QAM-DETR SYSTEM FOR DCASE 2026 TASK 6: QUALITY-AWARE MAMBA DETR FOR QUERY-BASED AUDIO MOMENT RETRIEVAL

Technical Report

JeongRae Kim[1]*, Ho jun Jung* [1]*, Yewon Park*[2]*, Changwon Lim*[1,2*]

[1] Department of Statistics and Data Science, Chung-Ang University, Seoul, Korea
[2] Department of Applied Statistics, Chung-Ang University, Seoul, Korea
{kjk632, winner947, mico02, clim}@cau.ac.kr

ABSTRACT

This technical report describes our system for DCASE 2026 Challenge Task 6: Audio Moment Retrieval from Long Audio. The task aims to localize the temporal segment in a long audio recording that matches a given natural-language query. Our system builds on a DETR-style moment localization framework using precomputed pretrained audio and text features. LAION-CLAP is used as the primary audio-language representation, while MS-CLAP, WavLM, and RoBERTa features are selectively integrated to provide complementary acoustic and linguistic information. To improve query-conditioned localization, we use 3-way cross-attention-based cross-modal fusion, a BiMamba-based temporal encoder, lightweight multi-scale temporal fusion, and quality-aware candidate ranking. For selected final submissions, a frozen audio-language LLM verifier is further applied as a post-processing re-ranker without modifying predicted temporal boundaries. Experiments on the CASTELLA development splits show that the proposed components improve the DETR-style baseline, and the final ensemble systems further enhance temporal localization performance.

Index Terms— Audio moment retrieval, temporal localization, audio-text cross-attention, BiMamba

1. INTRODUCTION

Audio moment retrieval aims to localize the temporal segment in a long audio recording that corresponds to a given natural-language query. Unlike conventional audio-text retrieval, which mainly retrieves or ranks entire audio clips, this task requires predicting precise start and end timestamps of the query-relevant event. Therefore, an effective system should capture audio-text semantics, model long-range temporal context, and rank candidate temporal windows reliably.

DCASE 2026 Challenge Task 6 addresses this problem using long audio recordings paired with natural-language queries [1]. The task is challenging because target events may occupy only a short portion of a long recording, and queries may describe high-level semantic concepts rather than simple acoustic patterns. Since the evaluation emphasizes accurate temporal localization, especially at high IoU thresholds, both boundary prediction and confidence estimation are important.

Our system follows a DETR-style moment localization framework. We use LAION-CLAP [2] as the primary audio-language representation and selectively integrate MS-CLAP, WavLM, and RoBERTa features to exploit complementary acoustic and linguistic information. To improve query-conditioned localization, we introduce a 3-way cross-attention-based cross-modal fusion module, replace the Transformer encoder with a BiMamba-based temporal encoder, and apply lightweight multi-scale temporal fusion. We also add a quality prediction branch for candidate ranking, and for selected final submissions, use a frozen audio-language LLM verifier only as a post-processing re-ranker without modifying temporal boundaries or updating model parameters.

*Corresponding author

2. METHOD

Our system is built upon the DETR-style audio moment retrieval baseline provided for DCASE 2026 Task 6 [1]. While we largely retain the decoder structure of the baseline, we modify several key components of the model, including the input representation, audio-text interaction module, temporal encoder, training strategy, and inference ranking scheme. The overall architecture of the proposed system is illustrated in Fig. 1. The main modifications are summarized as follows.

2.1. Multi-feature cross-modal fusion

To avoid relying on a single audio-text representation, we use multiple pretrained features from both modalities. Each feature stream is first projected into a shared hidden dimension using an individual projection layer. LAION-CLAP is used as the primary audio-text representation, while auxiliary audio and text features are selectively integrated through feature-level learnable scalar gates. Each auxiliary feature stream has one sigmoid-normalized scalar gate that is shared across all temporal tokens, rather than being estimated separately for each token. This lightweight gating scheme allows the model to adjust the contribution of each auxiliary representation while preserving the main LAION-CLAP representation. The same gating strategy is applied separately to audio and text feature streams.

