ABILITY-AWARE MULTIMODAL RETRIEVAL-AUGMENTED GENERATION FOR DCASE 2026 TASK 5: AUDIO-DEPENDENT QUESTION ANSWERING

Technical Report

Yuelan Cheng[1]*, Jinzheng Zhao*[1]*, Rong Wan*[1]*, Peiwei Chang*[1]*, Yongqiang Chen*[1]*, Wenwu Wang*[1*]

[1] University of Surrey, Guildford, United Kingdom
{yuelan.cheng, j.zhao, r.wan, p.chang, yongqiang.chen, w.wang}@surrey.ac.uk

ABSTRACT

Audio-Dependent Question Answering (ADQA) requires models to answer questions by grounding their decisions on acoustic evidence. In this report, we present an ability-aware multimodal retrieval-augmented generation (RAG) system for DCASE 2026 Task 5. The system uses fine-grained labels for indicating reasoning ability to decide when retrieval should be applied, and retrieves audio-text question-answer examples as multimodal in-context demonstrations. This allows the model to benefit from relevant external examples while avoiding misleading retrieval for question types where retrieval is less reliable. Experiments on ADQA-Bench show that ability-aware multimodal RAG improves accuracy from 63.52% to 65.07% for MOSS-Audio and from 62.48% to 64.47% for Qwen3-Omni. We also evaluate confidence-aware voting for low-reliability labels, but it does not improve overall performance, slightly decreasing accuracy from 65.07% to 65.01%. These results indicate that ability-aware audio-text retrieval is effective for ADQA, whereas the limited benefit of confidence-aware voting suggests that the remaining errors are not sufficiently mitigated by simple sampling-based ensembling under the current setting.

***Index Terms*—** Audio-Dependent Question Answering, Audio-Text Retrieval, Retrieval-Augmented Generation, Multimodal In-Context Learning, MOSS-Audio

1. INTRODUCTION

Audio-Dependent Question Answering (ADQA) is a challenging multimodal reasoning task that requires models to answer natural language questions by grounding their decisions on acoustic evidence. Unlike conventional Audio Question Answering (AQA), where some questions can be answered from textual priors in the question and answer choices, ADQA emphasizes genuine audio comprehension by requiring the correct answer to depend on the input audio. This task covers a wide range of audio understanding abilities, including speech content, sound events, music, speaker characteristics, and prosodic cues.

Recent large audio-language models [1, 2], such as MOSS-Audio [3], have shown promising capabilities in following audio-grounded instructions and generating natural language responses. Nevertheless, ADQA remains challenging not only because it requires acoustic grounding, but also because different questions rely on different types of audio evidence. For example, speech-content questions require lexical understanding, sound-event questions require event-level acoustic perception, music questions may depend on timbre, rhythm, or mood, while speaker-related questions often rely on fine-grained prosodic or phonetic cues. This heterogeneity makes a uniform inference strategy suboptimal, as the same prompting or reasoning procedure may not be equally suitable for all question types.

Retrieval-Augmented Generation (RAG) adapts inference by conditioning the model on retrieved external context rather than relying solely on parametric knowledge [4]. Recent multimodal RAG studies have extended this idea beyond text by retrieving visual or multimodal evidence for generation and question answering [5, 6, 7]. Emerging speech- and audio-centered RAG methods have also explored direct speech retrieval for spoken question answering and text-audio hybrid retrieval for spoken dialogue modeling [8, 9]. However, these methods mainly focus on retrieving external evidence, speech passages, or audio-text knowledge, while the use of retrieved audio-question examples as multimodal in-context demonstrations for ADQA remains underexplored. Since textually similar questions may rely on different acoustic evidence, retrieved examples can introduce irrelevant reasoning patterns. This motivates our reasoning ability-aware RAG strategy, which applies retrieval according to the audio reasoning ability required by each question.

In ADQA, retrieved audio-question pairs can serve as multimodal in-context demonstrations, helping the model follow reasoning patterns associated with a particular type of audio understanding. However, retrieval is not inherently reliable for all ADQA questions. Since textual similarity between questions does not necessarily imply similarity in the underlying acoustic evidence, retrieved examples may share surface-level wording with the target question while requiring different listening cues. Such acoustically mismatched examples can bias the model toward irrelevant reasoning patterns and degrade prediction quality. This motivates an ability-aware retrieval strategy, where RAG is enabled and reference examples are selected according to the audio reasoning ability required by each question.

