ENHANCED AUDIO MOMENT RETRIEVAL APPROACH FOR DCASE 2026 TASK 6

TECHNICAL REPORT

Takumi Ogawa, Nobuyuki Ohhashi, Yota Kurisuno, Minami Takenaka, Ririko Miyamoto

Yokohama City University, Yokohama, Kanagawa, 236-0027, Japan
{y255607c, y265609b, y265616b, y265623b, y265631c}@yokohama-cu.ac.jp

ABSTRACT

This technical report presents our solution to Task 6 of the DCASE 2026 Challenge, which focuses on retrieving specific moments within long audio recordings that correspond to a given textual query. We address this task by building on the official DETR-based baseline system provided by the DCASE 2026 Task 6 organizers and introducing three main components: M2D-CLAP audio features, a CG-DETR-based moment detection model enhanced with boundary modeling adapted from BAM-DETR, and a post-processing method for boundary refinement and candidate re-ranking. These components are designed to improve audio representation, temporal localization, and final moment selection. Our submitted system achieves an R1@0.7 of 41.13% on the development-testing set.

***Index Terms*—** Self-supervised learning, M2D-CLAP,CG-DETR , BAM-DETR, Boundary refinement

1. INTRODUCTION

This technical report presents our solution to Task 6 of the DCASE 2026 Challenge, which addresses the problem of Audio Moment Retrieval (AMR). AMR is the task of identifying the temporal segment in a long audio recording that corresponds to a given textual query. Different from conventional audio classification tasks, this task requires not only recognizing the existence of an audio event, but also estimating its start and end times.

The key technical emphases of our system include:

- **Audio feature extraction:** replacing the baseline CLAP audio features with M2D-CLAP features to obtain more detailed and robust audio representations.
- **Moment detection model improvement:** adopting CG-DETR as the base moment detection model and enhancing it with boundary modeling adapted from BAM-DETR to improve the prediction of start and end boundaries.
- **Boundary refinement and re-ranking:** refining the predicted start and end times using frame-level relevance scores and re-ranking candidate moments to select a more reliable final prediction.

In this work, we use the official DCASE 2026 Task 6 baseline as our base model and introduce three improvements to address these challenges. First, we replace the baseline CLAP audio features with M2D-CLAP features, which are obtained from a self-supervised audio model and are expected to capture more detailed acoustic patterns. Second, we adopt CG-DETR as the base moment detection model to improve query-aware moment detection, and enhance it with boundary-oriented modeling adapted from BAM-DETR to improve boundary prediction and candidate selection. Third, we apply a post-processing method based on frame-level relevance scores. These scores indicate how strongly each audio frame is related to the query, and we use them to refine predicted boundaries and re-rank candidate moments.

This paper is organized as follows. Section 2 describes the proposed approach, including each component of our system. Section 3 presents the experimental results. Section 4 concludes this report.

2. METHODS

2.1. Query Expansion via Large Language Models

To address the ambiguity and variability of queries in the CASTELLA dataset, we explored an LLM-based query augmentation strategy, based on the methodology introduced by Vo et al.[1]. Specifically, we utilized the Qwen2.5-Omni model to generate two types of augmentations for each original query:

- **Paraphrase**: Replacing words with synonyms to capture diverse linguistic expressions while preserving the original semantic meaning.
- **Acoustic-focused Augmentation**: Rewriting the query to explicitly emphasize acoustic characteristics (e.g., pitch, timbre, rhythm) for better alignment with the audio representations.

To integrate the predictions from the three queries, which are the original and the two augmented variants, the final pre-

diction score was calculated by adding the confidence scores output by the model.

Although the query augmentation demonstrated accuracy improvements in some of the metrics, we ultimately opted not to include this technique in our final submitted system due to a degradation of the R1@0.7 score.

2.2. Audio Feature Extraction with M2D-CLAP

In order to improve audio feature extraction, we replaced the baseline CLAP audio features with M2D-CLAP features. While the baseline CLAP audio encoder learns audio representations from supervised audio-text pairs, M2D-CLAP introduces self-supervised masked audio modeling. In this approach, parts of the audio spectrogram are masked, and the model is trained to predict the masked regions [2]. This allows the audio encoder to learn detailed audio features directly from the audio signal itself, rather than relying only on predefined labels. As a result, M2D-CLAP is less constrained by label-level supervision and can capture more detailed audio features.

Since this task requires identifying when an audio event starts and ends, we expected M2D-CLAP features to be more suitable because they can capture detailed changes in the audio signal. In our experiments, replacing CLAP features with M2D-CLAP features improved the performance of the baseline model.

2.3. Boundary-Aware Temporal Moment Localization(BAM-DETR

Temporal moment localization aims to identify the temporal interval in a video corresponding to a given text query. Existing methods such as QD-DETR represent a target moment using its center and duration:

$$B = (c, l)$$

where c and l denote the center and length of the interval, respectively.

However, this representation suffers from *center ambiguity*. Since the model tends to focus on highly salient regions, the predicted center does not always correspond to the actual temporal center of the target moment. As a result, when salient events appear near the beginning or end of the interval, localization performance may degrade and the predicted interval can have lower IoU with the ground truth.

To address this limitation, BAM-DETR[3] introduces a boundary-oriented formulation. Instead of predicting center and duration, a temporal moment is represented as:

$$B = (p, d_s, d_e)$$

where p denotes an anchor point inside the interval, and d_s and d_e represent the temporal distances from the anchor point to the start and end boundaries, respectively. By estimating interval boundaries directly, the model becomes less sensitive to center displacement and achieves more robust localization.

