Dependent Question Answering task. Qwen3-Omni-30B-A3B-Instruct is used as an offline teacher to generate structured reasoning targets containing audio evidence, option-level judgments, and final answers for training MOSS-Audio-8B-Thinking. The student model is optimized through full-response SFT and GRPO-style post-training, followed by choice-permuted eight-sample voting at inference time. Our complete system improves the original MOSS-Audio baseline by 10.58 absolute percentage points. Under eight-sample inference, the resulting 8B student achieves 62.79% accuracy on the development set, numerically comparable to the 62.48% official Qwen3-Omni baseline.

***Index Terms*—** audio-dependent question answering, audio-language models, structured reasoning, supervised fine-tuning, group-relative optimization

1. INTRODUCTION

Audio-dependent question answering [1, 2, 3] requires a model to jointly understand an audio recording, a natural-language question, and a set of candidate answers. Unlike text-only multiple-choice reasoning, the correct answer must be grounded in acoustic events, spoken content, or other properties of the input audio. In practice, audio-language models may still rely on textual priors, produce reasoning that is weakly connected to the audio, or exhibit sensitivity to the ordering of candidate choices. These issues make both reliable audio grounding and robust answer selection important for the DCASE 2026 audio-dependent question answering task [4].

We address these challenges through structured reasoning distillation and group-relative post-training [5]. We first use Qwen3-Omni-30B-A3B-Instruct [6] as an offline teacher to transform the provided Gemini-generated chain-of-thought (CoT) annotations into a unified format containing audio evidence, option-level judgments, and the final answer. The teacher generation process is independently repeated three times with different random seeds, providing multiple reasoning annotations for the same question. After automatic validation, the resulting structured data are used to fine-tune MOSS-Audio-8B-Thinking [7] with full-response supervision.

Starting from the supervised model, we apply a GRPO-style group-relative objective that jointly rewards answer correctness, option-level judgment quality, answer–judgment consistency, and output format validity. Candidate choices are randomly permuted during post-training to reduce dependence on their original positions. At inference time, we generate eight stochastic predictions under different choice orders, map every prediction back to the original option order, and obtain the final answer through majority voting.

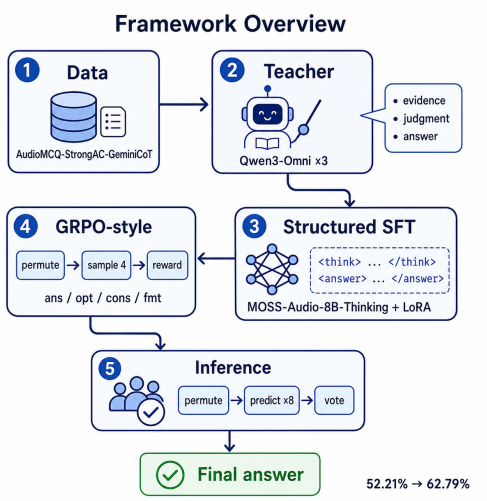

Figure 1: The overview of the AQA pipeline.

On the official development set, the complete system achieves an accuracy of 62.79%, compared with 52.21% for the original MOSS-Audio baseline. The results show that structured reasoning supervision substantially improves audio-dependent question answering, while choice-order randomized multi-sample aggregation further increases the robustness of the final prediction.

2. METHOD

2.1. System Overview

As Figure 1 shows, Our system consists of four components: multi-sample structured annotation generation, full-response SFT, group-relative post-training, and choice-order randomized vote aggregation. The first three stages produce the final student model, while the last stage is applied only during inference.

2.2. Structured Reasoning Distillation

The original training data contain an audio recording, a question, multiple candidate answers, a ground-truth answer, and a Gemini

CoT annotation. Instead of directly using the free-form CoT as the SFT target, we employ Qwen3-Omni-30B-A3B-Instruct to transform it into a unified structured representation.

Given the audio, question, candidate options, ground-truth answer, and the original Gemini reasoning, the teacher model produces three fields:

- `audio_evidence`: a concise description of the audio content that is directly relevant to the question;
- `option_judgment`: an explicit assessment of every candidate as *supported*, *contradicted*, or *unsupported*, together with a short justification;
- `answer`: the text of the selected candidate option.

We independently run the teacher generation process three times using different random seeds. Therefore, the same question may be associated with multiple valid reasoning paths rather than a single deterministic annotation. These independently sampled annotations increase the linguistic and reasoning diversity of the SFT data while preserving the same ground-truth answer.

Teacher generations are automatically validated before training. We discard an annotation if any required field is missing, if the generated answer does not match the ground-truth answer, if the option-level judgments do not cover all candidates, or if the audio-evidence field falls outside the predefined length range. Starting from 58,440 candidate annotations, 58,020 valid structured targets are retained. Consequently, each original question contributes up to three SFT examples.

