GATED MULTI-FEATURE FUSION FOR DCASE 2026 TASK 6

Technical Report

Kazushi Nakazawa

Advanced Media, Inc.
Tokyo, Japan
k-nakazawa@advanced-media.co.jp

ABSTRACT

This report describes our submission to DCASE 2026 Challenge Task 6, Audio Moment Retrieval from Long Audio. The task is to retrieve the temporal segment in a long audio recording that best matches a natural-language query, and systems are ranked primarily by `recall1@0.7`. Our approach is a DETR-style audio moment retrieval model that combines frozen audio and text representations through a lightweight gated fusion layer. Although residual reranking and external audio-language verifier scores improved some validation runs, a final sweep on the CASTELLA development-testing split selected four non-reranked checkpoint variants for submission. The best submitted system combines MS-CLAP, LAION-CLAP, BEATs, EAT, M2D-CLAP PCA, and VAD audio features with MS-CLAP and LAION-CLAP text features. It achieved `recall1@0.7` = 30.14, `recall1@0.5` = 49.29, `mAP` = 24.27, and `mAP@0.75` = 22.71 on CASTELLA development-testing. For the hidden evaluation set, whose query file contains 177 queries over 100 recordings, we generated and format-validated four output files.

***Index Terms*—** Audio moment retrieval, long audio, DCASE 2026, multi-feature fusion, gated fusion

1. INTRODUCTION

Audio moment retrieval (AMR) aims to localize, in a long audio recording, the segment described by a free-form text query [1]. Compared with clip-level audio-text retrieval, AMR additionally requires temporal boundary estimation. Compared with conventional sound event detection, it is not limited to a fixed event vocabulary: queries can describe sound sources, actions, acoustic scenes, or temporal relations in natural language. DCASE 2026 Task 6 evaluates this setting with temporal-IoU-based ranking metrics, using `recall1@0.7` as the primary metric [2].

Our system follows a DETR-style set-prediction framework [3] adapted to language-based moment retrieval in video and audio [4, 5, 1, 6]. The main empirical finding from development was that pretrained representation diversity is helpful, but only when the fusion and selection rules remain conservative. Features from EAT [7], BEATs [8], CLAP-style models [9, 10, 11], M2D-CLAP [12, 13], PaSST [14], and RoBERTa [15] were useful in different configurations, whereas high-capacity fusion and aggressive reranking were less stable. We therefore submitted four checkpoint variants built around simple gated multi-feature fusion.

2. TASK SETTING AND DATA

Given an audio recording and a query, the system outputs one or more temporal windows. For `recall1@0.7`, only the top-ranked window is judged, and it is counted as correct when its IoU with a ground-truth moment is at least 0.7. We also report `recall1@0.5`, `mAP`, and `mAP@0.75` as diagnostic metrics.

We used the DCASE 2026 Task 6 development data. They consist of Clotho-Moment, a large synthetic dataset constructed by overlaying Clotho audio-caption pairs [16] onto background recordings [1], and CASTELLA, a manually annotated long-audio dataset derived from AudioCaps source videos [17, 18]. In our processed CASTELLA split, there were 2182 training queries, 352 validation queries, and 1347 test queries. Following the task terminology, we refer to CASTELLA validation as development-validation and CASTELLA test as development-testing [2, 18]. The official evaluation data contain 100 recordings, but their ground-truth windows are hidden; consequently, official evaluation metrics cannot be computed locally. For each submitted system, we produced one format-validated output file using the required `[start, end]` window format.

We did not use visual information from source videos, manually inspect or annotate evaluation samples, or call closed LLM APIs. Optional metadata were not used for model training, checkpoint selection, or reranking; only identifiers, durations, and query text were used for inference and output formatting. All pretrained resources used by the submitted inference systems are on the task webpage's allowed-resource list [2]: MS-CLAP [9], LAION-CLAP [11], BEATs [8], EAT [7], M2D-CLAP [13], PaSST [14], and RoBERTa [15]. Audio Flamingo Next [19] was used only during development as an external verifier and is excluded from the submitted inference pipeline and metadata parameter counts. All pretrained models were frozen, and VAD features were computed by our own signal-processing front end without an external pretrained model.