In this technical report, we describe our MOSS-Audio-8B-based system submitted to DCASE 2026 Task 5. The system adopts an ability-aware RAG framework, where each question is first assigned to a fine-grained reasoning ability label and retrieval is then controlled according to the predicted label. When retrieval is enabled, retrieved audio-question examples are included in the prompt as multimodal in-context demonstrations; otherwise, the model performs direct audio-grounded inference without retrieved context. We also investigate a confidence-aware voting strategy for labels with lower reliability. However, our ablation results show that vot-

*Thanks to ABC agency for funding.

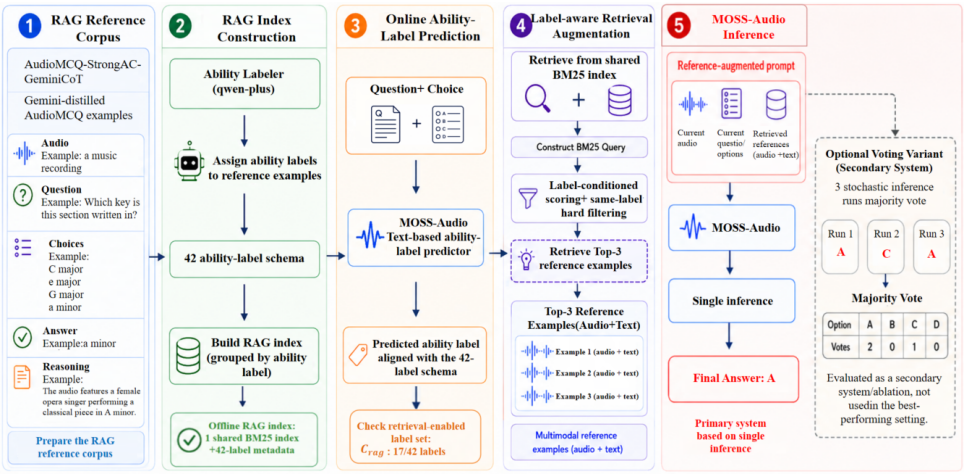

Figure 1: System overview of the proposed reasoning-ability aware RAG framework.

ing does not improve the overall accuracy, suggesting that repeated stochastic inference alone provides limited benefit under the current setting. These findings indicate that ability-aware retrieval is useful for selecting relevant in-context examples in ADQA.

2. SYSTEM OVERVIEW

We use MOSS-Audio [3] as the backbone for audio question answering and combine it with ability-aware retrieval. Each sample is assigned a fine-grained reasoning ability label based on its question and candidate options, such as accent region, sound event sequence, music instrument identification, or prosodic analysis. This label determines whether retrieval is applied and how reference examples are selected from a shared labeled lexical retrieval index based on the Okapi BM25 ranking function [10].

As shown in Fig. 1, the system consists of two stages: offline index construction and online inference. In the offline stage, a RAG corpus, including training examples and Gemini-distilled AudioMCQ examples, is annotated with fine-grained reasoning-ability labels and stored in a shared BM25 retrieval index. Each retrieval entry is associated with a multimodal reference example consisting of an audio clip and its textual annotations, including the question, candidate options, ground-truth answer, and reasoning trace. In the online stage, the reasoning ability label of each test sample is predicted using only the question and candidate options. When retrieval is enabled, relevant reference examples with the same labels are retrieved and incorporated into the prompt as in-context demonstrations, which, together with the test audio, question, and candidate options, are used for final answer prediction by MOSS-Audio.

3. METHODOLOGY

This section details the two stages of the proposed RAG framework based on reasoning ability. The offline stage constructs a shared labeled BM25 retrieval index from the RAG corpus, while the online stage uses the predicted label of reasoning ability for each test sample to perform ability-aware label-conditioned retrieval. The retrieved multimodal examples are then used as in-context references to guide MOSS-Audio answer prediction. We also evaluate a confidence-aware voting strategy as an additional inference variant.

