Encoder-aware Verifier Fusion with Boundary Refinement for Audio Moment Retrieval

DCASE 2026 Challenge Task 6 Technical Report

Mohammad Nur Hossain Khan[1*], Subrata Biswas[1], and Bashima Islam[1]

[1]Worcester Polytechnic Institute , {mkhan, sbiswas, bislam}@wpi.edu
[*]Corresponding author

Abstract

We describe our two submissions to the DCASE 2026 Challenge Task 6: Audio Moment Retrieval from Long Audio. Both build on the UVCOM [2] retriever and add complementary post-processing stages, but they target different encoder families: (i) a **MS-CLAP** [3] pipeline closer to the baseline, combining DAPO RL reranking, confidence-gated swap, and a 2-checkpoint refinement fusion; and (ii) a **LAION-CLAP** [4] pipeline that swaps the audio encoder, applies our *verifier-fusion* principle, and concludes with a single-checkpoint refinement SFT. The core insight is that the highest-quality rerank signal comes not from training a reranker over the retriever's candidates, but from fusing the retriever's candidate scores with an *architecturally independent* predictor (a standalone span-SFT Qwen2.5-Omni-7B) whose predictions have never seen the retriever's ranking. On the CASTELLA test set (public dev-test), our LAION pipeline reaches R1@0.7=32.07 and mAP=27.02 — the strongest configuration in our system grid. The MS-CLAP pipeline reaches R1@0.7=25.32, complementing the LAION submission with a different encoder family and a different refinement variant.

1 Introduction

Language-based Audio Moment Retrieval (AMR) takes a long audio recording and a natural-language query, and predicts the start and end times of the queried event [1]. We address the task at three stages: (a) a DETR-style retriever produces a ranked set of candidate windows; (b) one or more language-model components vote on or refine those windows; (c) an offline combination produces the final top-1 prediction.

Our central methodological observation is about *where the lift comes from* when stacking learned components on top of a strong retriever. Across an extensive system grid we find that:

1. **Encoder choice dominates.** Replacing MS-CLAP with LAION-CLAP [4] on the UVCOM retriever yields +8.26 R1@0.7 on CASTELLA test (28.66 vs. 20.64). BEATs [5] performs comparably to LAION-CLAP; speech-pretrained WavLM-large [6] transfers poorly.

2. **Verifier-fusion beats trained reranking.** A *separately-trained* span-prediction Qwen2.5-Omni-7B [7] (R1@0.7=22.99 standalone, weaker than UVCOM) produces, when its predictions are used as an IoU-overlap vote on UVCOM candidates, an additional +2.72 R1@0.7. This exceeds every DAPO-style reranker we trained (7 variants), all of which saturated near the retriever's top-1.

3. **Refinement is distribution-bound.** A refinement SFT trained to refine UVCOM-produced approximate windows composes well with the verifier-fusion stage (which also

produces UVCOM-flavored windows) but *hurts* when applied to LLM-generated spans from a DAPO reranker.

Our two submissions instantiate these findings with two complementary encoder choices.

2 System overview

Figure **??** summarises both submission pipelines. Both share four building blocks: a UVCOM retriever with LAD [1] matching, a standalone span-SFT Qwen2.5-Omni-7B [7], a DAPO-trained reranker, and a refinement SFT Qwen. The pipelines differ in which blocks they combine and in the audio encoder fed to UVCOM.

2.1 Retriever: UVCOM-LAD

We use the UVCOM [2] architecture with a DETR-style decoder of $Q = 30$ learnable query slots. We modify the Hungarian assignment with a **Length-Aware Decoder (LAD)** term that induces duration specialization across query slots:

$$\widetilde{C}_{q,n} = C_{q,n} + \mathbf{1}[\,\text{bucket}(n) \neq \text{bucket}(q)\,] \cdot \tau, \tag{1}$$

where buckets partition the gold-span duration distribution and $\tau = 10^6$ makes cross-bucket matches infeasible. We use $K = 3$ buckets with bin edges $\{0, 5, 15, \infty\}$ seconds (CASTELLA gold-span median $\approx 4\,\text{s}$). For audio features we evaluate both MS-CLAP [3] (submission 1) and LAION-CLAP HTSAT-tiny [4] (submission 2); text features come from the MS-CLAP text encoder in both cases.

2.2 Reranker (submission 1 only): DAPO with confidence gate

A Qwen2.5-Omni-7B span-SFT adapter is first trained to emit an integer index over the retriever's top-10 candidates ($N \approx 2.1k$ CASTELLA train queries, supervised by gold-IoU). The reranker is then fine-tuned with DAPO [8], a Group-Relative-Policy-Optimization variant with Clip-Higher, dynamic resampling, BNPO loss, and soft over-long filtering. The reward is a length-margin IoU score with a deviation penalty to suppress unjustified picks of distant candidates.