3. BASE RETRIEVAL MODEL

3.1. Feature Groups

The model operates on precomputed audio and text features. Audio is represented as a sequence of one-second time steps, up to 300 steps per recording. Each feature group is linearly projected to the model dimension and then fused by the gated module described below. Across the four submissions, the audio features are MS-CLAP [9, 10], LAION-CLAP [11], BEATs [8], EAT [7], PCA-compressed M2D-CLAP [12, 13], PCA-compressed PaSST [14],

Table 1: Feature groups used by at least one submitted system. A: audio, T: text.

Mod.	Feature	Dim.
A	MS-CLAP	768
A	LAION-CLAP	512
A	BEATs	768
A	EAT	768
A	M2D-CLAP PCA	512
A	PaSST PCA	512
A	VAD	8
T	MS-CLAP	768
T	LAION-CLAP	512
T	RoBERTa	768

Table 2: Representative training hyperparameters.

Setting	Value	Setting	Value
Epochs	200	Batch	32
Eval batch	32	Optimizer	AdamW
LR	1×10^{-4}	Weight decay	1×10^{-4}
LR drop	150	Grad. clip	0.1
Enc. layers	2	Dec. layers	2
Hidden dim.	256	Heads	8
FFN dim.	1024	Dropout	0.5
Clip length	1 s	Max clips	300
Max query	32	Max labels	5
Query slots	20	Seed	2023

and VAD. The text features are MS-CLAP, LAION-CLAP, and RoBERTa [15]. Table 1 lists the feature groups used by at least one submitted system. M2D-CLAP and PaSST are compressed to 512 dimensions with PCA before fusion to reduce dimensionality and improve training stability.

3.2. Gated Fusion

Let x_i be the temporal sequence for the i-th audio feature group, and let c_{text} denote the query context. We compute a scalar gate for each group from pooled audio statistics and the text context:

$$g_i = \sigma\left(f_i\left(\text{pool}\left(x_i\right), c_{\text{text}}\right)\right), \qquad \tilde{x}_i = g_i x_i. \tag{1}$$

The gated streams are concatenated and projected before the Transformer encoder-decoder [20]. This low-capacity design lets the model suppress unreliable streams without introducing a large controller. In our local experiments, bottleneck gating, query-conditioned gating, second-stage refinement, and group-wise dropout did not consistently improve the selection metric.

3.3. Candidate Generation and Training

The retrieval backbone uses direct set prediction [3, 5]. A fixed set of learned query slots attends to the fused audio-text representation and predicts temporal windows with confidence scores. Window coordinates are estimated in normalized time and converted back to seconds for evaluation. All submitted systems use `num_queries=20`. Following DETR-style moment retrieval baselines [1, 4], training combines L_1 and generalized IoU [21] losses for boundary regression with a cross-entropy loss for candidate confidence, plus the implementation's auxiliary saliency and quality/rerank-head losses.

The model was optimized with AdamW [22]; Table 2 shows the representative configuration.

4. MODEL SELECTION

Initial experiments on development-validation explored feature fusion, conservative residual reranking, and single external verifier scores. Small residual corrections based on pairwise candidate features and verifier similarities from M2D-CLAP [13], MS-CLAP [10], and Audio Flamingo Next [19] improved some validation checkpoints. However, a broader sweep on CASTELLA development-testing selected a simpler M2D-PCA512+VAD base checkpoint over the reranked variants. We therefore submitted four base checkpoints without external-verifier reranking.

The sweep covered 30 representative candidates, including validation-frontier checkpoints and additional high-validation models that had not yet been evaluated on development-testing. Systems were ordered by development-testing `recall1@0.7`. For the fourth submission, a tie in `recall1@0.7` was resolved using `recall1@0.5` and `mAP`. This procedure was used only to choose among public development candidates; the hidden evaluation set remains the final benchmark.

5. SUBMITTED SYSTEMS

Table 3 summarizes the four submitted systems and their CASTELLA development-testing scores. Task6_1 and Task6_3 share the M2D-PCA512+VAD architecture but use different checkpoints. Task6_2 is the most feature-rich model, adding PaSST and RoBERTa. Task6_4 is a lighter EAT-centered alternative. All four systems outperform the official baseline on development-testing, whose best `recall1@0.7` is 13.59 [2, 1].

6. DISCUSSION

The development-testing sweep changed the final selection: validation-favorable residual reranking was replaced by simpler base checkpoints. This suggests that external verifier scores can be useful diagnostics, but should be used cautiously when the selection split changes.

