Quality- and Boundary-Aware Cross-Modal Refinement for Long-Audio Moment Retrieval

Technical Report

Yingzhao Hou[1], Anda Liu[1], Zhongqin Shu[1], Xin Guo[1], Yuzheng Wu[1], Yiwei Liu[1], Xiaolan Xia[1], Xueqin Luo[2], Gongping Huang[1,*]

[1]School of Electronic Information, Wuhan University, China

[2] CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China

Abstract

This technical report describes our submission to DCASE 2026 Challenge Task 6: Audio Moment Retrieval from Long Audio. Based on the official QD-DETR baseline, we improve the system through three aspects: cross-modal semantic refinement, temporal localization optimization, and long-context training data construction. First, we adapt the Comprehensive Integration Module (CIM) and Multi-Aspect Contrastive Learning (MCL) from UVCOM to enhance audio-language semantic interaction. Second, we introduce several task-specific designs, including dense all-window saliency supervision, auxiliary audio-text similarity learning, localization-quality-aware candidate resorting, short-window proposal generation, and candidate-level boundary hard negative learning, which jointly improve proposal generation, boundary estimation, and candidate ranking. Third, we construct Clotho-Moment-Long, a long-context extension of Clotho-Moment following the same data generation pipeline while extending the audio context to 300 s and introducing repeated event occurrences and sparser event placement to better simulate realistic and challenging long-audio retrieval scenarios. The resulting model is first trained on Clotho-Moment-Long and then fine-tuned on the real-world CASTELLA dataset. On the CASTELLA development-testing split, the final fused system improves Recall1@0.7 from the official CASTELLA+Clotho-Moment baseline value of 13.59 to 25.09, while increasing average mAP from 12.06 to 18.61.

Index Terms—audio moment retrieval, long audio understanding, audio-text grounding, cross-modal refinement

1. Introduction

DCASE 2026 Task 6 addresses audio moment retrieval from long audio recordings [1]. Given a long audio recording and a natural-language query, the system must localize the temporal segment corresponding to the query. The task is challenging because the model must understand query semantics, distinguish target events from long background audio, and predict accurate temporal boundaries.

The official baseline is based on QD-DETR [5] and uses organizer-provided MS-CLAP 2023 audio features extracted with a sliding window, together with the corresponding text-query features [4]. Building on this baseline, we adopt a UVCOM-inspired cross-modal semantic refinement framework [6] to enhance audio-language semantic alignment for long-audio moment retrieval. We adapt this backbone from video moment retrieval to audio moment retrieval by replacing video clip tokens with MS-CLAP audio-window tokens and redesigning the training and inference strategy for long-audio localization. In addition, We introduce several task-specific enhancements for DCASE 2026 Task 6. Dense all-window saliency supervision provides richer temporal supervision across candidate windows, while auxiliary audio-text similarity learning further strengthens audio-language semantic alignment. To improve localization accuracy, we employ localization-quality-aware candidate reranking and candidate-level boundary hard negative learning. Short-window proposal generation is also introduced to better capture short-duration target events. Finally, we construct the Clotho-Moment-Long dataset, a long-context extension of Clotho-Moment for long-audio training. It follows the same feature extraction pipeline as the baseline, while extending the audio context and introducing repeated event occurrences and sparser event placement to better simulate realistic long-audio retrieval scenarios. The ablation results show that these components address different parts of the retrieval pipeline and that the final fusion gives the best performance.

2. System Description

2.1 Backbone: QD-DETR with Cross-modal Semantic Refinement

The Overall architecture of the proposed long-audio moment retrieval system is illustrated in Fig. 1. Our system builds on the official QD-DETR baseline and adopts the Comprehensive Integration Module (CIM) and an MCL-style auxiliary contrastive loss from UVCOM [6] for audio-language cross-modal refinement. Unlike the original video-text setting, UVCOM is adapted to audio-only long-audio moment retrieval by replacing video tokens with one-second MS-CLAP audio-window tokens and concatenating temporal endpoint features to en-

*Corresponding author: Gongping Huang (gongpinghuang@whu.edu.cn). Code repository: https://github.com/HYZ-0514/dcase2026-quality-boundary-aware-lamr.

