MIAQA SUBMISSION FOR DCASE 2025 CHALLENGE TASK 5: A REINFORCEMENT LEARNING DRIVEN AUDIO QUESTION ANSWERING METHOD

Technical Report

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Xingwei Sun, Tianzi Wang, Junbo Zhang, Jian Luan*

MiLM Plus, Xiaomi Corp., China
{ligang5, liujizhong1, dinkelheinrich}@xiaomi.com

ABSTRACT

This technical report presents an audio question answering (AQA) method submitted to DCASE 2025 Challenge Task 5. Recent studies have shown that reinforcement learning (RL) can enhance the audio reasoning capabilities of large audio language models (LALMs). Thus, we employ a RL strategy to optimize our AQA model. The MiAQA submission is based on our preliminary study [1][1]. We apply the group relative policy optimization (GRPO) algorithm to Qwen2.5-Omni-7B. The model directly generates responses after implicit reasoning, without relying on complex, explicit chain-of-thought (CoT). To enhance data diversity, the training data combines human-annotated datasets with weakly labeled datasets generated by large language models (LLMs). Using only a single model and 35k training samples, MiAQA achieves up to 78.0% accuracy on the DCASE 2025 AQA development set.

Index Terms— AQA, LALMs, RL, GRPO, data generation

1. INTRODUCTION

Audio question answering (AQA) [2] is a multimodal task that involves understanding and reasoning based on audio content to generate accurate responses to questions. AQA systems must comprehend diverse acoustic environments, integrate relevant external knowledge (e.g., facts or auditory information on marine mammals or daily sounds), and reason over sound events and context to generate accurate answers.

The latest breakthroughs in large language models (LLMs) have greatly enhanced their reasoning abilities, particularly in mathematics and coding. However, research on audio understanding and reasoning still lags behind. Although Large audio language models (LALMs) , such as Qwen2.5-Omni [3] and Audio Flamingo 2 [4], have been released one after another, they typically focus on general capabilities and lack task-specific optimization for audio understanding and reasoning. AQA can be considered an advanced technology [2] built upon automated audio captioning (AAC). Although researchers have achieved strong AAC performances [5, 6, 7, 8], AQA remains highly challenging, as it combines auditory and linguistic modalities, making it well-suited for evaluating complex reasoning. The AQA task requires the ability to extract meaningful insights from raw audio signals, infer implicit relationships, and provide contextually relevant answers. Exploring LALM-based AQA models presents a promising direction for further research.

In this technical report, we present our exploration of audio question answering, where the group relative policy optimization (GRPO) algorithm [9] is applied to Qwen2.5-Omni-7B [3], instead of Qwen2-Audio-7B-Instruct [10] used in the previous study [1]. The model generates responses directly through implicit reasoning, without relying on complex, explicit chain-of-thought (CoT). To enhance data diversity, the training data combines human-annotated datasets with weakly labeled datasets generated by LLMs. The data generation prompts are provided in the Appendix. MiAQA achieves up to 78.0% accuracy on the DCASE 2025 AQA development set using only a single model and 35k training samples.

2. METHOD

2.1. Large Audio Language Models

LALMs generally support two main functions: audio understanding and audio generation. As AQA is an audio reasoning task, our discussion focuses exclusively on the audio understanding capabilities of LALMs. We adopt Qwen2.5-Omni-7B [3], a state-of-the-art LALM, as the backbone of our AQA model. Since the Thinker of Qwen2.5-Omni-7B is trained via supervised fine-tuining, reinforcement learning (RL) is expected to further enhance its performance.

2.2. Group Relative Policy Optimization

We conduct the GRPO algorithm [9] along with a prompt template to enhance its reasoning capabilities. The prompt template is:

- $[Question]$. Please choose the answer from the following options: $[Options]$. Output the final answer in $< answer >$ $< /answer >$.

