FLAM-CONDITIONED QD-DETR WITH PARAPHRASE POOLING AND WEIGHTED-BOXES-FUSION ENSEMBLING FOR AUDIO MOMENT RETRIEVAL

Technical Report

Yaozhong Kang *Runwu Shi* *Benjamin Yen* *Kazuhiro Nakadai*

Department of Systems and Control Engineering, Institute of Science Tokyo
Tokyo, Japan
{kangyaozhong, shirunwu, benjamin, nakadai}@ra.sc.eng.isct.ac.jp

ABSTRACT

We describe our submission to DCASE 2026 Task 6, Audio Moment Retrieval (AMR) from long audio. Starting from the official QD-DETR baseline, we leave the detection head untouched and instead strengthen the two stages that bound its boundary-precise recall: how audio and text are encoded, and how predictions are combined. Replacing the MS-CLAP encoder with FLAM (Frame-Wise Language-Audio Modeling), whose features are text-aligned at a high frame rate, is the single largest source of our gain. We then add several inexpensive sources of prediction diversity, namely query paraphrasing, saliency-curve candidates, and augmentation-trained students, and fuse them with a one-dimensional Weighted-Boxes-Fusion (WBF) ensemble that an iterative leave-one-out sweep prunes to its most complementary members. On the held-out test split the final ensemble more than doubles the baseline, raising Recall1 at 0.7 IoU (R1@0.7) from 13.59 to 33.23 and mean average precision (mAP) from 12.06 to 24.51.

Index Terms— *Audio moment retrieval, temporal grounding, FLAM, QD-DETR, weighted boxes fusion, ensemble*

1. INTRODUCTION

Audio Moment Retrieval (AMR) localizes the temporal interval(s) within a long audio recording that match a natural-language query [1]. DCASE 2026 Task 6 [2] pairs a large synthetic training corpus, Clotho-Moment, derived from the Clotho captioning dataset [3], with a smaller human-annotated benchmark, CASTELLA [4]. Following the challenge, development-training adds the benchmark's train split to the synthetic corpus, while development-validation and development-testing are its validation and test splits; all numbers below are on development-testing. The official baseline adapts QD-DETR [5], a query-dependent detection-transformer (DETR) head from the video-moment-retrieval literature [6], to audio by encoding the waveform and query with the contrastive MS-CLAP encoder [7]. What makes the task hard is boundary precision: target events are often short and embedded in cluttered audio, so a system must place interval edges within a fraction of a second rather than merely find the right region. This is what the strict 0.7-IoU recall metric rewards, and it is where the baseline, built on a roughly 1 fps encoder, is weakest.

We keep the QD-DETR detection head and rebuild the rest of the pipeline around three findings (Figure 1). First, boundary-precise recall (high intersection-over-union, IoU) is limited at the feature-encoder stage; once a high-frame-rate text-aligned encoder is in place, the detection head is comparatively saturated. Second, AMR offers several cheap sources of prediction diversity, namely text paraphrasing, saliency curves, and augmentation-trained students, that an ensemble stage can exploit. Third, Weighted-Boxes-Fusion turns that diversity into boundary-precise predictions far more effectively than non-maximum suppression (NMS). Section 2 describes the system, Section 3 reports results and an oracle analysis that characterizes the remaining headroom.

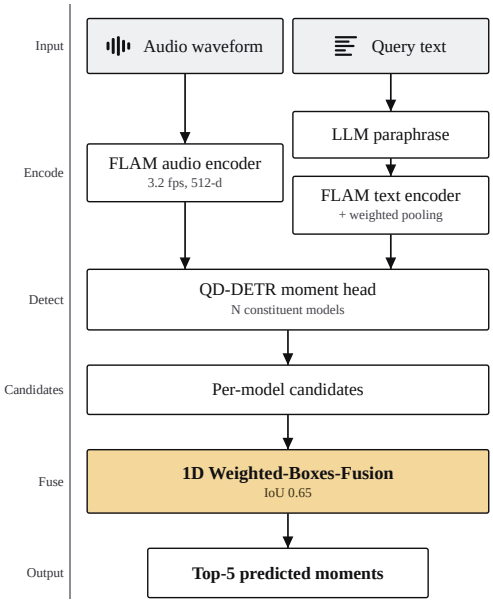

Figure 1: Overview of the inference pipeline.