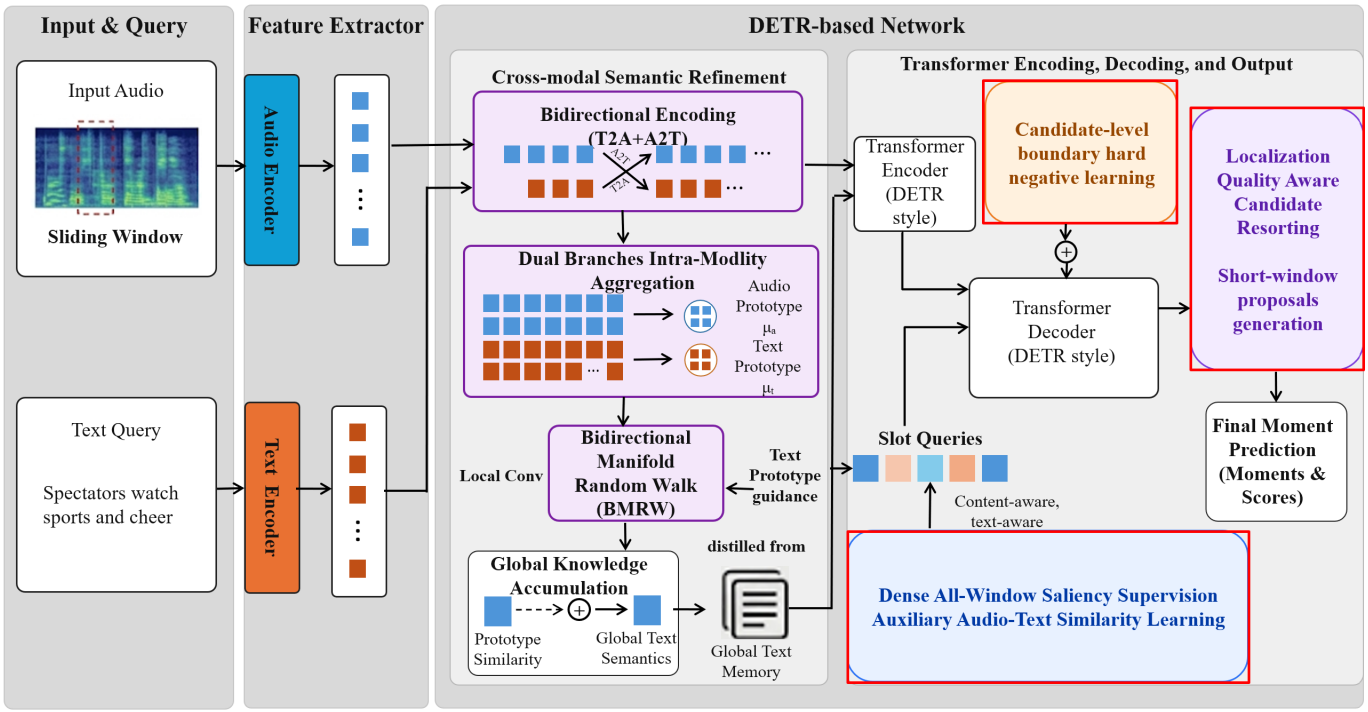

Figure 1: Overall architecture of the proposed long-audio moment retrieval system.

code window positions. The CIM contains Dual Branches Intra-Modality Aggregation (DBIA), Local Relation Perception (LRP), and Global Knowledge Accumulation (GKA), which are adapted from video-text modeling to audio-text modeling.

The refined audio features are subsequently fed into the QD-DETR temporal grounding head, preserving the original DETR-style span prediction and Hungarian matching framework while benefiting from stronger audio-language semantic alignment.