$[Question]$ and $[Options]$ denote the question and options given by the dataset. and For each question, the model is guided to generate a response within the $< answer >< /answer >$ tags, optimized by reinforcement learning. GRPO eliminates the need to train an additional value function approximation model in proximal policy optimization (PPO) [11]. GRPO uses the average reward of sampled response from the policy model as the baseline in computing the advantage. Specifically, given an input question q, a group of responses $\{o_1, o_2, \cdots, o_G\}$ is first sample, and their corresponding rewards corresponding rewards $\{r_1, r_2, \cdots, r_G\}$ are computed using the reward model. The advantage is subsequently computed as Equation (1).

$$\hat{A}_{i,t} = \widetilde{r}_i = \big(r_i - \mathrm{mean}(\mathbf{r})\big)/\mathrm{std}(\mathbf{r}) \qquad (1)$$

*Corresponding author.

[1] https://github.com/xiaomi-research/r1-aqa

The policy model is subsequently optimized by maximizing the Kullback-Leibler objective as Equation (2). where π_θ and π_{old} are the current and former policy, and ϵ and β are hyper-parameters introduced in PPO. Responses are evaluated by a rule-based reward function in terms of their format and correctness:

- If the answer is correct, the model obtains an accuracy reward of plus one;
- If the answer is given in $< answer >< /answer >$, the model obtains a format reward of plus one;
- In other cases, the model does not obtain any rewards;
- The final reward is the sum of accuracy and format rewards.

2.3. Data Generation

To increase the amount of high-quality training data, we use Qwen3-32B [12] to generate question-answer (QA) pairs from Clotho-AQA [13] and TACOS [14]. For Clotho-AQA, the simple answers, such as "yes" and "no", are rewritten into richer descriptions. In addition, each QA pair is expanded to include four answer options. For TACOS, the audio captions are transformed into QA pairs, each with four answer options. More details on data generation are provided in the Appendix.

3. EXPERIMENTS

3.1. Datasets

Several datasets are combined in specific proportions for training (detailed in the experimental results). If only a portion of a dataset is used, uniform sampling is applied. When multiple datasets are combined, they are randomly shuffled. The datasets are described as follows:

- **AVQA** [15]: Designed for audio-visual question answering by providing multimodal information in real-life video scenarios. A total of 38k[2] audio-text pairs from the training set are considered for training with "video" in the questions replaced by "audio".
- **Clotho-AQA** [13]: A dataset for AQA consisting of 1991 audio files, each ranging from 15 to 30 seconds in duration. A total of 7k generated QA pairs are available for training, instead of the original QA pairs.
- **TACOS** [14]: An audio captioning dataset containing 12k audio recordings and 47k temporally aligned captions. All 47k generated QA pairs (including those from the test set) are available for training.
- **DCASE 2025 AQA Dataset** [2]: The training set includes 8k samples, including bioacoustics QA [16], temporal soundscape QA, and complex QA [17], which can be used for training.

The development set of the DCASE 2025 AQA dataset is used for evaluating model performance. We select the submitted models based on the evaluation results of the development set.

3.2. Implementation Details

The models are trained using eight NVIDIA H800 GPUs, with each device processing a batch size of 1. The total number of training steps is set to 2000, with a learning rate of 1×10^{-6} and a temperature of 1.0. All hyperparameters are listed in Table 1. Model checkpoints are saved every 100 steps, and the best-performing checkpoint is selected for comparison and submission.

Table 1: Hyperparameters of the GRPO implementation.

Setting	**Value**
Batch Size per Device	1
Gradient Accumulation Steps	2
Training Steps	2000
Learning Rate	1×10^{-6}
Temperature	1.0
Maximum Response Length	512
Number of Responses per GRPO Step	8
Kullback-Leible Coefficient	0.04

3.3. Results

The metric is top-1 accuracy, the proportion of questions where the predicted answer exactly matches the ground truth.

Table 2 gives the comparison on the DCASE 2025 AQA development set, where Part 1, Part 2 and Part 3 denote the bioacoustics QA, temporal soundscapes QA and complex QA subsets. Experimental findings suggest that, for GRPO training, the quality of data matters more than the amount, and the ratio of data types is crucial. A comparison between "Not Submitted" and "Submission 1" shows that although AVQA was an important training dataset in our previous work [1], its impact is no longer significant. In fact, it may even have a negative effect when compared to high-quality generated datasets such as Clotho-AQA and TACOS. This highlights the effectiveness of high-quality generated data in training AQA models, in some cases even surpassing the benefits of manually annotated but simpler datasets (e.g., Clotho-AQA).