2. SYSTEM DESCRIPTION

2.1. Feature encoder: FLAM

We replace MS-CLAP with FLAM (Frame-Wise Language-Audio Modeling) [8], which produces frame-level text-aligned audio embeddings at 3.2 frames per second in a 512-dim joint audio–text space. At 3.2 fps FLAM resolves sub-second boundaries that MS-CLAP at roughly 1 fps cannot, and in single-model transfer experi-

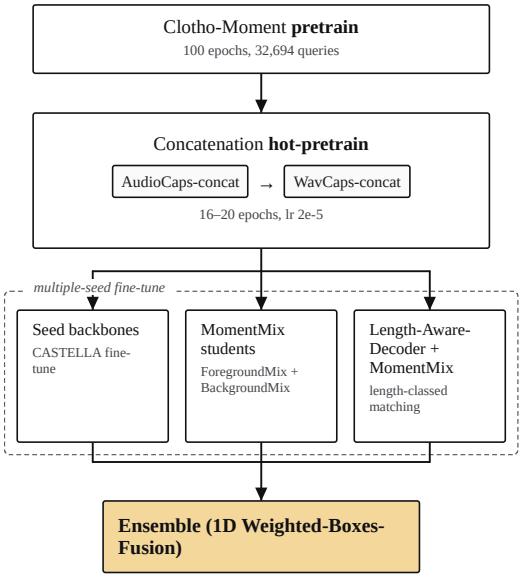

Figure 2: The three-stage training ladder.

ments it raised validation R1@0.7 by about 5 points. We pass FLAM audio features to the QD-DETR cross-attention encoder and FLAM text features to the query branch. We avoid concatenating features of different frame rates (for example FLAM with CLAP), which regressed in every trial because the loader truncates to the shorter sequence and discards FLAM's temporal resolution.

2.2. Training ladder

Each constituent model follows a three-stage schedule (Figure 2).

Clotho-Moment pretrain. The head is pretrained on the 32,694 synthetic queries, the only stage with supervision at scale. It learns to localize moments from FLAM features on densely annotated synthetic soundscapes, giving every later stage a consistent starting point before the model sees any real recording.

Concatenation hot-pretrain. Continued pretraining on long sequences synthesized by concatenating AudioCaps [9] clips (AC-concat) and WavCaps [10] `AudioSet_SL` clips (WC-concat) into 30 to 150 s soundscapes. This stage narrows the duration and clutter gap between the synthetic corpus and the real benchmark and yields the strongest single-checkpoint backbones, with WC-concat ahead of AC-concat.

CASTELLA fine-tuning. On the benchmark's merged train and validation set, with a 200-query random holdout reserved only for checkpoint selection by moment-retrieval mAP.

We train four seeds per recipe; with seed variance around 2 R1@0.7 points, averaging across seeds pays off. All runs use the default lighthouse QD-DETR configuration [11] and AdamW at learning rate 10^{-4}, reduced to 2×10^{-5} from the concatenation stage onward. The pretrain, concatenation, fine-tuning, and student stages run for 100, 16–20, 200, and 100 epochs.

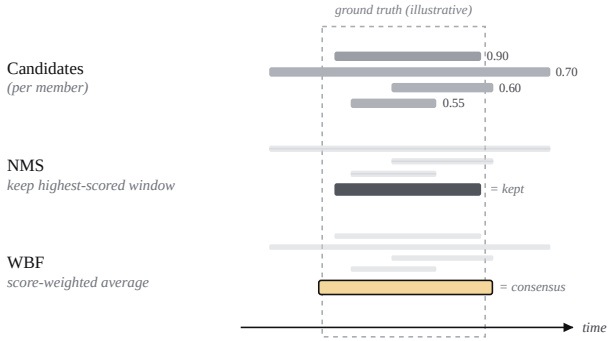

Figure 3: Why Weighted-Boxes-Fusion beats NMS on boundary precision.

2.3. Diversification mechanisms

Paraphrase pooling. At inference we generate three paraphrases per query with a local LLM (Qwen2.5-7B-Instruct [12]) at sampling temperature 0.8, encode the original and the paraphrases with the FLAM text encoder, and pool the L2-normalized vectors by a weighted mean (original weight 1.0, paraphrase weight 0.5) followed by a final L2 normalization. This is a free, training-free source of text-side diversity that helps most seeds by 0.3 to 0.9 R1@0.7.

Saliency-curve candidates. QD-DETR exposes a per-frame saliency curve alongside its span predictions. We convert each curve to candidate windows by min-max normalization, width-3 smoothing, peak detection above 0.85, expansion while above 0.7, and a 10 s cap, then add the result as a separate ensemble member. The same thresholds apply to every checkpoint.

Augmentation-trained students. We add MomentMix [13] students, where ForegroundMix shuffles in-moment frames and BackgroundMix replaces out-of-moment frames with random crops from other recordings ($\epsilon_{cut} = 16$, $p = 0.7$, 100 epochs), and Length-Aware-Decoder plus MomentMix joint students, which partition QD-DETR's ten query slots into four length classes with class-stratified bipartite matching. These students trade a little exact-boundary precision for boundary diversity that the fusion stage exploits, so they enter the ensemble with a reduced span weight of 0.75.