2.2 Task-specific Improvements

Based on the QD-DETR grounding head and the adapted UVCOM refinement backbone, we introduce several task-specific components for long-audio moment retrieval. These components target three practical issues in this task: incomplete local audio-text supervision, inaccurate candidate ranking, and boundary ambiguity for short or partially overlapped events.

Dense all-window saliency supervision. The original saliency supervision samples positive and negative clips. During CASTELLA fine-tuning, we instead use dense all-window saliency labels. Every one-second clip inside any ground-truth relevant window is marked as positive, while clips outside the relevant windows are treated as negative. This gives the model a more complete local relevance signal, especially for events spanning multiple adjacent audio windows.

Auxiliary audio-text similarity learning. We also keep an auxiliary audio-text similarity branch to encourage global and local cross-modal alignment. The global audio memory and global text representation are projected into a normalized embedding space and trained with a symmetric contrastive objective. At the local level, clips inside ground-truth moments are encouraged to have higher similarity with the query. This auxiliary supervision helps align text semantics with both global recording context and local acoustic evidence.

Localization-quality-aware candidate resorting. As shown in Fig. 1, this module is attached after the DETR-style decoder and before the final moment prediction output. The original decoder ranks candidate windows mainly by foreground classification confidence, which reflects semantic matching but does not always indicate boundary accuracy. Inspired by the quality-based scoring strategy in BAM-DETR [7], we add a lightweight quality head to estimate the localization reliability of each candidate:

$$q_i = \sigma(\mathrm{MLP}_{\mathrm{quality}}(h_i)), \tag{1}$$

where h_i is the decoder feature of the i-th candidate. The quality score is supervised by the temporal IoU between the matched prediction $\hat{W}_i$ and its matched ground-truth window W_i^{m}:

$$\mathcal{L}_{\mathrm{quality}} = \frac{1}{M}\sum_{i=1}^{M}\left|q_i - \mathrm{IoU}(\hat{W}_i, W_i^{\mathrm{m}})\right|. \tag{2}$$

During inference, candidates are re-ranked by combining the foreground probability and the predicted localization quality:

$$s_i^{\mathrm{final}} = p_i^{\mathrm{fg}} \cdot q_i^{\gamma}, \tag{3}$$

where p_i^{fg} is the foreground classification probability and $\gamma = 1.5$. This module only changes the ranking order of candidate windows

Candidate-level boundary hard negative learning. This objective is motivated by boundary-aware temporal grounding and negative sample mining in temporal localization [7, 8]. This module is a training-side boundary accuracy objective attached to the candidate windows produced by the QD-DETR decoder. After the decoder predicts candidate temporal windows and foreground scores, we compute an additional boundary discrimination loss and add it to the total training objective as $\mathcal{L}_{\text{boundary}}$. The loss is back-propagated through the QD-DETR temporal grounding head and the UVCOM-style refinement backbone, but it does not introduce a new decoder branch, change Hungarian matching, or modify inference-time post-processing.

The objective focuses on medium-quality candidates that roughly cover the target event but still contain boundary errors. For each predicted candidate window $\hat{W}_i$, we find its best-overlap ground-truth window:

$$W_i^* = \arg\max_{W \in \mathcal{G}} \text{IoU}(\hat{W}_i, W), \tag{4}$$

where $\mathcal{G}$ denotes the set of ground-truth windows for the current query, and W denotes a ground-truth window in this set. The candidate is selected only when

$$0.40 \leq \text{IoU}(\hat{W}_i, W_i^*) \leq 0.75. \tag{5}$$

For each selected candidate, we construct boundary hard negatives from the matched ground-truth window W_i^*. These negatives simulate four common boundary errors: over-expansion, under-coverage, temporal shift, and single-boundary offset, and are filtered to keep their IoU with the ground truth in $[0.45, 0.75]$. The selected candidate is trained to be closer to the ground truth than the constructed boundary-error negative:

$$\mathcal{L}_{\text{bd}}^{(i)} = \max\left(0, m + d(\hat{W}_i, W_i^*) - d(W_i^-, W_i^*)\right), \tag{6}$$

where W_i^- denotes a constructed boundary hard negative window, $m = 0.03$, and $d(\cdot,\cdot)$ combines center-width span distance and generalized temporal IoU distance. We average losses within each error type and then across types to avoid one boundary-error pattern dominating the objective.