2.3. Structured Reasoning Supervised Fine-Tuning

We use MOSS-Audio-8B-Thinking as the student model. Each training input contains the audio recording, the question, and the candidate options. The target response is organized as

$$y = \langle \texttt{think} \rangle\, [e,\ j]\, \langle / \texttt{think} \rangle \langle \texttt{answer} \rangle a \langle / \texttt{answer} \rangle, \quad (1)$$

where e denotes the audio evidence, j denotes the option-level judgments, and a denotes the final answer.

The explicit `think` block aligns the structured reasoning target with the native generation format of the student model. The final answer remains outside the reasoning block so that it can be reliably parsed during evaluation.

We perform parameter-efficient fine-tuning using LoRA [8] with rank 16 and scaling factor 32. The language-model projection layers are adapted, while the audio encoder and audio projector remain frozen. The language modeling loss is computed over the complete assistant response, including the structured reasoning and final answer:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t \in \mathcal{Y}} \log p_\theta \left(y_t \mid x, y_{<t} \right), \quad (2)$$

where x contains the audio and textual prompt, and $\mathcal{Y}$ denotes the assistant-response tokens. Prompt and audio-input tokens are masked from the loss.

2.4. Group-Relative Post-Training

Although SFT teaches the desired reasoning format, it assigns equal token-level importance to both correct and imperfect generated reasoning patterns. We therefore continue training the SFT model using a GRPO-style group-relative objective that directly evaluates sampled responses.

For each training question, we randomly permute the candidate choices and sample a group of $G = 4$ responses. The order of the choices is updated consistently in the prompt and in the answer mapping. For a sampled response y_i, the total reward is

$$R_i = 0.65 R_i^{\mathrm{ans}} + 0.20 R_i^{\mathrm{opt}} + 0.10 R_i^{\mathrm{con}} + 0.05 R_i^{\mathrm{fmt}}. \quad (3)$$

The four reward terms are defined as follows:

- R^{ans} evaluates whether the generated final answer matches the ground-truth candidate;
- R^{opt} evaluates the agreement between the generated option-level judgments and the teacher-derived structured reference;
- R^{con} evaluates whether the selected answer is consistent with the model's own option judgments, e.g., whether the selected candidate is labeled as supported;
- R^{fmt} evaluates structural validity, including the presence of the required fields, coverage of all candidate options, and successful answer parsing.

Answer correctness receives the largest weight so that improvements in reasoning structure cannot compensate for an incorrect final answer. The remaining terms encourage interpretable and internally consistent reasoning.

The rewards are standardized within each response group:

$$A_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon}, \quad (4)$$

where μ_R and σ_R are the mean and standard deviation of the four rewards in the same group. The group-relative training loss is

$$\mathcal{L}_{\mathrm{GRPO}} = -\frac{1}{G} \sum_{i=1}^{G} A_i \left[\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \log p_\theta \left(y_{i,t} \mid x_i, y_{i,<t} \right) \right]. \quad (5)$$

This objective increases the likelihood of responses whose rewards are above the group average and decreases the likelihood of below-average responses. Averaging token log-probabilities within each response reduces the direct influence of response length.

2.5. Choice-Order Randomized Multi-Sample Inference

Multiple-choice audio models may be sensitive to the positions of the candidate options, while stochastic decoding can produce different answers for the same input. We address both issues through choice-order randomized self-consistency.

For each evaluation question, we generate eight stochastic responses. The candidate choices are randomly reordered for every vote, using distinct permutations whenever possible. After generation, the predicted answer is matched to an option under the current permutation and then mapped back to its index in the original candidate ordering. The final prediction is obtained by voting over these original option indices:

$$\hat{a} = \arg\max_{k} \sum_{v=1}^{8} \mathbb{I} \left[\pi_v^{-1} \left(\hat{a}^{(v)} \right) = k \right], \quad (6)$$

where π_v denotes the choice permutation used for the v-th generation and $\hat{a}^{(v)}$ is the corresponding predicted option.

Mapping every prediction back to the original ordering ensures that votes refer to option semantics rather than temporary option positions. This procedure combines stochastic self-consistency with choice-order augmentation and produces the final submitted answer.

Table 1: Development-set results of the main system variants. The preliminary experiments validate the effectiveness of structured reasoning supervision, while the final-system block shows the progressive construction of our submitted system. All results use single-sample decoding except the last row.