2.4. Ensembling: 1D Weighted-Boxes-Fusion

We fuse all members with a 1D adaptation of Weighted-Boxes-Fusion (WBF) [14]. NMS keeps one survivor per cluster and discards the others, so it forwards a single member's boundary. WBF instead averages all overlapping windows weighted by their scores and produces a consensus boundary (Figure 3). The fusion-ablation block of Table 1(a) shows the effect: at the 42-member set WBF adds 5.4 Recall1@0.7 and 2.2 mAP over NMS at no training cost, and the high-IoU Recall1@0.7 metric gains the most because it is boundary-limited. We use confidence type *avg* and an IoU clustering threshold of 0.65 that scales up with ensemble size. Most members carry unit weight; the augmentation-trained students and a few saliency members enter at a reduced 0.75, which down-weights heads whose boundaries are less reliable.

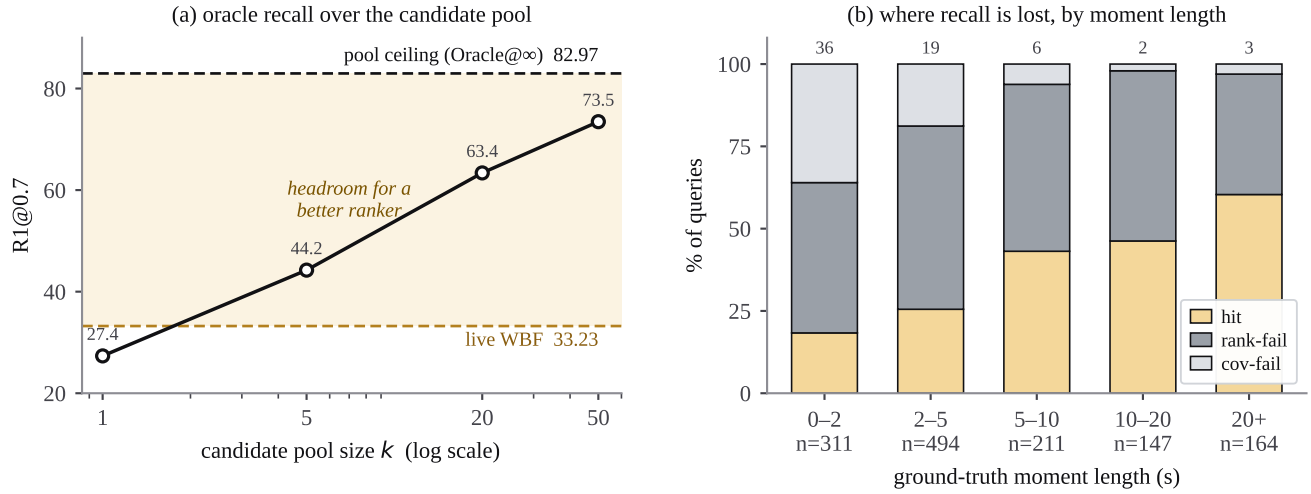

Figure 4: Error analysis on development-testing: oracle headroom (a) and failure modes by moment length (b).

(a) Results, fusion ablation, staged construction

Configuration	R1@0.5	R1@0.7	mAP
Baseline			
DCASE 2026 baseline	25.61	13.59	12.06
Fusion ablation (42-member)			
NMS	37.83	25.40	21.36
WBF	43.63	30.82	23.54
Staged construction			
+ MomentMix, Length-Aware students	44.84	31.88	24.17
+ 8-pass drop sweep	**45.89**	**33.23**	**24.51**

(b) Ensemble composition

Backbone family	Seeds	Members
Base (CM, CASTELLA)	2023/42/1234	12
Base, seed 7	7	4
Length-rebalanced (saliency)	2023	2
AudioCaps-concat	2023/42	8
WavCaps-concat	2023/42/7/1234	16
MomentMix students	42/7/1234	12
Length-Aware + MomentMix	42/7	8
Members before sweep		62
Members after sweep		**51**
Fusion IoU threshold		0.65

Table 1: Main results (a) and ensemble composition (b) on development-testing (CASTELLA test, 1327 queries; the DCASE baseline reaches R1@0.7 of 13.59). In (a), the fusion-ablation rows fix a 42-member set and vary only NMS versus WBF, and the staged-construction rows build up to our submitted 51-member ensemble (WBF IoU 0.65). R1@τ is Recall1 at IoU τ, higher is better. In (b), each backbone contributes up to four members, one per combination of original or paraphrase text and span or saliency output.