Despite the best performance being obtained with the DCASE training set alone, models trained with additional datasets are also submitted to reduce the risk of overfitting. In addition, we apply weight averaging strategy in both Submission 3 and Submission 4 to further ensure model robustness.

4. CONCLUSION

This technical report describes the MiAQA submission for the DCASE 2025 Challenge Task 5. Our method builds upon our preliminary study [1] and applies the GRPO algorithm to Qwen2.5-Omni-7B. The model generates responses directly through implicit reasoning. To enhance data diversity, the training data combines human-annotated datasets with weakly labeled datasets generated by Qwen3-32B. Using only a single model and 35k training samples,, MiAQA achieves up to 78.0% accuracy on the DCASE 2025 AQA development set.

5. REFERENCES

[1] 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.

[2] C.-H. H. Yang, S. Ghosh, Q. Wang, J. Kim, H. Hong, S. Kumar, G. Zhong, Z. Kong, S. Sakshi, V. Lokegaonkar,

[2] The AVQA training set contains 40k samples; 2k failed to download.

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|q)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t},\right.\right.\right.$$
$$\left.\left.\left.\text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})},1-\epsilon,1+\epsilon\right)\hat{A}_{i,t}\right]-\beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\}\right] \tag{2}$$

Table 2: Comparison on the DCASE 2025 AQA development set. The percentage sign in the scores has been omitted.

Model	Details / Training Samples	Part 1 Dev	Part 2 Dev	Part 3 Dev	Dev Total
Qwen2-Audio-7B	Direct Inference	30.0	39.2	49.6	45.0
Audio Flamingo 2	Direct Inference	53.9	31.7	49.5	45.7
Gemini 2.0 Flash	Direct Inference	42.0	46.3	56.6	52.5
Unsubmitted Version	AVQA 8k + Clotho-AQA 7k + TACOS 12k + DCASE 8k	64.7	58.8	79.1	72.8
Submission 1	Clotho-AQA 7k + TACOS 12k + DCASE 8k	68.8	60.1	79.9	74.0
Submission 2	DCASE 8k	75.9	67.0	82.4	78.0
Submission 3	Mean Weights of Unsubmitted Version and Submission 2	71.4	61.7	81.1	75.4
Submission 4	Mean Weights of Submission 1 and Submission 2	71.4	64.2	81.2	76.1

et al., “Multi-domain audio question answering toward acoustic content reasoning in the DCASE 2025 challenge,” *arXiv preprint arXiv:2505.07365*, 2025.

[3] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin, “Qwen2.5-Omni technical report,” *arXiv preprint arXiv:2503.20215*, 2025.

[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] J. Kim, M. Jeon, J. Jung, S. H. Woo, and J. Lee, “EnCLAP++: Analyzing the EnCLAP Framework for Optimizing Automated Audio Captioning Performance,” *arXiv preprint arXiv:2409.01201*, 2024.

[6] W. Chen, Z. Ma, X. Li, X. Xu, Y. Liang, Z. Zheng, K. Yu, and X. Chen, “SLAM-AAC: Enhancing Audio Captioning with Paraphrasing Augmentation and CLAP-Refine through LLMs,” in *Proc. ICASSP*, 2025, pp. 1–5.

[7] J. Liu, G. Li, J. Zhang, H. Dinkel, Y. Wang, Z. Yan, Y. Wang, and B. Wang, “Enhancing automated audio captioning via large language models with optimized audio encoding,” in *Proc. Interspeech*, 2024, pp. 1135–1139.

[8] J. Liu, G. Li, C. Liu, J. Zhang, H. Dinkel, Y. Wang, Z. Yan, Y. Wang, and B. Wang, “Leveraging CED encoder and large language models for automated audio captioning,” DCASE2024 Challenge, Tech. Rep. 42, May 2024.

[9] 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.

[10] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, “Qwen2-Audio technical report,” *arXiv preprint arXiv:2407.10759*, 2024.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.

[12] 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.

[13] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, “Clotho-AQA: A crowdsourced dataset for audio question answering,” in *Proc. EUSIPCO*, 2022, pp. 1140–1144.