Finally, we apply score-aware weighting using the detached foreground confidence p_i^{fg}, which is consistent with the candidate foreground score used in localization-quality-aware candidate resorting:

$$\omega_i = \frac{\max(p_i^{\text{fg}}, \epsilon)}{\frac{1}{N_{\text{bd}}}\sum_{j=1}^{N_{\text{bd}}} \max(p_j^{\text{fg}}, \epsilon)}, \quad \mathcal{L}_{\text{boundary}} = \frac{1}{N_{\text{bd}}} \sum_{i=1}^{N_{\text{bd}}} \omega_i \mathcal{L}_{\text{bd}}^{(i)}. \tag{7}$$

Here, N_{bd} is the number of selected boundary candidates and $\epsilon = 0.05$. Since p_i^{fg} is detached, it only weights the boundary loss and does not create an additional inference ranking path.

Short-window proposal post-processing. To improve localization of short target events, we add short-window proposals around local saliency peaks during inference. The proposal lengths are 1, 2, and 3 seconds, and the short-proposal base coefficient is set to 0.25. Candidate windows are clipped to the audio duration, rounded to one-second boundaries, and filtered with temporal non-maximum suppression.

2.3 Training Objective

The final objective combines the original DETR-style localization losses with the task-specific objectives:

$$\begin{aligned}\mathcal{L} = \lambda_{\text{span}}\mathcal{L}_{\text{span}} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} \\ + \lambda_{\text{sal}}\mathcal{L}_{\text{saliency}} + \lambda_{\text{sim}}\mathcal{L}_{\text{similarity}} \\ + \lambda_{\text{quality}}\mathcal{L}_{\text{quality}} + \lambda_{\text{bd}}\mathcal{L}_{\text{boundary}}.\end{aligned} \tag{8}$$

Here, $\mathcal{L}_{\text{span}}$ and $\mathcal{L}_{\text{giou}}$ are the original QD-DETR temporal localization losses. $\mathcal{L}_{\text{span}}$ is the L1 loss for center-width span regression, and $\mathcal{L}_{\text{giou}}$ is the generalized temporal IoU loss between predicted and ground-truth windows. Together with the original foreground classification loss $\mathcal{L}_{\text{cls}}$ and saliency loss $\mathcal{L}_{\text{saliency}}$, their weights follow the official QD-DETR baseline configuration [1, 5]. Specifically, we set $\lambda_{\text{span}} = 10$, $\lambda_{\text{giou}} = 1$, $\lambda_{\text{cls}} = 4$, and $\lambda_{\text{sal}} = 1$. For the auxiliary audio-text similarity loss, we set $\lambda_{\text{sim}} = 0.5$. For the task-specific quality and boundary objectives, we set $\lambda_{\text{quality}} = 1.0$ and $\lambda_{\text{bd}} = 0.05$. The boundary candidate range is $[0.40, 0.75]$, and the generated negative IoU range is $[0.45, 0.75]$. These task-specific coefficients and boundary ranges are selected on the CASTELLA development-validation split.

3. Experimental Setup

3.1 Data and Features

Clotho-Moment-Long Dataset. We build on the original Clotho-Moment dataset [2], which synthesizes audio moment retrieval samples by overlaying Clotho foreground audio-caption pairs onto Walking Tours background recordings. Using the public Lighthouse Wrapper generation pipeline, we construct Clotho-Moment-Long by extending the audio context to 300 s, increasing the mean inter-onset interval to 45 s, and allowing each sampled event to recur 2-4 times with probability $p = 0.35$. These modifications produce longer, sparser, and multi-moment retrieval scenarios. The resulting dataset contains 6,274 audio clips and 14,714 query-moment annotations.