System variant	Student backbone	#T	Struct. SFT	<think>	GRPO	Permuted 8-vote	Acc. (%)
Preliminary validation of structured reasoning supervision							
Base model	MOSS-Audio	0	–	–	–	–	52.21
Single-teacher structured SFT	MOSS-Audio	1	✓	–	–	–	56.94
Progressive construction of the final system							
Three-run teacher annotation SFT	MOSS-Audio-8B-Thinking	3	✓	–	–	–	58.55
+ native thinking format	MOSS-Audio-8B-Thinking	3	✓	✓	–	–	58.62
+ group-relative post-training	MOSS-Audio-8B-Thinking	3	✓	✓	✓	–	58.99
+ choice-order self-consistency	MOSS-Audio-8B-Thinking	3	✓	✓	✓	✓	**62.79**

Note: #T denotes the maximum number of valid teacher annotations per question. The single-teacher system uses MOSS-Audio, whereas the remaining systems use MOSS-Audio-8B-Thinking. The final row applies choice-permuted eight-sample voting.

Table 2: Comparison with the official baseline systems on the development set.

System	Accuracy (%)
Official baselines	
Fun-Audio-Chat [9]	56.81
Kimi-Audio [10]	46.36
MiMo-Audio [11]	54.57
Qwen3-Omni [6]	<u>62.48</u>
Step-Audio 2 Mini [12]	50.53
Overall weighted average	54.15
Our system	
MOSS-Audio baseline [7]	52.21
Final system	**62.79**

Note: The random-guess accuracy is 25.46%. The underlined value denotes the strongest official baseline.

3. EXPERIMENTS

3.1. Datasets and models

We use the AudioMCQ-StrongAC-GeminiCoT [4, 1, 2, 3] training set provided for the challenge. Each training instance contains an audio recording, a multiple-choice question, its candidate answers, the ground-truth answer, and a Gemini-generated CoT annotation. As described in Section 2.2, Qwen3-Omni-30B-A3B-Instruct is employed only as a teacher model to transform the original free-form reasoning into structured audio evidence, option-level judgments, and final answers. We run the teacher generation process three times with different random seeds, resulting in 58,440 candidate annotations. After automatic validation, 58,020 annotations are retained for supervised fine-tuning.

The final student model is MOSS-Audio-8B-Thinking. It is first adapted with structured reasoning SFT and subsequently optimized using our GRPO-style group-relative objective. Qwen3-Omni is used only during offline annotation generation and is not required during either development-set evaluation or competition inference. For the preliminary experiment reported in Table 1, we additionally use the original MOSS-Audio model to verify the effectiveness of structured reasoning supervision. All reported accuracies are evaluated on the official DCASE 2026 Task 5 development set using top-1 accuracy.

3.2. Implementation details

For clarity, we summarize only the most important hyperparameters here; the complete configuration for teacher generation, SFT, GRPO-style post-training, and inference is provided in Table 3 near the end of the report. All stages use BF16 precision. Teacher annotations are generated in three independent runs with a temperature of 0.3 and top-p of 0.9. For structured SFT, we train MOSS-Audio-8B-Thinking for one epoch using LoRA with rank 16 and $\alpha = 32$, a learning rate of 5×10^{-5}, and a maximum sequence length of 6144. LoRA is applied only to the language-model layers, while the audio encoder and audio projector remain frozen.

The GRPO-style stage is initialized from the SFT model and uses groups of four sampled responses, a learning rate of 2×10^{-6}, and online random permutation of the candidate choices. The reward weights for answer correctness, option-level judgment, answer–judgment consistency, and format validity are 0.65, 0.20, 0.10, and 0.05, respectively. During final inference, we generate eight stochastic predictions per question with temperature 0.7 and top-p 0.9, map the predictions back to the original choice order, and aggregate them by majority voting.

3.3. Experimental Results

Table 2 compares our final system with the official development-set baselines. Among the official systems, Qwen3-Omni achieves the highest accuracy of 62.48%. Our final MOSS-Audio-8B-Thinking system reaches 62.79%, exceeding the strongest official baseline by 0.31 absolute percentage points. It also outperforms the official weighted-average baseline by 8.64 percentage points. Notably, Qwen3-Omni is used only for offline structured annotation generation in our pipeline; the submitted inference model is based on MOSS-Audio-8B-Thinking.

Table 1 summarizes the development-set results of the main system variants. The original MOSS-Audio model achieves an accuracy of 52.21%. Fine-tuning it with one structured teacher annotation per question improves the accuracy to 56.94%. This result provides an early validation that supervising the model with explicit audio evidence, option-level judgments, and final answers is more effective than directly using the pretrained model for answer prediction.