After gated feature integration, we apply a 3-way cross-attention-based cross-modal fusion module before the temporal encoder. Instead of simply concatenating audio and text features, the module progressively exchanges information among audio features, text query features, and cross-modal representations. Specifically, the audio feature first attends to the text query to emphasize query-relevant temporal regions. The text feature is then refined by attending to the audio sequence, and the resulting cross-modal rep-

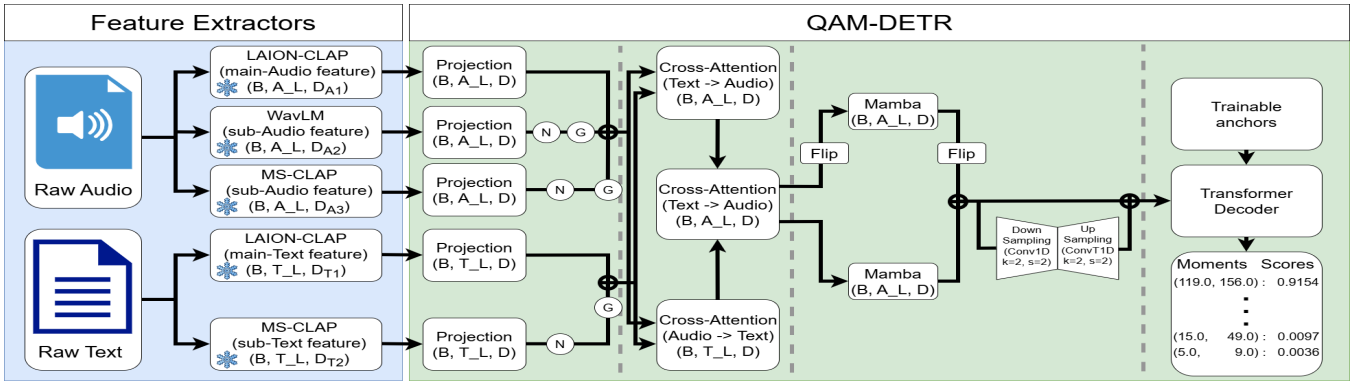

Figure 1: Overall architecture of QAM-DETR. Pretrained audio and text features are projected, gated, fused through 3-way cross-attention, encoded by BiMamba with multi-scale temporal fusion, and decoded into candidate moments with ranking scores.

resentation is used to update the audio sequence before temporal encoding. This fusion module strengthens audio-text interaction for moment localization without introducing an additional explicit alignment loss.

Let x_{main} denote the projected LAION-CLAP feature and x_j denote the j-th projected auxiliary feature. The fused representation is computed as

$$x_{\text{fused}} = x_{\text{main}} + \sum_{j=1}^{J} \alpha_j x_j, \tag{1}$$

where $\alpha_j = \sigma(g_j)$ is the learnable scalar gate for the j-th auxiliary feature stream. Here, g_j is a trainable parameter and $\sigma(\cdot)$ denotes the sigmoid function. We apply this gating scheme separately for audio and text feature streams.

For audio-text interaction, we introduce a 3-way cross-modal fusion module before the temporal encoder. This design is inspired by prior multimodal alignment work that uses three cross-attention blocks to progressively exchange semantic information across modalities [3]. Instead of simply concatenating audio and text features, our module applies multiple cross-attention blocks to exchange information among audio features, text query features, and cross-modal representations.

Specifically, the audio feature first attends to the text query to enhance query-relevant temporal regions. Then, the text feature is further refined by attending to the temporal context of the audio sequence. Finally, the enhanced cross-modal representation is used to update the audio sequence and is passed to the temporal encoder. This 3-way cross-modal fusion strengthens cross-modal interaction for moment localization without introducing an additional explicit alignment loss.

2.2. BiMamba encoder with multi-scale temporal fusion

To model temporal dependencies in long audio sequences, we replace the Transformer encoder in the baseline with a BiMamba-based temporal encoder. This design is motivated by Mamba-based moment retrieval models that combine cross-modal alignment with efficient long-range sequence modeling [4]. In audio moment retrieval, the target event may be short, but its semantic meaning and boundaries often depend on surrounding acoustic context. Therefore, the encoder should use information from both preceding and following segments.

The BiMamba encoder consists of stacked bidirectional Mamba blocks. Each block has two temporal branches: one processes the sequence in the original temporal order, and the other processes the reversed sequence. The backward output is reversed back to the original order, concatenated with the forward output, and projected to the model hidden dimension using a linear layer. A residual connection and layer normalization are then applied to stabilize training. This bidirectional structure allows each temporal frame to incorporate both past and future context while keeping the sequential modeling cost lower than full self-attention over long audio sequences.