Dataset Setting and Features. The model is first pretrained on Clotho-Moment-Long and then fine-tuned on CASTELLA [3]. The CASTELLA development-validation split is used for model selection, and the development-testing split is used only for evaluation. For CASTELLA, we use organizer-provided MS-CLAP 2023 features, while features for Clotho-Moment-Long are extracted using the same MS-CLAP encoder and sliding-window configuration. Each audio recording is represented as a sequence of 1-second window embeddings with a 1-second hop, and each window embedding has 768 dimensions. Audio features are concatenated with temporal endpoint features to provide explicit temporal position information for moment localization. The maximum audio length is 300 windows, and the maximum text length is 32 tokens.

3.2 Training and Inference Settings

All systems are trained with AdamW. The learning rate is 10^{-4}, weight decay is 10^{-4}, batch size is 32, and gradient clipping is

Table 1: Results on the CASTELLA development-testing split.

System	R1@.5	R1@.7	mAP	mAP@.5	mAP@.75
B0: official baseline, CASTELLA only	23.16	10.32	9.11	20.34	6.96
B1: official baseline, CASTELLA + Clotho-Moment	25.61	13.59	12.06	23.60	10.72
S1: official baseline, CASTELLA + Clotho-Moment-Long	29.84	17.97	14.60	26.44	14.07
S2: UVCOM-style backbone, CASTELLA + Clotho-Moment-Long	32.29	20.34	16.06	28.53	15.72
S3: S2 + quality-aware resorting	36.38	24.65	18.42	**31.98**	17.66
S4: S2 + all-window saliency	33.33	21.31	16.42	28.88	15.70
S5: S4 + short proposals	33.33	21.31	16.87	28.92	16.21
S6: S5 + boundary hard negative learning	32.67	21.08	16.96	28.59	16.71
S7: S5 + quality-aware resorting	35.93	23.24	17.03	30.60	15.74
S8: S6 + quality-aware resorting	**37.49**	**25.09**	**18.61**	31.97	**17.86**

set to 0.1. The random seed is 2023, and training is performed for 200 epochs. The transformer hidden dimension is 256. The model uses 2 encoder layers, 2 decoder layers, 8 attention heads, feed-forward dimension 1024, dropout 0.1, input dropout 0.5, and 10 decoder queries.

During inference, each query returns up to 10 ranked candidate windows. The final system uses localization-quality-aware resorting and short-window proposal post-processing.

4. Results and Ablation

Table 1 reports results on the CASTELLA development-testing split. B0 and B1 are official baseline results. The remaining rows are local evaluations obtained with the same standalone evaluation script and local `submission_metrics.json` files. All values are percentages.

The ablation table separates the adopted backbone from the task-specific improvements. Compared with S1, S2 improves Recall1@0.7 from 17.97 to 20.34 and average mAP from 14.60 to 16.06, showing the benefit of using the UVCOM-style backbone. S3 shows that quality-aware resorting is effective for candidate ordering, improving Recall1@0.7 to 24.65 and mAP to 18.42 on top of S2.

S4 and S5 evaluate localization-oriented improvements. All-window saliency improves Recall1@0.7 to 21.31, and short proposals further improve average mAP from 16.42 to 16.87. Adding candidate-level boundary hard negative learning in S6 improves average mAP from 16.87 to 16.96 and mAP@0.75 from 16.21 to 16.71, although Recall1@0.5 and Recall1@0.7 slightly decrease. This indicates that the boundary objective mainly improves candidate-window quality and high-threshold localization rather than directly improving the top-ranked prediction.

