TRAINING-FREE INFERENCE-TIME EXPLORATION FOR AUDIO-DEPENDENT QUESTION ANSWERING

Technical Report

Zeyu Yin[1]*, Qi Cao*[1]*, Pingsong Deng*[1]*, Yizhou Tan*[1]*, Shengchen Li*[1]

[1] Xi'an Jiaotong-Liverpool University, School of Advanced Technology, Suzhou, China,
{qi.cao23, zeyu.yin22, pingsong.deng23, yizhou.tan22}@student.xjtlu.edu.cn, shengchen.li@xjtlu.edu.cn

ABSTRACT

This technical report presents our submission systems for DCASE 2026 Task 5. We focus on training-free inference-time strategies for audio-dependent question answering, where the model must answer multiple-choice questions based on the given audio rather than textual priors alone. Our main submitted system uses Qwen3-Omni with stable prompt-program majority voting. For each question, we run five complementary prompt variants, including option elimination, acoustic-focused reasoning, counterfactual verification, and self-refinement, and select the final answer by deterministic majority voting over normalized option-text predictions. On the development set, this system achieves 68.33% accuracy, showing that different prompt programs can recover complementary correct answers. We further analyze confidence-based branch selection and find that entropy- and margin-based selectors do not improve over majority voting. In addition, we explore audio-token attention enhancement on MOSS-Audio-8B-Thinking, which improves audio-grounded reasoning by increasing attention to audio tokens during decoding. Overall, our results suggest that inference-time exploration is a practical training-free direction for audio-dependent question answering, while better branch selection remains an important direction for future work.

***Index Terms*—** audio-dependent question answering, large audio-language models, test-time scaling

1. INTRODUCTION

Large Audio-Language Models (LALMs) have recently become an important direction for general audio understanding and multimodal reasoning. By connecting audio encoders with large language models, recent systems such as SALMONN [1], Qwen3-Omni [2], and MOSS-Audio [3] can process speech, environmental sounds, music, and natural language instructions within a unified generation framework. This capability makes LALMs suitable for a wide range of audio-centric tasks, including sound event detection [4], audio captioning, speech understanding, and audio question answering (AQA) [5]. Unlike conventional audio classification systems that map an input signal to a fixed taxonomy, LALMs can answer flexible natural-language questions and combine acoustic perception with linguistic reasoning.

Audio-dependent question answering remains challenging because the correct answer must be grounded in the audio rather than inferred from the question text alone. In multiple-choice settings, a model may exploit linguistic priors from the question and answer options, especially when the acoustic evidence is subtle or ambiguous. This issue is related to the cross-modal imbalance observed in LALMs, where the language model may attend more strongly to text tokens than to audio tokens during reasoning [6]. As a result, improving audio-grounded reasoning requires not only stronger language reasoning, but also more reliable use of acoustic evidence.

A common way to improve LALMs is to rely on training-based methods, such as large-scale audio-text pretraining, instruction tuning, supervised fine-tuning, or reinforcement learning. These approaches can improve instruction following and audio-grounded reasoning, as reflected in recent audio-language model systems [2, 3]. However, they usually require large annotated datasets, heavy GPU resources, and careful optimization. Such requirements make them less practical when the target benchmark is new, the size of task-specific data is limited, or the model parameters are frozen.

Another line of work improves model behavior at inference time. In language models, chain-of-thought prompting [7], zero-shot reasoning prompts [8], self-consistency [9], tree-of-thought search [10], and automatic prompt engineering [11] show that additional inference-time computation can improve reasoning reliability without modifying model parameters. These methods suggest that generating and aggregating multiple reasoning paths can be a practical alternative to expensive training, especially under limited-resource settings.