[14] P. Primus, F. Schmid, and G. Widmer, “TACOS: Temporally-aligned audio captions for language-audio pretraining,” *arXiv preprint arXiv:2505.07609*, 2025.

[15] 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 *ACM International Conference on Multimedia*, 2022, p. 3480–3491.

[16] J. Kim, H. Yun, S. H. Woo, C.-H. H. Yang, and G. Kim, “Wow-bench: Evaluating fine-grained acoustic perception in audio-language models via marine mammal vocalizations,” 2025.

[17] 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 *Proc. ICLR*, 2025.

6. APPENDIX

The data generation prompt for Clotho-AQA [13] is shown in Table 3, where {INPUT_JSON_DATA} represents the metadata associated with Clotho-AQA, including the original question and answer, audio keywords, and the file name. Similarly, the data generation prompt for TACOS is presented in Table 4, where {INPUT_JSON_DATA} denotes the metadata associated with TACOS [14], which includes the strong caption, audio keywords, audio superclass, audio subclass, and file description.

~~~
I’m training an audio understanding model that needs many high-quality question-answer pairs. The model can only be input with the raw audio, questions and options.

## Your Task
Generate a QA pair with 3 wrong options based on the following audio metadata.

## Audio Metadata
{INPUT_JSON_DATA}

## Critical Instruction
1. The question must have exactly 4 options (A, B, C, D) with only 1 correct answer
2. Do not change the question in the audio metadata
3. Generate 3 wrong options based on the given question and answer
4. All options should be plausible but distinguishable
5. Distractors (incorrect options) should be reasonable, not obviously wrong
6. Refine all options (including the correct option), none of the options are allowed to be a single word

## Output Format
Please provide the QA pair in this JSON format:
```json
[
  {
    ”question”: ”Question text”,
    ”options”: {
      ”A”: ”Option A text”,
      ”B”: ”Option B text”,
      ”C”: ”Option C text”,
      ”D”: ”Option D text”
    },
    ”answer”: ”A|B|C|D”
  },
]
```

## Output Example
```json
[
  {
    ”question”: ”What is the main topic being discussed in this recording?”,
    ”options”: {
      ”A”: ”A family vacation”,
      ”B”: ”A school reunion”,
      ”C”: ”A wedding ceremony”,
      ”D”: ”A business conference”
    },
    ”answer”: ”D”
  },
]
```

Now, generate a QA pair.
~~~

Table 3: Data generation prompt for Clotho-AQA.

~~~~
I’m training an audio understanding model that needs many high-quality question-answer pairs. The model can only receive the raw audio
and questions as the input.

## Your Task
Generate a QA pair from the following audio metadata, especially based on ”caption”. The QA pair should have exactly 4 options with
only 1 correct answer.

## Audio Metadata
{INPUT_JSON_DATA}

## Critical Instruction
1. Question must have exactly 4 options (A, B, C, D) with only 1 correct answer
2. All options should be plausible but distinguishable
3. Distractors (incorrect options) should be reasonable, not obviously wrong
4. The question must focus on the content and information in the audio
5. The correct answer must be based on the caption in the audio metadata
6. AVOID querying about technical quality, recording environment, or potential applications
7. Focus on what people would genuinely want to know about the content
8. The QA pair should be written as if created by someone who ONLY LISTENED to the audio without seeing any metrics

## Output Format
Please provide the QA pair in this JSON format:
```json
[
  {
    ”question”: ”Question text”,
    ”options”: {
      ”A”: ”Option A text”,
      ”B”: ”Option B text”,
      ”C”: ”Option C text”,
      ”D”: ”Option D text”
    },
    ”answer”: ”A|B|C|D”
  },
]
```

## Output Example
```json
[
  {
    ”question”: ”What is the main topic being discussed in this recording?”,
    ”options”: {
      ”A”: ”A family vacation”,
      ”B”: ”A school reunion”,
      ”C”: ”A wedding ceremony”,
      ”D”: ”A business conference”
    },
    ”answer”: ”D”
  },
]
```

Now, generate a QA pair.
~~~~

Table 4: Data generation prompt for TACOS.