S7 applies quality-aware resorting to S5 and serves as a resorting-only comparison on the short-proposal branch. It improves Recall1@0.7 from 21.31 to 23.24 and average mAP from 16.87 to 17.03, but mAP@0.75 decreases from 16.21 to 15.74. The final fused system S8 combines boundary hard negative learning with quality-aware resorting and achieves the best Recall1@0.5, Recall1@0.7, average mAP, and mAP@0.75. Compared with S5, S8 improves Recall1@0.7 from 21.31 to 25.09 and mAP from 16.87 to 18.61. Compared with the official CASTELLA+Clotho-Moment baseline, S8 improves Recall1@0.7 from 13.59 to 25.09 and average mAP from 12.06 to 18.61.

These results show that the proposed improvements address different parts of the retrieval pipeline. The UVCOM-style backbone improves query-aware long-audio representations. Dense all-window saliency and short proposals improve local evidence modeling and short-event localization. Boundary hard negative learning improves the quality of candidate windows, especially reflected by mAP@0.75. Quality-aware resorting improves candidate ordering. Their fusion in S8 is most effective because improved candidate windows can be promoted more reliably in the ranked output.

5. Submitted System and Rule Compliance

The official evaluation output is generated with `config.yml`. The checkpoint is `results/best_checkpoint.pth`, and inference uses the final fused setting with localization-quality-aware candidate resorting and short-proposal post-processing. The output file is `results_evaluation/private_submission.jsonl`. It contains 176 evaluation queries and predicts up to 10 ranked windows per query.

The system uses organizer-provided MS-CLAP 2023 audio and text features for CASTELLA. For Clotho-Moment-Long, compatible MS-CLAP features are extracted locally using the same preprocessing configuration. Clotho-Moment-Long is used for pretraining, while CASTELLA is used for fine-tuning and development evaluation. We use the QD-DETR baseline code as the starting point for the DETR-style temporal grounding model, and use UVCOM as the inspiration for cross-modal semantic refinement.

No visual information from the original videos is used. The hidden evaluation data are not annotated, LLM APIs such as ChatGPT or Gemini are not used, and the evaluation data are not modified.

6. Conclusion

This report presented a system for DCASE 2026 Task 6. Starting from the official QD-DETR baseline, we adopt a UVCOM-style backbone and add task-specific improvements for saliency supervision, auxiliary similarity learning, candidate ranking, short-event proposal generation, and boundary hard negative learning. The final fused system S8 combines boundary-aware candidate learning with quality-aware resorting, allowing the model to generate higher-quality candidate windows and better promote them in the ranked output. On the CASTELLA development-testing split, the final system improves Recall1@0.7 from the official CASTELLA+Clotho-Moment baseline value of 13.59 to 25.09, with average mAP increasing from 12.06 to 18.61.

References

[1] DCASE Community, “DCASE 2026 Challenge Task 6: Audio Moment Retrieval from Long Audio,” 2026. [Online]. Available: https://dcase.community/ch

allenge2026/task-audio-moment-retrieval-from-long-audio

[2] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Language-based Audio Moment Retrieval," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[3] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long Audio Dataset with Captions and Temporal Boundaries," arXiv:2511.15131, 2026.

[4] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: Learning Audio Concepts from Natural Language Supervision," 2023.

[5] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-Dependent Video Representation for Moment Retrieval and Highlight Detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] Y. Xiao, Z. Luo, Y. Liu, Y. Ma, H. Bian, Y. Ji, Y. Yang, and X. Li, "Bridging the Gap: A Unified Video Comprehension Framework for Moment Retrieval and Highlight Detection," arXiv:2311.16464, 2023.

[7] P. Lee and H. Byun, BAM-DETR: Boundary-Aligned Moment Detection Transformer for Temporal Sentence Grounding in Videos," in *Proc. European Conference on Computer Vision (ECCV)*, 2024, pp. 220–238.

[8] Z. Wang, L. Wang, T. Wu, T. Li, and G. Wu, Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2613–2623.