A pick-distribution diagnostic on CASTELLA val showed that DAPO's idx-$\geq$2 picks have a 1:3 help-to-hurt ratio. We add a **confidence gate** at inference: accept idx-1 picks only when $\mathit{score}_{\text{cand}_1} / \mathit{score}_{\text{cand}_0} \geq 0.65$, and reject all idx-$\geq$2 picks. Picks rejected by the gate fall back to a Qwen-span-swap: if the standalone Qwen prediction disagrees with the current top-1 *and* concurs with any UVCOM top-10 candidate (IoU ≥ 0.5), swap to that candidate. This gated+swap submission without refinement reaches CASTELLA R1@0.7=23.76.

2.3 Verifier fusion (submission 2 only)

We observe that a standalone span-SFT Qwen [7], trained on (audio, query) $\rightarrow \langle s, e \rangle$, never sees UVCOM's candidates. Its predictions are therefore *architecturally independent* of UVCOM's ranking. We define a fused score

$$\mathit{final}_c = \mathit{uvcom}_c + \beta \cdot \text{IoU}(c, \hat{y}_{\text{Qwen}}), \tag{2}$$

re-rank the retriever's top-10 by final_c, and take the new top-1. With $\beta = 0.5$, this lifts the LAION-CLAP retriever from R1@0.7=28.66 to 31.62 on CASTELLA test *without any additional training*. We also evaluated multi-verifier extensions (adding a second Qwen-DAPO prediction as a second IoU term); the lift saturates with one *architecturally independent* verifier (any candidate-conditioned LLM trained on the same candidates contributes negligible new signal).

2.4 Refinement SFT

A Qwen2.5-Omni-7B refinement adapter takes (audio, query, approximate window) and outputs a refined $\langle s, e \rangle$. We trained two variants: **v1**, on UVCOM-on-MS-CLAP candidates + perturbed-gold + gold-anchor (12k training rows); and **v2**, on UVCOM-on-LAION-CLAP candidates with the same mix. Both produce a sweep across training steps. By validation loss, ckpt-200 is the best v2 ckpt and ckpts-200/700 the best two v1 ckpts. Submission 1 uses a per-row median fusion of refinement-v1 ckpts-200 and 700; submission 2 uses refinement-v2 ckpt-200 alone.

3 Two submissions

3.1 Submission 1 (MS-CLAP track)

Pipeline: UVCOM-LAD MS-CLAP CASTELLA-finetuned → top-10 candidates → DAPO v5 reranker with confidence gate (§2.2) → Qwen-swap → refinement v1 ckpt-200 *and* ckpt-700, median-fused per row.
Audio encoder: MS-CLAP [3], matching the Munakata et al. baseline [1].
Why this submission: same encoder family as the published baseline, isolates the contribution of {LAD + DAPO+gate + Qwen-swap + refinement} above retriever alone.
CASTELLA test (public dev-test): R1@0.5=37.27, **R1@0.7=25.32**, mAP=19.89, mAP@0.5=31.61.

3.2 Submission 2 (LAION track) — primary

Pipeline: UVCOM-LAD LAION-CLAP CASTELLA-finetuned → top-10 candidates → standalone span-SFT Qwen produces a candidate-blind prediction → verifier fusion at $\beta = 0.5$ (§2.3) → refinement v2 ckpt-200 on top-1.
Audio encoder: LAION-CLAP HTSAT-tiny [4] (512-dim audio features extracted on the DCASE 2026 evaluation audios released to us after request).
Why this submission: the encoder swap delivers the largest single improvement we observed (+8 R1@0.7 over MS-CLAP), and the verifier-fusion technique is the strongest training-free intervention in our system grid.
CASTELLA test (public dev-test): R1@0.5=48.71, R1@0.7=32.07, mAP=27.02, mAP@0.5=43.70.

3.3 Cross-dataset validation: UnAV-100 subset

To verify that submission 2's recipe is not over-tuned to CASTELLA, we evaluated the same pipeline on the UnAV-100 subset [1], a held-out AMR benchmark released alongside CASTELLA. Using the published MS-CLAP audio features and the Clotho-Moment-pretrained UVCOM-LAD (matching the published AM-DETR training regime), our system reaches **R1@0.7=49.45** on UnAV-100, exceeding the published AM-DETR baseline (R1@0.7=39.00) by **+10.45**. With verifier fusion ($\beta = 1.0$), the result rises to R1@0.7=50.55. The LAD + verifier-fusion combination thus generalizes across the CASTELLA (real long-audio) and UnAV-100 (synthetic short-audio) distributions.

4 Experiments and ablations

Table 1 reports our system grid on CASTELLA test.

