Audio-Dependent Question Answering with Attention-Anchored Reinforcement Learning on MiMo-Audio

DCASE 2026 Challenge -- Task 5: Audio-Dependent Question Answering (ADQA) Technical Report

Hongjin Song
Beijing Institute of Technology (BIT), Beijing, China
3120252100@bit.edu.cn

Abstract

We present our submission to DCASE 2026 Task 5 (Audio-Dependent Question Answering). Our system is a single end-to-end large audio-language model (LALM) built on the open MiMo-Audio-7B base model and adapted with reinforcement learning (RL) on the AudioMCQ-StrongAC-GeminiCoT training set. The central motivation of ADQA is to penalize *textual hallucination*, where a model answers from language priors rather than genuine audio perception. To directly counter this failure mode, we introduce an **attention-anchoring reward** that, in addition to the usual answer-correctness and output-format rewards, explicitly rewards rollouts whose chain-of-thought attends to the audio token positions. The reward is gated on correctness so that the model cannot game the anchor by "pretending to listen". The policy is optimized with Group Relative Policy Optimization (GRPO). The model reasons step-by-step inside `<think>...</think>` and emits a single option letter inside `<answer>...</answer>`, which is mapped to the option text for submission. We report results on the official development set and submit predictions on the 3,000-question evaluation set.

1. Introduction

Large audio-language models routinely score well on audio multiple-choice benchmarks, yet a substantial fraction of that accuracy can survive even when the audio is removed or replaced, indicating that models exploit text-only shortcuts and common-sense priors. DCASE 2026 Task 5 (ADQA) is explicitly constructed to remove such shortcuts through a four-step Audio-Dependency Filtering (ADF) process, so that correct answers genuinely require listening.

Our design follows from this premise: if the benchmark demands real audio grounding, the training objective should *reward* audio grounding directly, not merely answer correctness. We therefore augment a standard reinforcement-learning-from-verifiable-rewards (RLVR) recipe with an attention-based "anchoring" signal computed from the model's own cross-attention onto the audio frames during its reasoning.

2. System Description

2.1 Base model

We use **MiMo-Audio-7B** (Xiaomi), an autoregressive audio-language model that consumes audio as discrete RVQ tokens interleaved with text in a single decoder stream, together with its companion audio tokenizer. No architectural changes are made; the system is end-to-end and operates directly on the raw waveform.

2.2 Prompting and output format

For each question the model receives the audio followed by a text prompt containing the question and the four options (A-D).

A system instruction asks the model to reason concisely inside `<think>...</think>` and then output exactly one letter inside `<answer>...</answer>`. The audio is provided as a single contiguous segment of RVQ tokens; no auxiliary time markers or timestamp annotations are injected, consistent with the training set, which contains no temporal labels. At submission time the predicted letter is mapped back to the full option text via direct string matching, with a layered fallback (strict `<answer>` tag -> letter inside the answer block -> first letter after `</think>` -> last stand-alone A-D letter) to recover an answer from malformed outputs.

2.3 Training data

We train on **AudioMCQ-StrongAC-GeminiCoT** (the official ADQA training set), comprising ~19.5k audio multiple-choice items spanning speech, sound, and music, each with four options, a ground-truth answer, and a Gemini-generated chain-of-thought reference. From this set we sample a balanced subset of prompts (weighted by question type) and generate K=4 rollouts per prompt from the base policy with temperature sampling, yielding 3,114 rollout groups used for RL.

2.4 Reward design

Each rollout receives a scalar reward combining three verifiable/diagnostic terms:

- **Answer correctness (RLVR):** 1 if the parsed option matches the ground-truth answer, else 0.
- **Format:** rewards the presence of a well-formed `<think>...</think>` reasoning block followed by an `<answer>X</answer>` tag, with a repetition penalty to discourage degenerate looping.
- **Attention anchoring:** a scalar in [0,1] measuring how much the response's reasoning attends to the audio. It is a weighted mixture of three components computed from the model's deep-layer cross-attention:

- **CON (concentration, weight 0.6):** the fraction of the response's attention mass that lands on audio key-positions versus all key-positions -- the primary, most robust term;
- **AR (activation ratio, weight 0.25):** the fraction of audio positions receiving above-threshold attention (breadth of audio coverage);
- **ENT (entropy floor, weight 0.15):** the normalized entropy of the response->audio attention, rewarded *above* a floor to prevent attention collapse.

The anchoring term is **gated on correctness** -- it is only added when the rollout's answer is correct -- so the model cannot inflate the reward by attending to audio while answering wrongly. This makes audio grounding a complement to, not a substitute for, getting the answer right.

2.5 Optimization

We optimize the policy with **Group Relative Policy Optimization (GRPO)**: within each group of K rollouts for the same prompt, advantages are computed by normalizing rewards across the group, and the policy is updated with a clipped objective plus a KL penalty to the reference (base) model. Training uses 8×GPU DeepSpeed ZeRO-3, bf16, gradient checkpointing, learning rate 1e-6 with cosine schedule and 5% warmup, clip e=0.2, KL coefficient 0.01. Audio is pre-encoded to RVQ codes once and cached for the collator.

3. Results

The system is evaluated with the same think-style inference pipeline used in training. We report accuracy on the official development set (DCASE2026-Task5-DevSet, 1,607 questions with ground-truth answers); the answer letter is extracted with the layered fallback described in §2.2 and matched against the gold option text.

System	Dev accuracy
MiMo-Audio-7B + attention-anchored GRPO (ours)	_to be filled_

Predictions on the 3,000-question evaluation set are provided in the accompanying submission file.

4. Conclusion

We adapt MiMo-Audio-7B to the ADQA task with a reinforcement-learning recipe whose reward explicitly anchors the model's reasoning to the audio signal, directly targeting the textual-hallucination failure mode that the benchmark is designed to expose. The anchoring reward is gated on answer correctness to avoid reward hacking, and is optimized with GRPO on the AudioMCQ-StrongAC-GeminiCoT training set without any timestamp supervision.

References

[1] DCASE 2026 Challenge, Task 5: Audio-Dependent Question Answering. https://dcase.community/challenge2026/

[2] ADQA-Bench / AudioMCQ-StrongAC-GeminiCoT. https://huggingface.co/datasets/Harland/AudioMCQ-StrongAC-GeminiCoT

[3] MiMo-Audio: Xiaomi MiMo-Audio Technical Report.

[4] Shao et al., DeepSeekMath: Group Relative Policy Optimization (GRPO), 2024.