BAM-DETR further employs a Dual-pathway Decoder. The model first generates a coarse temporal proposal and then separately estimates the anchor position and boundary distances. The anchor pathway captures global temporal information to determine the anchor location, while the boundary pathway focuses on local regions around predicted boundaries for refinement. This decoupled design enables more accurate temporal localization.

In addition, BAM-DETR introduces Quality-based Ranking. Conventional approaches rank candidate intervals using only query matching confidence, which may favor short or fragmented predictions. BAM-DETR instead predicts a localization quality score and optimizes it with respect to IoU during training. Final predictions are selected according to the estimated localization quality, improving the reliability of interval selection.

2.4. Boundary Refinement and Re-ranking

The model described in the previous section produces frame-level saliency scores $\mathbf{s} \in \mathbb{R}^L$ as a by-product of the encoder, where L is the number of audio frames. Each score s_t reflects the relevance of frame t to the query, computed as the inner product between the frame-level audio memory and the global audio-text representation. Although $\mathbf{s}$ is used only as an auxiliary training signal in the original model, we exploit it explicitly at inference time to refine predicted boundaries. Since this approach requires no additional training and operates purely as post-processing on top of any model that outputs saliency scores, it is orthogonal to and compatible with any architectural improvements applied to the base model.

For each of the top-K candidate moments (s_0, e_0) output by the model, we search for an adjusted span (s', e') within a local window of radius R frames that maximises the following score:

$$\begin{aligned} f(s', e') = &\left(\bar{s}_{[s':e']} - \bar{s}_{\mathrm{ctx}}\right) \\ &- \beta \left(|s' - s_0| + |e' - e_0|\right) \\ &- \gamma \left|(e' - s') - (e_0 - s_0)\right|, \end{aligned} \tag{1}$$

where $\bar{s}_{[s':e']}$ is the mean saliency inside the candidate span and $\bar{s}_{\mathrm{ctx}}$ is the mean saliency over a context region of fixed width immediately outside both boundaries. The first term encourages the refined span to cover frames with high saliency relative to its surroundings, while the second and third terms penalise large boundary shifts and duration changes, respectively. The hyperparameters R, β, and γ are tuned on the validation split.

After refining all K candidates, we re-rank them by a composite score:

$$\mathrm{score}_k = \sigma_k + \lambda\, f(s'_k, e'_k) - \alpha\, k, \tag{2}$$

Table 1. Ablation study on the CASTELLA development-testing split.

	M2D-CLAP	CG-DETR+BAM	Boundary Ref.	R1@0.5	R1@0.7	mAP	mAP@0.5	mAP@0.75
Baseline				25.61	13.59	12.06	23.60	10.72
	✓			43.73	26.43	21.84	39.22	20.11
	✓	✓		54.12	39.64	29.26	44.39	29.19
Ours	✓	✓	✓	**54.19**	**41.13**	**30.79**	**44.78**	**30.74**

where σ_k is the original model confidence score of the k-th candidate and k is the original rank. The final prediction is the top-ranked candidate after re-ranking.

3. EVALUATION

3.1. Experimental Setup

We evaluate on the CASTELLA development-testing split. All models are pretrained on Clotho-Moment for 200 epochs and then fine-tuned on the CASTELLA training split for 200 epochs. The best checkpoint during fine-tuning is selected based on validation mAP, and the boundary refinement hyperparameters are tuned by grid search on the validation split.

3.2. Results

Table 1 presents an incremental ablation study in which each proposed component is added to the official DCASE 2026 Task 6 baseline one at a time.

Replacing the baseline CLAP features with M2D-CLAP yields the largest single improvement, raising R1@0.7 from 13.59% to 26.43%. Because M2D-CLAP is trained with self-supervised masked audio modeling, it learns richer frame-level representations than the label-supervised CLAP features in the baseline, which is particularly beneficial for precise temporal localisation.

Replacing QD-DETR with CG-DETR incorporating a BAM-style decoder provides a similarly large gain, pushing R1@0.7 further to 39.64%. The query-adaptive cross-attention mechanism and a dedicated quality-ranking head improve the accuracy of the predicted boundaries and the ordering of candidate moments.

Adding saliency-guided boundary refinement and re-ranking as a final post-processing step brings R1@0.7 to 41.13% and mAP to 30.79%. While the gain from this step is smaller in absolute terms, it is consistent across all IoU thresholds, suggesting that the frame-level saliency scores provide information about moment boundaries that complements the model's direct predictions.

4. CONCLUSION

In this report, we presented our system for DCASE 2026 Task 6, which addresses the problem of audio moment retrieval from long audio recordings. We introduced three improvements over the official baseline: replacing MS-CLAP with M2D-CLAP audio features, adopting a boundary-aware detection architecture based on BAM-DETR, and applying saliency-guided boundary refinement and re-ranking as a post-processing step. Ablation experiments on the CASTELLA development-testing split confirmed that each component contributes independently, with the feature replacement and architecture change providing the largest gains and the post-processing step offering consistent improvement across all IoU thresholds. Our final system achieves R1@0.7 of 41.13% and mAP of 30.79%, substantially outperforming the baseline.

5. REFERENCES

[1] Van-Thinh Vo, Minh-Khoi Nguyen, Minh-Huy Tran, Anh-Quan Nguyen-Tran, Duy-Tan Nguyen, and Loi Nguyen, "Enhanced multimodal video retrieval system: Integrating query expansion and cross-modal temporal event retrieval," 2025.

[2] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, Masahiro Yasuda, Shunsuke Tsubaki, and Keisuke Imoto, "M2d-clap: Masked modeling duo meets clap for learning general-purpose audio-language representation," 2024.

[3] Pilhyeon Lee and Hyeran Byun, "BAM-DETR: Boundary-aligned moment detection transformer for temporal sentence grounding in videos," 2023.