3.1. Reasoning Ability-Aware RAG Index Construction

We first construct a retrieval index to provide external reference examples for MOSS-Audio. The RAG corpus consists of training examples and Gemini distilled AudioMCQ examples. Each example contains an audio clip, a question, candidate choices, the correct answer, and a reasoning trace. Since different questions require different audio understanding abilities, directly retrieving from the whole corpus may introduce irrelevant demonstrations. Therefore, we associate each example with a fine-grained reasoning ability label and use this label to guide retrieval during online inference.

Specifically, we used MOSS-Audio [3] as an LLM-based labeler to assign an ability label to each example. The labeler takes only the question and candidate options as input and outputs one of the 42 predefined reasoning ability labels. The ground-truth answer and reasoning trace are not used for ability label prediction; instead, they are stored as part of the reference example retrieved.

Rather than constructing separate BM25 indexes for different ability categories, we built a single shared BM25 index $\mathcal{I}$ over all labeled examples. Each example is converted into a textual document:

$$D_i = [\mathrm{dom}_i;\ \ell_i;\ Q_i;\ O_i;\ A_i;\ R_i], \tag{1}$$

where Q_i, O_i, A_i, and R_i denote the question, candidate options, correct answer, and reasoning text, respectively. The terms ℓ_i and dom_i denote the ability label and ability domain assigned to example i. Including the label and domain tokens in D_i allows the BM25 score to reflect lexical overlap between the predicted ability information and indexed examples.

The complete multimodal example is stored as:

$$E_i = (x_i^a, Q_i, O_i, A_i, R_i), \tag{2}$$

where x_i^a denotes the audio clip. BM25 is only applied to the textual document D_i, while the corresponding multimodal example E_i is used as the reference example retrieved during MOSS-Audio inference.

3.2. Ability-Aware Label-Conditioned BM25 Retrieval

During online inference, given a test audio clip, a question, and candidate options, the system first predicts the ability label $\hat{c}$ of the test sample using only the question and candidate options. Retrieval is not applied uniformly to all samples. Instead, the predicted label is checked against a label-level retrieval policy to determine whether retrieval should be enabled. If retrieval is disabled for the predicted label, MOSS-Audio directly answers the test question without retrieved examples.

Let $\mathcal{C}$ denote the full set of 42 ability labels. The retrieval system uses a single shared labeled BM25 index rather than separate indexes for individual labels. Retrieval is enabled only for a subset of 17 labels, denoted as $\mathcal{C}_{\text{rag}} \subset \mathcal{C}$, including `music_timbre_technique`, `prosody_intonation`, and `music_tempo_bpm`. If the predicted label $\hat{c}$ belongs to $\mathcal{C}_{\text{rag}}$, RAG is enabled; otherwise, MOSS-Audio directly predicts the answer without retrieved references.

For a retrieval-enabled sample, the BM25 query is constructed from the predicted ability information, the test question, and its candidate options:

$$q = [\text{dom}_{\text{test}};\ \hat{c};\ Q_{\text{test}};\ O_{\text{test}}], \tag{3}$$

where Q_{test} and O_{test} denote the test question and candidate options, and dom_{test} denotes the predicted ability domain.

For a query q and a document $D_i \in \mathcal{I}$, the BM25 score is computed over the shared index as:

$$\text{BM25}(q, D_i) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t, D_i)(k_1 + 1)}{f(t, D_i) + k_1 \left(1 - b + b \frac{|D_i|}{\text{avgdl}}\right)}, \tag{4}$$

where $f(t, D_i)$ is the frequency of term t in document D_i, $|D_i|$ is the document length, avgdl is the average document length over the shared BM25 index $\mathcal{I}$, and k_1 and b are BM25 hyperparameters.

Let ℓ_i denote the ability label of document D_i. To bias retrieval toward examples requiring the same audio-understanding ability, we apply label-conditioned re-scoring:

$$S(q, D_i) = \text{BM25}(q, D_i)\ + \lambda \cdot I[\ell_i = \hat{c}]\ - \mu \cdot I[\ell_i \neq \hat{c}], \tag{5}$$

where $I[\cdot]$ is the indicator function, and λ and μ control the bonus and penalty for label consistency. Documents whose audio path is identical to that of the test sample are excluded to prevent self-retrieval.