Motivated by this idea, we explore training-free inference-time strategies for DCASE 2026 Task 5. Our submitted system uses Qwen3-Omni with stable prompt-program majority voting. For each question, we run five complementary prompt variants and select the final answer by deterministic majority voting. This system does not require additional training data, supervised fine-tuning, reinforcement learning, or parameter updates. In addition, we analyze uncertainty-based selection and audio-token attention enhancement as auxiliary inference-time explorations. Overall, our work presents a simple training-free submission system and examines the potential and limitations of inference-time exploration for audio-dependent question answering.

2. METHOD

We study training-free inference-time strategies for audio-dependent question answering. The goal is to improve the reliability of large audio-language models without additional training data or parameter updates. One of our main submitted system uses Qwen3-Omni with stable prompt-program majority voting. In addition, we submit MOSS-Audio-based systems using audio-token attention enhancement as an auxiliary inference-time intervention.

2.1. Stable Prompt-Program Majority Voting

Our submitted system is based on prompt-program branching. For each audio clip a, question q, and answer option set $\mathcal{O} = \{o_1, o_2, \ldots, o_K\}$, we run the same Qwen3-Omni model with five fixed prompt programs. Each branch is designed to emphasize a different reasoning behavior while keeping the model parameters unchanged.

The five branches are:

- **Original prompt**: the standard multiple-choice prompt.
- **Option elimination**: the model compares each option with the audio evidence.
- **Acoustic expert**: the model focuses on acoustic cues such as timbre, rhythm, pitch, intensity, repetition, and source identity.
- **Counterfactual verification**: the model checks what should be heard for each option to be true.
- **Self-refinement**: the model verifies and revises its initial baseline answer.

Each branch is required to output only the exact option text rather than the option letter. Let the normalized output of branch b be $\hat{y}_b$. The final answer is selected by majority voting:

$$\hat{y} = \arg\max_{y \in \mathcal{O}} \sum_{b=1}^{B} I(\hat{y}_b = y), \tag{1}$$

where $B = 5$ is the number of branches. If multiple answers receive the same number of votes, ties are broken using a fixed branch order: original prompt, option elimination, acoustic expert, counterfactual verification, and self-refinement.

2.2. Confidence-Based Branch Selection

The oracle accuracy of multiple prompt branches can be much higher than simple majority voting if at least one branch predicts the correct answer. This motivates an additional question: whether we can select the most reliable branch automatically for each sample. To investigate this, we evaluate confidence-based branch selection using teacher-forced option likelihoods.

For each branch output, we compute the likelihood of each answer option under the model by teacher forcing the option text. Let $p(o_k \mid a, q)$ denote the normalized likelihood of option o_k. We then compute the entropy of the option distribution:

$$H = -\sum_{k=1}^{K} p(o_k \mid a, q) \log p(o_k \mid a, q). \tag{2}$$

A lower entropy indicates that the model assigns probability mass more confidently to one option. We also compute the margin between the two most likely options:

$$m = p(o_{(1)} \mid a, q) - p(o_{(2)} \mid a, q), \tag{3}$$

where $o_{(1)}$ and $o_{(2)}$ are the highest- and second-highest-likelihood options.

Based on these signals, we test several selection rules, including choosing the lowest-entropy branch, choosing the highest-margin branch, and weighting votes by entropy or margin. These methods are used only for analysis and are not in the final submitted system.

2.3. Audio Token Attention Enhancement

We also explore a separate inference-time intervention for MOSS-Audio-8B-Thinking. During autoregressive decoding, the model may underuse the audio input and rely mainly on the textual question. Inspired by MATA [12], we increase the attention assigned to audio tokens by adding a positive bias to their pre-softmax attention logits.

Given an audio input a and textual question q, the model converts them into a token sequence $\mathbf{x} = (x_1, \ldots, x_T)$. We identify the positions of audio tokens as

$$\mathcal{A} = \{j : x_j \in \mathcal{V}_{\text{audio}}\}, \tag{4}$$

where $\mathcal{V}_{\text{audio}}$ denotes the set of special audio tokens.