The M2D-PCA512+VAD models were especially strong on high-IoU metrics. The all-feature model remained competitive and achieved the second-best `recall1@0.7`, but adding PaSST and RoBERTa did not surpass the best M2D-PCA512+VAD checkpoint. The EAT-centered model was retained as a diverse fourth candidate because it tied the next M2D-PCA512+VAD checkpoint on `recall1@0.7` and achieved a higher `recall1@0.5`.

The negative results were also informative. Raw M2D features without PCA compression, EAT+VAD without stronger complementary features, more complex gates, direct learned reranking, and naive score averaging did not reliably improve the target metric. Overall, the results support a conservative design: combine diverse frozen representations, but keep fusion and selection simple.

7. CONCLUSION

We submitted four DETR-style AMR systems based on lightweight gated multi-feature fusion. The best system achieved `recall1@0.7` = 30.14 on CASTELLA development-testing by combining CLAP-style, BEATs, EAT, M2D-CLAP PCA, and

Table 3: Submitted systems and CASTELLA development-testing results (%). System indices follow the labels `Nakazawa_AM_task6_{1--4}`. The baseline rows reproduce the official Task 6 baseline results [2].

Label	Description	recall1@0.5	recall1@0.7	mAP	mAP@0.75
Baseline	CASTELLA only	23.16	10.32	9.11	6.96
Baseline	CASTELLA + Clotho-Moment	25.61	13.59	12.06	10.72
Task6_1	M2D-PCA512+VAD, `best_map_075` checkpoint	49.29	**30.14**	24.27	22.71
Task6_2	All-feature gated model, raw best checkpoint	47.07	29.47	23.79	22.07
Task6_3	M2D-PCA512+VAD, `best_map` checkpoint	49.59	29.03	23.73	21.96
Task6_4	EAT-centered `num_queries=20` gated model	49.29	28.88	22.76	21.02

Table 4: Parameter counts reported in the metadata. Frozen parameters correspond to the feature extractors used by each submitted system.

Label	Trainable	Frozen
Task6_1	9.89 M	587.91 M
Task6_2	10.75 M	798.71 M
Task6_3	9.89 M	587.91 M
Task6_4	9.62 M	498.87 M

VAD audio features with CLAP-style text features. We also generated format-validated outputs for all four systems on the hidden evaluation set; the official scores will be computed by the organizers.

Acknowledgment

AI assistance was used only for English editing and drafting support for this report, not for model training, inference, model selection, or annotation. All experiments, metadata values, selection decisions, and final technical claims were reviewed and verified by the author.

8. REFERENCES

[1] 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.

[2] DCASE Community, "DCASE 2026 Challenge Task 6: Audio moment retrieval from long audio," https://dcase.community/challenge2026/task-audio-moment-retrieval-from-long-audio, 2026, accessed: 2026-06-15.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, vol. 12346. Springer, 2020, pp. 213–229.

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

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

[6] 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. Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 2024, pp. 53–60. [Online]. Available: https://aclanthology.org/2024.emnlp-demo.6/

[7] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. International Joint Conferences on Artificial Intelligence Organization, 2024, pp. 3807–3815. [Online]. Available: https://doi.org/10.24963/ijcai.2024/421

[8] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 5178–5193. [Online]. Available: https://proceedings.mlr.press/v202/chen23ag.html

[9] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[10] 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.

[11] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," 2023.

[12] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Learning representations by encouraging both networks to model the input," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[13] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, and K. Imoto, "M2D-CLAP: Masked modeling duo meets CLAP for learning general-purpose audio-language representation," in *Proc. Interspeech*, 2024, pp. 57–61.

[14] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. Interspeech*, 2022, pp. 2753–2757.

[15] 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," 2019.

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

[17] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. Conf. North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019, pp. 119–132. [Online]. Available: https://aclanthology.org/N19-1011/

[18] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long audio dataset with captions and temporal boundaries," 2026.

[19] S. Ghosh, A. Goel, K. Jayakumar, L. Koroshinadze, N. Anand, Z. Kong, S. Gururani, S.-g. Lee, J. Kim, A. Aljafari, C.-H. H. Yang, S. Kim, R. Duraiswami, D. Manocha, M. Shoeybi, B. Catanzaro, M.-Y. Liu, and W. Ping, "Audio flamingo next: Next-generation open audio-language models for speech, sound, and music," 2026.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[21] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 658–666.

[22] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7