DAPO + refinement does not compose. Applying refinement on top of the DAPO reranker's output *hurts* R1@0.7 (30.94→30.33). We trace this to a distribution mismatch: refinement was

Table 1: Ablations on CASTELLA test (n=1322 evaluable after filtering missing audios). Pipelines used in our DCASE submissions are **bold**.

Pipeline	R1@0.5	R1@0.7	mAP
Retrievers (UVCOM-LAD, encoder swap)			
UVCOM-LAD MS-CLAP [3]	33.48	20.64	17.27
UVCOM-LAD LAION-CLAP [4]	45.46	28.90	24.47
UVCOM-LAD BEATs [5]	44.33	28.59	23.54
UVCOM-LAD WavLM-large [6]	33.74	20.95	17.32
MS-CLAP pipeline (submission 1)			
+ DAPO+gate+Qwen-swap	37.19	23.76	18.81
+ **Refinement v1 (2-ckpt)**	**37.27**	**25.32**	**19.89**
LAION-CLAP pipeline (submission 2)			
+ Verifier fusion $\beta = 0.5$	49.32	31.62	25.59
+ **Refinement v2 ckpt-200**	**48.71**	**32.07**	**27.02**
Negative results			
LAION + DAPO → Refinement v2	47.05	30.33	26.05
LAION+LAION cross-encoder fusion	42.59	27.53	22.88

trained on UVCOM-produced approximate windows, while DAPO outputs are LLM-generated spans not in UVCOM's top-10. Refinement-on-fusion works because the verifier-fusion stage only re-ranks UVCOM candidates and preserves their distribution; the refinement model sees the same kind of input it was trained on.

Cross-encoder UVCOM score fusion does not work. Naively summing UVCOM-LAION-CLAP and UVCOM-BEATs scores hurts R1@0.7 at every mixture coefficient. Both retrievers share architecture, training data (Clotho-Moment+CASTELLA), and loss; their errors are correlated and score-blending adds noise. This is the negative-result version of the verifier-fusion principle: fusion needs *structural* independence, not just different input features.

5 Implementation details

All retrievers are trained on Clotho-Moment for 200 epochs and then fine-tuned on CASTELLA train+val for $\sim$ 50 epochs. All Qwen2.5-Omni-7B [7] adapters are LoRA (rank 16, $\alpha = 32$, `all-linear` target). The DAPO reranker is trained for 300 steps (2 $\times$ H200 GPUs, $\sim$ 24 h); the refinement SFT runs are 2 epochs ($\sim$ 9 h). Inference at test time uses three Qwen passes per query (standalone span SFT → verifier vote → refinement), which takes $\sim$ 3 h end-to-end on the 177-query DCASE 2026 evaluation set on one H200.

For the DCASE 2026 evaluation set we used the LAION-CLAP feature extractor on the raw audios released to us by the task organizers; the MS-CLAP features and text features were taken directly from the official Zenodo release. Both submissions cover all 177 evaluation queries (no fallback to UVCOM-only required).

6 Conclusion

Our two DCASE 2026 Task 6 submissions reflect a clear methodological finding: the largest lifts in AMR come from *(i)* swapping to a stronger audio encoder (LAION-CLAP [4] or BEATs [5] over MS-CLAP [3]); *(ii)* fusing the retriever's ranking with the IoU vote of a *structurally independent* span predictor (our *verifier-fusion*); and *(iii)* a single refinement-SFT pass over the resulting top-1, trained on the same window distribution the retriever produces. We did not find

further gains from training a candidate-conditioned reranker on top of an already-strong retriever, and refinement does not generalize across window distributions. The LAION-CLAP submission, which combines all three findings, is our primary submission at CASTELLA R1@0.7=32.07.

References

[1] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Language-based Audio Moment Retrieval," *arXiv preprint arXiv:2409.15672*, 2024.

[2] Y. Xiao, J. Yan, H. Chen, R. Lin, X. Liu, T. Sun, "Bridging the Gap: A Unified Video Comprehension Framework for Moment Retrieval and Highlight Detection," *CVPR*, 2024.

[3] B. Elizalde, S. Deshmukh, M. Al Ismail, H. Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," *ICASSP*, 2023.

[4] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, S. Dubnov, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," *ICASSP*, 2023.

[5] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, F. Wei, "BEATs: Audio Pre-Training with Acoustic Tokenizers," *ICML*, 2023.

[6] S. Chen, C. Wang, Z. Chen, et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Topics Signal Process.*, 2022.

[7] Qwen Team, "Qwen2.5-Omni: Technical Report," *arXiv preprint arXiv:2503.20215*, 2025.

[8] ByteDance Seed, "DAPO: An Open-Source LLM Reinforcement Learning System at Scale," *arXiv preprint arXiv:2503.14476*, 2025.