For decoder layer ℓ, the original attention logit is

$$A^{(\ell)}_{h,i,j} = \frac{\mathbf{q}^{(\ell)}_{h,i} \cdot \mathbf{k}^{(\ell)}_{h,j}}{\sqrt{d}} + M_{i,j}, \tag{5}$$

where h is the attention head, i is the query position, j is the key position, d is the head dimension, and $M_{i,j}$ is the causal or padding mask. We add an audio-token bias:

$$\widetilde{A}^{(\ell)}_{h,i,j} = A^{(\ell)}_{h,i,j} + B_{i,j}, \tag{6}$$

where

$$B_{i,j} = \begin{cases} \beta, & j \in \mathcal{A}, \\ 0, & \text{otherwise}. \end{cases} \tag{7}$$

The modified attention weights are then computed as

$$\alpha^{(\ell)}_{h,i,j} = \frac{\exp(\widetilde{A}^{(\ell)}_{h,i,j})}{\sum_{t=1}^{T} \exp(\widetilde{A}^{(\ell)}_{h,i,t})}. \tag{8}$$

Since the bias is added before softmax, it increases the unnormalized attention weight of audio tokens by a factor of $\exp(\beta)$. This encourages the model to rely more strongly on acoustic evidence during decoding. We test different bias strengths β and injection layers ℓ. We also evaluate a category-aware setting, where the bias value is selected according to question categories on the development set.

3. EXPERIMENTAL RESULTS

3.1. Prompt-Program Branching Results

We evaluate the Qwen3-Omni prompt-program branches on the DCASE 2026 Task 5 development set. Table 1 reports the performance of individual branches, majority-voting systems, and the oracle upper bound.

Table 1: Qwen3-Omni prompt-program results on development set.

System	Accuracy
Option elimination	67.89%
Acoustic expert	66.46%
Counterfactual verification	67.64%
Self-refinement	66.77%
Static 4-branch majority	68.51%
Submitted 5-branch majority	68.33%
5-branch oracle	75.61%

Among the individual prompt programs, option elimination performs best with 67.89%. The submitted five-branch majority system achieves 68.33% without any model training. The static four-branch majority obtains 68.51%, which is slightly higher on the development set.

The oracle accuracy of the five branches reaches 75.61%. This indicates that different prompt programs can recover complementary correct answers across different samples. However, the gap between majority voting and oracle accuracy shows that better branch selection could further improve the final prediction.

3.2. Confidence-Based Branch Selection

The result in Table 1 suggests that the branch pool often contains a correct answer even when majority voting fails. We therefore evaluate whether confidence signals can help select better branches. For each sample, we compute teacher-forced option likelihoods and derive two signals: entropy over the option distribution and the margin between the top-two options. Lower entropy and higher margin are treated as indicators of higher confidence.

Table 2: Branch selection ablation on the development set.

Method	Accuracy
5-branch majority voting	68.33%
4-branch majority voting	68.51%
Select branch with lowest option entropy	59.37%
Select branch with largest option margin	59.18%
Vote weighted by option entropy	57.87%
Vote weighted by option margin	57.93%
Development-tuned hybrid selection	68.33%

The results show that confidence-based branch selection does not improve over simple majority voting. Although the branch oracle is much higher than the final majority result, entropy and margin are not reliable indicators of audio-grounded correctness. In particular, the model may assign high likelihood to an answer that is linguistically plausible but not supported by the audio. Therefore, our final submitted system uses deterministic majority voting.

3.3. Audio Attention Enhancement Results

Finally, we evaluate audio-token attention enhancement on MOSS-Audio-8B-Thinking. Table 3 reports the results for different injection layers and bias strengths.

The best fixed audio-attention setting is obtained by applying a bias of 2.0 at layer 0, which improves the MOSS-Audio-8B-Thinking baseline from 62.60% to 64.97%. This configuration is used in our submitted system_2. In addition, the category-aware setting achieves an oracle accuracy of 66.46% on the development set, suggesting that different question categories may benefit from different audio-attention biases. Based on this observation, our submitted system_3 adopts a category-aware audio-attention strategy.