We first retain the top-K' candidates according to $S(q, D_i)$, and then discard candidates whose ability label does not match $\hat{c}$. Among the remaining same-label candidates, we rank the examples by $S(q, D_i)$ and select the K highest-scoring examples as the final reference set. In our implementation, we use $K' = 10$ and $K = 3$. Each $E_{(j)}$ is the multimodal reference example corresponding to a retrieved document. This procedure restricts the final reference set to examples requiring similar reasoning abilities while still using BM25 lexical similarity to rank candidates.

3.3. Reference-Augmented Answer Prediction

After retrieval, the selected reference examples are incorporated into the prompt as in-context demonstrations. Each retrieved example contains an audio clip together with its question, candidate options, answer, and reasoning trace. The test audio, question, and candidate options are then appended after the retrieved references, and MOSS-Audio is asked to predict the final answer. If retrieval is disabled by the label-level policy, MOSS-Audio directly answers the question using only the test audio, question, and candidate options. The model prediction is finally mapped to one of the candidate option IDs.

3.4. Confidence-Aware Voting Inference

In addition to the single-pass setting, we evaluate a confidence-aware voting strategy to examine whether repeated inference improves prediction stability for difficult ability labels. Each predicted ability label is checked against a historical-accuracy-based rule. Labels with relatively high historical accuracy use single-pass inference, while labels with lower historical accuracy are assigned to voting mode.

For samples assigned to voting mode, MOSS-Audio performs three independent inference runs with the same input. The final answer is selected by majority voting over the predicted option IDs. If all three runs produce different option IDs, the prediction from the first run is used as a deterministic fallback. This voting strategy is treated as an evaluated inference variant rather than the default setting of the proposed system.

4. EXPERIMENTS

4.1. Dataset

AudioMCQ-StrongAC-GeminiCoT. We construct the RAG index using AudioMCQ-StrongAC-GeminiCoT [11], the official training set provided for DCASE 2026 Task 5. It contains 19k audio-dependent multiple-choice QA samples selected by the StrongAC split, where answers rely strongly on acoustic cues. The GeminiCoT annotations provide audio-grounded reasoning annotations, which we use to construct retrieval documents and multimodal in-context examples.

AudioMCQ. We further use AudioMCQ [11] as an auxiliary source for selected ability labels. AudioMCQ is a large-scale audio multiple-choice question answering dataset containing more than 570k samples. Instead of using the full dataset, we only select samples related to several underrepresented or difficult labels, including age group, music mood, and syllable-related questions. For these selected samples, we use Gemini 3.1 Pro to generate distilled reasoning annotations. The resulting supplementary data are added to the retrieval pool to provide additional multimodal demonstrations for these specific ability categories.

4.2. Experimental Setup

MOSS-Audio-8B-Thinking is used as the primary backbone model, and runs on a single NVIDIA RTX 3090 GPU (24GB) without LoRA fine-tuning. The audio inputs are resampled to 16kHz mono. Decoding is performed with sampling (temperature = 0.6, top-p = 0.9, top-k = 20, max = 1024 new tokens) to enable diversity across voting rounds. For RAG, the top-3 examples (from a candidate pool of 10) are retrieved via same-label BM25 retrieval. For voting, labels with historical accuracy below 60% trigger 3 sampled inference rounds aggregated by majority vote; other labels use single-pass inference. To further validate the effectiveness of the ability-aware RAG mechanism across different backbones, single-pass (non-voting) inference experiments are additionally conducted with Qwen3-Omni [2] on a single NVIDIA A100 GPU, applying the same ability-aware RAG strategy.

Table 1: Results of ability-aware RAG on different backbone models.

Model	RAG	Accuracy	Retrieval Used
MOSS-Audio	No	63.52%	–
MOSS-Audio	Yes	65.07%	345
Qwen3-Omni	No	62.48%	–
Qwen3-Omni	Yes	64.47%	791

5. RESULTS

5.1. Results of Ability-Aware RAG

Table 1 reports the effect of ability-aware RAG on different backbone models. For the main MOSS-Audio system, ability-aware RAG improves the accuracy from 63.52% to 65.07%, yielding a gain of 1.55 percentage points. This suggests that ability-aware retrieval provides useful in-context references for audio question answering. We also evaluate Qwen3-Omni as an additional backbone to examine whether the retrieval strategy generalizes beyond MOSS-Audio. Ability-aware RAG improves the Qwen3-Omni accuracy from 62.48% to 64.47%, corresponding to a gain of 1.99 percentage points. These results suggest that the proposed retrieval strategy is not specific to a single backbone model.