After BiMamba encoding, we apply a lightweight multi-scale temporal fusion module to handle events of different durations. The module combines the original high-resolution temporal representation with downsampled contextual representations and then upsamples them back to the original temporal resolution. The fused representation is added through residual connections, allowing the model to capture both short local events and longer temporal patterns while preserving frame-level temporal resolution for boundary prediction.

2.3. Quality-aware training and re-ranking

We retain the DETR-style Hungarian matching framework of the baseline and add a quality prediction branch to estimate the localization reliability of each predicted window. After matching decoder predictions with ground-truth moments, the model is trained with foreground/background classification, L1 span regression, temporal GIoU, saliency, and quality regression losses. The quality target is defined as the temporal IoU between each matched prediction and its ground-truth window, and the IoU target is detached from the computation graph. This allows the quality branch to learn localization reliability without directly changing the span regression target.

During inference, the predicted quality score is combined with the foreground probability to obtain a quality-aware decoder score:

$$s_{\text{dec}} = p^{\text{fg}} \cdot \hat{q}. \tag{2}$$

Candidate windows are first ranked according to s_{dec}. For Submit 2, Submit 3, and Submit 4, we further apply a frozen audio-language LLM verifier to re-rank the top-K candidates, where $K = 10$. Submit 1 does not use LLM-based post-processing and relies only on decoder-based ensemble confidence.

Configuration	val R1@0.5	val R1@0.7	val mAP	val mAP@0.5	val mAP@0.75	test R1@0.5	test R1@0.7	test mAP	test mAP@0.5	test mAP@0.75
DETR-style baseline with LAION-CLAP	39.32	22.22	17.89	35.69	15.80	33.01	16.42	13.98	28.94	12.02
+ BiMamba encoder	46.19	29.20	23.10	42.11	21.80	38.66	22.17	18.32	34.33	16.96
+ Quality regression loss	50.46	34.09	26.33	44.83	25.28	42.59	25.32	20.39	36.84	18.98
+ Multi-scale temporal fusion	51.08	34.60	26.06	45.87	24.57	44.05	26.65	21.03	37.59	19.92
+ 3-way cross-attention	51.25	34.43	26.73	45.73	25.47	45.26	27.65	21.82	38.73	20.54
+ Multi-feature integration	**52.33**	**35.46**	**28.11**	**46.76**	**27.08**	**47.40**	**29.00**	**23.50**	**40.74**	**22.19**

Table 1: Ablation results on the CASTELLA validation and test splits. All scores are averaged over five random seeds.

For LLM verification, each candidate audio segment is cropped using its predicted start and end times and expanded on both sides by 50% of the candidate duration to provide additional acoustic context. The cropped audio, the text query, and the relative start and end positions of the candidate interval within the crop are given to Qwen2.5-Omni as a frozen verifier. The verifier is prompted to answer only "Yes" or "No" to whether the candidate interval accurately covers the moment described by the query. The normalized probability of the "Yes" answer, computed from the "Yes" and "No" logits, is used as the LLM relevance score s_{LLM}.

The final ranking score is computed as

$$s_{\mathrm{final}} = (1-\beta)s_{\mathrm{dec}} + \beta s_{\mathrm{LLM}}. \tag{3}$$

The interpolation weight was selected through hyperparameter tuning over $\beta \in \{0.30, 0.35, 0.45\}$ on the development splits, using Recall@0.7 as the primary criterion and mAP-based metrics as secondary criteria. The final values were $\beta = 0.35$ for Submit 2, $\beta = 0.45$ for Submit 3, and $\beta = 0.30$ for Submit 4.

The LLM-based module is used only as a post-processing reranker. It does not modify predicted temporal boundaries, update AMR model parameters, or use any ground-truth information from the hidden challenge evaluation set. The verifier uses only the provided audio content and text query, and no visual information from the original videos is used. Since Qwen2.5-Omni is an external pretrained model, its use is explicitly reported in the system description and submission metadata.

3. EXPERIMENTS

This section describes the experimental setup of our final system, including the datasets, feature extraction process, model configuration, and training procedure.

3.1. Datasets

We use Clotho-Moment for pretraining and CASTELLA for fine-tuning and model development. Clotho-Moment is a synthetic audio moment retrieval dataset constructed by overlaying short audio clips from Clotho onto long background recordings from Walking Tours [5]. It provides long audio recordings with text queries and temporal moment annotations, and is used only for pretraining.