Overall, the results suggest that inference-time exploration can improve audio-dependent question answering without model training. Prompt-program majority voting provides a simple and robust submitted system, while the oracle and attention-enhancement results indicate that further gains may be obtained with better evidence-grounded selection strategies.

Table 3: Development set results of Audio Token Attention Enhancement on MOSS-Audio-8B-Thinking.

Layer	Bias	Accuracy
0	2.0	64.97%
1	2.0	64.84%
0	1.5	64.65%
0	4.0	64.59%
0	4.5	64.53%
3	2.0	64.41%
0	2.2	64.28%
0	1.8	64.22%
18	2.0	64.16%
0	1.0	63.91%
0	3.0	63.91%
0	2.5	63.78%
–	–	62.60%

4. CONCLUSION AND FUTURE WORK

In this technical report, we presented a training-free inference-time exploration framework for DCASE 2026 Task 5. Our submitted system uses Qwen3-Omni with stable prompt-program majority voting. By running five complementary prompt variants and aggregating their predictions with deterministic majority voting, the system improves answer robustness without using additional training data or updating model parameters. Experimental results show that different prompt programs can produce complementary correct answers across different samples, and the oracle accuracy of the branch pool is substantially higher than the majority-vote accuracy.

We also explored uncertainty-based selection and audio-token attention enhancement. Entropy- and margin-based selectors did not improve over majority voting, suggesting that token-level option likelihood is not well calibrated for audio-grounded correctness. Audio-token attention enhancement showed positive results on MOSS-Audio-8B-Thinking, indicating that increasing the model's reliance on acoustic evidence can be useful for audio-dependent question answering.

In future work, we will focus on improving branch selection rather than simply increasing the number of inference branches. In particular, evidence-grounded confidence estimation and more reliable audio-aware selection mechanisms may help convert the high oracle potential of inference-time exploration into actual performance gains.

5. REFERENCES

[1] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "SALMONN: Towards generic hearing abilities for large language models," in *International Conference on Learning Representations*, 2024.

[2] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, Y. Lv, Y. Wang, D. Guo, H. Wang, L. Ma, P. Zhang, X. Zhang, H. Hao, Z. Guo, B. Yang, B. Zhang, Z. Ma, X. Wei, S. Bai, K. Chen, X. Liu, P. Wang, M. Yang, D. Liu, X. Ren, B. Zheng, R. Men, F. Zhou, B. Yu, J. Yang, L. Yu, J. Zhou, and J. Lin, "Qwen3-Omni technical report," 2025.

[3] C. Yang, C. Yu, H. Chen, J. Zhu, J. Chen, K. Chen, W. Wang, Y. Wang, Y. Jiang, Y. Jiang, Z. Lin, Z. Chen, Z. Fei, C. Liu,

D. Yu, J. Zhan, K. Yu, K. Huang, L. Fan, M. Chen, Q. Cheng, R. Li, S. Li, S. Wang, X. Zhao, Y. Gao, Y. Gong, Y. Zhang, Z. Xu, and X. Qiu, "MOSS-Audio technical report," 2026.

[4] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[5] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-AQA: A crowdsourced dataset for audio question answering," in *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1140–1144.

[6] J. Wang, Z. Ma, Z. Luo, T. Wang, M. Ge, X. Wang, and L. Wang, "Pay more attention to audio: Mitigating imbalance of cross-modal attention in large audio language models," 2025.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, 2022.

[8] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Advances in Neural Information Processing Systems*, 2022.

[9] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *International Conference on Learning Representations*, 2023.

[10] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," in *Advances in Neural Information Processing Systems*, 2023.

[11] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," in *International Conference on Learning Representations*, 2023.

[12] J. Wang, Z. Ma, Z. Luo, T. Wang, M. Ge, X. Wang, and L. Wang, "Pay more attention to audio: Mitigating imbalance of cross-modal attention in large audio language models," 2025. [Online]. Available: https://arxiv.org/abs/2509.18816