5.2. Ablation Study on Confidence-Aware Voting

Table 2 compares the single-pass system (MAR) and the voting-augmented system (MARS) based on MOSS-Audio on the same 1606 development samples, using identical ability-aware RAG settings. The two systems achieve nearly identical overall accuracy, with voting slightly decreasing accuracy from 65.07% to 65.01%. Voting changes 139 predictions in total: 69 originally incorrect samples are corrected, while 70 originally correct samples become incorrect, resulting in a net change of -1 sample.

Table 2: Ablation study on confidence-aware voting.

System	Accuracy (%)	N
MAR	65.07	1606
MARS	65.01	1606

To better understand this result, we further analyze the voting-triggered samples. As shown in Table 3, the non-voting subset ($n = 898$), corresponding to higher-accuracy labels, achieves 73.2% accuracy. In contrast, the voting-triggered subset ($n = 708$), corresponding to lower- accuracy labels, achieves only 54.7%. This confirms that the historical- accuracy threshold successfully identifies harder categories. However, within the voting-triggered subset, unanimous predictions ($n = 566$) are correct for only 58.8% of the time, and split votes (2:1, $n = 142$) are correct for only 38.0%. This suggests that repeated inference often reinforces systematic errors rather than correcting random mistakes.

At the label level, voting improves several categories, including `prosody_pause` ($+7.9\%$, $n = 63$), `syllable_phonology` ($+7.7\%$, $n = 26$), and `music_instrument` ($+5.3\%$, $n = 57$). However, it also decreases performance on `music_genre` (-6.7%, $n = 30$) and `music_rhythm` (-4.5%, $n = 22$). These mixed label-level results explain why voting does not improve the overall benchmark accuracy. Overall, the main performance gain comes from ability-aware RAG, while confidence-aware voting provides only category-dependent and inconsistent benefits.

Table 3: Accuracy breakdown of the voting behavior.

Subset	n	Accuracy (%)
Non-voting labels	898	73.2
Voting labels	708	54.7
Unanimous votes	566	58.8
Split votes	142	38.0

6. CONCLUSION

This technical report presents a ability-aware multimodal RAG system for DCASE 2026 Task 5: Audio-Dependent Question Answering. The proposed system uses fine-grained reasoning ability labels to determine when retrieval should be applied and to select ability-matched reference examples from a shared labeled BM25 index. Experimental results show that ability-aware RAG improves overall accuracy for both evaluated backbone models, indicating the usefulness of ability-aware retrieval for ADQA. In contrast, confidence-aware voting does not lead to further improvement, suggesting that repeated stochastic inference provides limited benefit under the current setting. Future work will focus on improving ability-label prediction, refining the retrieval policy, and developing more acoustically aligned retrieval strategies for fine-grained ADQA.

7. REFERENCES

[1] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[2] 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.

[3] 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.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[5] M. Yasunaga, A. Aghajanyan, W. Shi, R. James, J. Leskovec, P. Liang, M. Lewis, L. Zettlemoyer, and W.-t. Yih, "Retrieval-augmented multimodal language modeling," *arXiv preprint arXiv:2211.12561*, 2022.

[6] W. Chen, H. Hu, X. Chen, P. Verga, and W. Cohen, "Murag: Multimodal retrieval-augmented generator for open question answering over images and text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5558–5570.

[7] M. M. Abootorabi, A. Zobeiri, M. Dehghani, M. Mohammadkhani, B. Mohammadi, O. Ghahroodi, M. S. Baghshah, and E. Asgari, "Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation," *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16 776–16 809, 2025.

[8] K. Mundnich, A. Lapastora, E. Soltanmohammadi, S. Ronanki, K. Han, *et al.*, "Speech retrieval-augmented generation without automatic speech recognition," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[9] Y. Chen, S. Ji, H. Wang, Z. Wang, S. Chen, J. He, J. Xu, and Z. Zhao, "Wavrag: Audio-integrated retrieval augmented generation for spoken dialogue models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 12 505–12 523.

[10] S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009, vol. 4.

[11] 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," *arXiv preprint arXiv:2509.21060*, 2025.