2.5. Ensemble composition and pruning

Table 1(b) lists the backbone families and their member counts. A model contributes up to four members, one for each combination of original or paraphrase-pooled text and span or saliency output; the length-rebalanced backbone contributes saliency-only members. After assembling each ensemble we run an iterative drop sweep that removes the most redundant members and re-scores after each pass. Each pass recovers 0.1 to 0.4 R1@0.7 and consistently flags paraphrase coverage of an already-represented seed as the most redundant axis; eight passes prune the ensemble from 62 to 51 members. The submitted systems use only the query text and the audio duration, with no other evaluation-set metadata.

3. RESULTS

3.1. Main result

Table 1(a) consolidates the headline numbers, the fusion ablation, and the staged construction of our system. The final ensemble lifts R1@0.7 from the baseline's 13.59 to 33.23 and mAP from 12.06 to 24.51, a 2.4 times gain on the high-IoU metric. We omit validation scores because the train and validation merge contaminates them. The fusion-ablation block isolates the effect of replacing NMS with WBF at a fixed member set, and the staged-construction block traces the system from the WBF base through the augmentation-trained students to the drop sweep, which adds a further 1.35 R1@0.7 while holding mAP steady; this 51-member ensemble is our primary submission, with a smaller 36-member ensemble (WBF IoU 0.60) entered as a secondary. Despite its size the ensemble stays cheap to run: the frozen FLAM encoder is computed once per recording and shared across members, each of which is only a lightweight QD-DETR head over the cached features.

3.2. Boundary tightness

Read another way, the gain concentrates in edge precision rather than mere localization. Of the queries the baseline places within 0.5 IoU, only 53% stay correct at the stricter 0.7 IoU threshold (13.59 of 25.61); for our system that share rises to 72% (33.23 of 45.89). The pipeline therefore improves not only how often the right

moment is found but how precisely its edges are drawn, which is exactly where the high-frame-rate encoder and consensus fusion are aimed and where the baseline lagged most.

3.3. Oracle analysis and remaining headroom

To locate the bottleneck we measured oracle recall over the candidate pool that feeds our system's fusion. Figure 4(a) reports the fraction of queries for which a 0.7-IoU hit exists among the top-k pool candidates, together with the live system. The pool contains a 0.7-IoU hit for 82.97% of queries, yet fusion surfaces one for 33.23%. Splitting the gap, a correct candidate is present but mis-ranked for 49.74% of queries (a ranking failure), while the pool has no correct candidate at all for 17.03% (a coverage failure). The dominant limitation is ranking rather than coverage.

Figure 4(b) stratifies the failure modes by ground-truth moment length. Coverage is the dominant failure only for moments shorter than 2 s, where 36.01% of queries have no correct candidate in the pool. For every longer bucket the pool nearly always contains a correct candidate and the loss is almost entirely a ranking problem. This points to two complementary directions. The ranking failures are in part a side effect of fusion: because WBF averages overlapping windows weighted by their scores, a single well-placed candidate can be dragged off-target by a cluster of plausible-but-wrong neighbours, and the fused confidence tracks agreement among members more closely than overlap with the ground truth. A lightweight re-scorer applied after fusion, trained to predict IoU and used only to reorder the surviving windows, would target this directly while leaving the encoder, head, and fusion stage untouched. The sub-2 s coverage gap, in turn, is consistent with the feature frame rate: at 3.2 fps a one-second event spans only three frames, so even a perfect ranker cannot recover boundaries the candidate generator never proposes, and denser features or an onset-driven candidate source would raise the coverage ceiling for this bucket.

4. REFERENCES

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

[2] "DCASE 2026 challenge website," https://dcase.community/challenge2026/, 2026.

[3] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[4] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long audio dataset with captions and temporal boundaries," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[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 Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 23 023–23 033.

[6] J. Lei, T. L. Berg, and M. Bansal, "QVHighlights: Detecting moments and highlights in videos via natural language queries," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[7] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[8] Y. Wu, C. Tsirigotis, K. Chen, C.-Z. A. Huang, A. Courville, O. Nieto, P. Seetharaman, and J. Salamon, "FLAM: Frame-wise language-audio modeling," in *Proc. Int. Conf. on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 267, 2025, pp. 67 719–67 740.

[9] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 119–132.

[10] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 32, pp. 3339–3354, 2024.

[11] T. Nishimura, S. Nakada, H. Munakata, and T. Komatsu, "Lighthouse: A user-friendly library for reproducible video moment retrieval and highlight detection," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 2024, pp. 53–60.

[12] Qwen Team, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2025.

[13] S. Park, J. Choi, K. Baek, and H. Shim, "MomentMix augmentation with length-aware DETR for temporally robust moment retrieval," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2026.

[14] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, p. 104117, 2021.