CASTELLA is a human-annotated long-audio moment retrieval dataset with free-form captions and temporal boundaries [6]. It contains 1,009 training recordings, 213 validation recordings, and 640 test recordings. We use the official train, validation, and test splits without adding any manual annotations. The model is first pretrained on Clotho-Moment and then fine-tuned on the CASTELLA training split. The CASTELLA validation split was mainly used for model selection, while the CASTELLA test split was used as an additional development benchmark for selecting robust final candidates. The hidden challenge evaluation set was used only for generating the final submissions.

3.2. Model setup

All audio and text features are extracted from pretrained models and precomputed before training. The feature extractors are kept frozen throughout pretraining, fine-tuning, validation, and inference. Audio features are extracted using a 1-second window with a 1-second hop, and no audio augmentation is applied. Each feature stream is projected into a shared hidden dimension inside the AMR model before gated feature fusion.

LAION-CLAP [2] is used as the primary audio-language representation because it provides audio and text embeddings learned in a shared semantic space. To complement this main representation, we additionally use MS-CLAP [7], WavLM [8], and RoBERTa [9]. MS-CLAP provides auxiliary CLAP-style audio-text features, WavLM provides fine-grained acoustic and temporal audio representations, and RoBERTa provides text-only semantic representations for query understanding. The auxiliary feature streams are integrated with the main LAION-CLAP representation through the feature-level learnable scalar gates described in Section 2.1.

For the final submission systems, we use three feature combinations, as summarized in Table 2. Feature combination A uses LAION-CLAP, MS-CLAP, and WavLM-base-plus audio features with LAION-CLAP and MS-CLAP text features. Feature combination B uses LAION-CLAP and MS-CLAP audio-text features. Feature combination C uses LAION-CLAP together with multiple WavLM variants for audio and combines LAION-CLAP text features with RoBERTa text features. These combinations are used to construct diverse model candidates and ensemble systems for the final submissions.

Audio-language LLMs are not used for feature extraction or model training. Qwen2.5-Omni [10] is used only as a frozen audio-language verifier for post-processing-based re-ranking in Submit 2, Submit 3, and Submit 4. It receives the predicted candidate audio crop and the corresponding text query, and re-ranks the already predicted temporal windows without modifying their boundaries or updating any model parameters.

3.3. Training

For the final submission system, the model is first pretrained on Clotho-Moment and then fine-tuned on the CASTELLA training split. The CASTELLA validation split was used as the primary criterion for model selection, and the CASTELLA test split was additionally used as a development benchmark to select robust final candidates. The hidden challenge evaluation set was not used during training, validation, or model selection, and was used only for generating the final submission results. We train all models with AdamW using precomputed frozen audio and text features, without audio augmentation. The training objective consists of classification, span regression, temporal GIoU, saliency, and quality regression losses.

Feature combo	Audio features	Text features
A	LAION-CLAP, MS-CLAP, WavLM-base-plus	LAION-CLAP text, MS-CLAP text
B	LAION-CLAP, MS-CLAP	LAION-CLAP text, MS-CLAP text
C	LAION-CLAP, WavLM-base, WavLM-base-plus, WavLM-large	LAION-CLAP text, RoBERTa

Table 2: Feature combinations used in the final submission systems.

System	val R1@0.5	val R1@0.7	val mAP	val mAP@0.5	val mAP@0.75	test R1@0.5	test R1@0.7	test mAP	test mAP@0.5	test mAP@0.75
Submit 1	58.81	49.43	44.33	60.38	45.53	55.23	42.91	38.02	52.51	38.55
Submit 2	64.77	53.12	45.97	63.01	47.03	59.02	45.14	39.29	54.52	39.86
Submit 3	67.05	50.00	45.53	63.77	46.37	58.87	45.14	38.34	53.99	39.13
Submit 4	66.48	54.55	46.41	64.48	48.50	54.57	41.28	37.02	52.07	38.06

Table 3: Final submission performance on the CASTELLA validation and test sets.

4. RESULTS

In this section, we report the experimental results of the proposed system. We first present the model selection results through an ablation study on the CASTELLA development set. Then, we report the performance of the final submitted system after applying the selected architecture and training configuration.