For the final system, we replace the student backbone with MOSS-Audio-8B-Thinking and use up to three independently sampled Qwen3-Omni annotations per question. This configuration reaches 58.55% with single-sample decoding. Since both the stu-

Table 3: Main configurations for teacher annotation, structured SFT, GRPO-style post-training, and final inference.

Configuration	Teacher	SFT	GRPO	Inference
Model and data				
Model	Qwen3-Omni 30B-A3B-Instruct	MOSS-Audio 8B-Thinking	SFT-initialized student	GRPO-trained student
Data / runs	AudioMCQ-StrongAC-GeminiCoT; 3 runs	58,020 structured annotations	Structured references	Official evaluation set
Optimization				
Precision / epochs	BF16 / –	BF16 / 1	BF16 / 1	BF16 / –
Learning rate	–	5×10^{-5}	2×10^{-6}	–
LoRA (r, α)	–	$(16, 32)$	Continued training	–
Batch / accumulation	32 / –	1 / 8	1 / 8	8 / –
Sampling and sequence settings				
Samples per question	1 per run	–	4	8
Temperature / top-p	0.3 / 0.9	–	0.7 / 0.9	0.7 / 0.9
Choice permutation	–	–	Online	Per vote
Maximum length (prompt / completion / total)	– / 1024 / –	– / – / 6144	2048 / 768 / 6144	– / 768 / –
Reward weights (Ans./Opt./Cons./Fmt.)	–	–	0.65 / 0.20 0.10 / 0.05	–

Note: LoRA is applied only to the language-model layers; the audio encoder and projector remain frozen. Final predictions are obtained by voting over eight choice-permuted generations.

dent backbone and the amount of teacher supervision are changed, this result should be interpreted as a system-level improvement rather than a strictly controlled ablation. Wrapping the structured rationale in the native `<think>` format results in 58.62%, indicating that the format alignment preserves the model's predictive performance while matching its original reasoning interface.

Starting from this SFT model, GRPO-style group-relative post-training further increases the single-sample accuracy to 58.99%. Finally, choice-order randomized eight-sample voting improves the result to 62.79%, yielding a gain of 3.80 absolute percentage points over single-sample decoding. The complete system therefore outperforms the original MOSS-Audio baseline by 10.58 absolute percentage points. These results suggest that structured reasoning supervision provides the main training-stage improvement, while multi-sample aggregation substantially improves prediction robustness at inference time.

4. CONCLUSION

We proposed a structured reasoning and post-training framework for MOSS-Audio-8B-Thinking, combining multi-sample teacher distillation, full-response SFT, GRPO-style optimization, and choice-order randomized voting. The complete system achieves 62.79% accuracy, improving the original MOSS-Audio baseline by 10.58 absolute percentage points. Notably, the resulting 8B inference model reaches performance comparable to, and slightly higher than, the official Qwen3-Omni-30B baseline, demonstrating that structured teacher supervision and robust decoding enable an 8B student backbone to approach the development-set performance of the official Qwen3-Omni baseline.

5. REFERENCES

[1] Z. Ma, Y. Ma, Y. Zhu, C. Yang, Y.-W. Chao, R. Xu, W. Chen, Y. Chen, Z. Chen, J. Cong, K. Li, K. Li, S. Li, X. Li, X. Li, Z. Lian, Y. Liang, M. Liu, Z. Niu, T. Wang, Y. Wang, Y. Wang, Y. Wu, G. Yang, J. Yu, R. Yuan, Z. Zheng, Z. Zhou, H. Zhu, W. Xue, E. Benetos, K. Yu, E.-S. Chng, and X. Chen, "MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix," 2025. [Online]. Available: https://arxiv.org/abs/2505.13032

[2] 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 *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[3] 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," 2026. [Online]. Available: https://arxiv.org/abs/2506.04779

[4] H. He, X. Du, R. Sun, Z. Dai, Y. Xiao, M. Yang, J. Zhou, X. Li, Z. Liu, Z. Liang, C. Wu, Q. He, T. Lee, X. Chen, W.-L. Zheng, W. Wang, M. Plumbley, J. Liu, and Q. Kong, "Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models," in *International Conference on Learning Representations (ICLR)*, 2026.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] T. F. Team, Q. Chen, L. Cheng, C. Deng, X. Li, J. Liu, C.-H. Tan, W. Wang, J. Xu, J. Ye, *et al.*, "Fun-audio-chat technical report," *arXiv preprint arXiv:2512.20156*, 2025.

[10] 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.

[11] 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.

[12] B. Wu, C. Yan, C. Hu, C. Yi, C. Feng, F. Tian, F. Shen, G. Yu, H. Zhang, J. Li, *et al.*, "Step-audio 2 technical report," *arXiv preprint arXiv:2507.16632*, 2025.