4.1. Model selection

We conduct ablation experiments to analyze the contribution of each component and to determine the final architecture. Starting from the DETR-style baseline with LAION-CLAP features, we progressively add the BiMamba encoder, quality regression loss, multi-scale temporal fusion, 3-way cross-attention, and multi-feature integration. All results are averaged over five random seeds and are summarized in Table 1.

Table 1 shows that the BiMamba encoder provides the largest and most consistent improvement over the baseline, indicating the importance of bidirectional long-range temporal modeling for audio moment retrieval. The quality regression loss further improves most localization metrics, especially high-IoU scores, suggesting that explicit localization-quality estimation helps rank candidate windows more reliably. The effects of multi-scale temporal fusion and 3-way cross-attention are more nuanced: they do not improve every validation metric, but they improve most test-side metrics, suggesting better generalization. Finally, multi-feature integration achieves the best overall performance among the ablated configurations, supporting the use of complementary pretrained audio and text representations. Based on these results, we use the full model with multi-feature integration, 3-way cross-attention, BiMamba-based temporal modeling, multi-scale temporal fusion, and quality-aware prediction as the base architecture for the final submission systems.

4.2. Final system performance

After selecting the final architecture through the ablation study, we construct the final submission systems using different feature combinations and ensemble strategies. The base architecture includes multi-feature integration, 3-way cross-attention, BiMamba-based temporal modeling, multi-scale temporal fusion, and quality-aware prediction. Table 2 summarizes the feature combinations used for the final submissions.

Submit 1 is a diverse ensemble system that combines candidate predictions from models trained with feature combinations A, B, and C. It ranks the merged candidate windows using decoder-based confidence scores without LLM post-processing. Submit 2 uses the same diverse ensemble candidate set as Submit 1, but additionally applies the frozen audio-language LLM verifier for post-processing-based re-ranking. Submit 3 is an ensemble system based on feature combination A with LLM-based re-ranking, while Submit 4 is an ensemble system based on feature combination B with LLM-based re-ranking. In all LLM-based submissions, the verifier only re-ranks the top candidate windows and does not modify their temporal boundaries.

Table 3 reports the performance of the four final submissions on the CASTELLA validation and test splits. The results in Table 1 and Table 3 serve different purposes. Table 1 reports non-ensemble ablation results averaged over five random seeds, whereas Table 3 reports final submission systems with ensemble-based candidate merging and, for Submit 2–4, LLM-based re-ranking. Therefore, the higher scores in Table 3 reflect the cumulative effect of the selected full architecture, feature-combination diversity, model ensembling, and post-processing, rather than the effect of a single architectural component.

5. CONCLUSION

We presented QAM-DETR for DCASE 2026 Task 6, a DETR-style audio moment retrieval system that combines multi-feature audio-text representations, 3-way cross-attention-based fusion, BiMamba temporal encoding, multi-scale temporal fusion, and quality-aware prediction. Ablation results show that the BiMamba encoder provides the largest gain over the baseline, while the remaining components further improve robustness and localization performance. For the final submissions, we construct ensemble systems with different feature combinations; Submit 1 uses decoder-based ensemble ranking, whereas Submit 2, Submit 3, and Submit 4 additionally apply a frozen audio-language LLM verifier only for post-processing-based re-ranking. Future work includes improving boundary refinement, stabilizing multi-feature fusion, and developing more efficient re-ranking strategies for long audio recordings.

6. REFERENCES

[1] DCASE Community, “Audio moment retrieval from long audio,” DCASE 2026 Challenge, Task 6. [Online]. Available: https://dcase.community/challenge2026/task-audio-moment-retrieval-from-long-audio, 2026, accessed: Jun. 9, 2026.

[2] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[3] A. Nadeem, A. Hilton, R. Dawes, G. Thomas, and A. Mustafa, "CAD: Contextual multi-modal alignment for dynamic AVQA," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 7251–7263.

[4] B. Yu, J. Li, Y. Di, *et al.*, "Mamba-based modulated fusion model for video moment retrieval," *Sci. Rep.*, vol. 16, p. 15847, 2026, art. no. 15847.

[5] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Language-based audio moment retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, pp. 1–5.

[6] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long audio dataset with captions and temporal boundaries," *arXiv preprint arXiv:2511.15131*, 2026.

[7] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 336–340.

[8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[10] 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.