Exploring Pretrained Audio-Text Encoders for Audio Moment Retrieval: DCASE 2026 Task 6

Snehit B. Chunarkar[1], *Krishnagiri Hamza*[2], *Chi-Chun Lee*[1],

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Electronics and Communication Engineering, SRM University, Chennai, India
snehit@gapp.nthu.edu.tw, hk1583@srmist.edu.in, cclee@ee.nthu.edu.tw

***Abstract*—We present our system submitted to DCASE 2026 Task 6: Audio Moment Retrieval (AMR), which aims to retrieve a temporally grounded moment within a long audio recording given a natural language query. Following the AM-DETR baseline framework, we adopt frame-level audio feature extraction by segmenting long recordings into non-overlapping one-second clips, yielding a temporally ordered sequence of clip-level embeddings. Building on this, our primary contribution is a systematic investigation of pretrained audio and text encoders as replacements for the baseline MS-CLAP features, including M2D-CLAP and LAION-CLAP for both audio and text. The selected feature representations are fused into an ensemble and fed to an AM-DETR-based retrieval head for temporal boundary regression. We further incorporate frame masking during training and an IoU-based loss to improve localisation. Our system achieves a Recall1@0.7 score of 26.43 on the CASTELLA test split, surpassing the baseline score of 13.59.**

***Index Terms*—audio moment retrieval, temporal grounding, ensemble learning**

1. INTRODUCTION

The ability to locate a specific event within a long, unstructured audio recording given only a natural language description is a practically important yet under-explored problem. Audio Moment Retrieval (AMR) [1] addresses this challenge by requiring a system to predict the start and end timestamps of the moment within a long audio clip that best corresponds to a free-form text query. Unlike conventional audio retrieval, which ranks short, pre-segmented clips, AMR operates on recordings that can span several minutes, demanding both fine-grained temporal localization and robust cross-modal understanding.

Interest in AMR is motivated by a wide range of real-world applications, including surveillance monitoring, sports broadcast indexing, meeting summarization, and accessibility tooling for audio archives. Despite its relevance, research in AMR remains limited compared to its visual counterpart, Video Moment Retrieval (VMR) [2], primarily due to the absence of suitable large-scale datasets and the difficulty of encoding temporal context in long audio signals.

DCASE 2026 Task 6 formalizes AMR as a community challenge, using the recently introduced CASTELLA dataset [3] for evaluation and the AM-DETR framework [1] as a reference baseline. The baseline employs MS-CLAP [4] for joint audio-text feature extraction alongside a DETR-based [5] moment prediction head. While this establishes a solid foundation, the choice of pretrained encoder is known to have a significant impact on downstream performance in audio-language tasks, and the baseline leaves this dimension largely unexplored.

In this work, we systematically investigate the effect of substituting pretrained audio and text encoders within the AM-DETR framework. We evaluated M2D-CLAP [6], LAION-CLAP [7], and BEATs [8] as audio encoders, and M2D-CLAP, LAION-CLAP, and T5 [9] as text encoders, both individually and in ensemble configurations. we further introduce frame masking during training and an IoU-based auxiliary loss to improve temporal boundary regression. Our best system, an ensemble of M2D-CLAP and LAION-CLAP features under a two-stage training strategy, achieves a Recall1@0.7 score of 26.43 on the CASTELLA test split, nearly doubling the two-stage baseline score of 13.59.

The remainder of this paper is organised as follows. Section 2 describes the datasets used. Section 3 details the proposed methodology. Section 4 outlines the experimental setup and training strategy. Section 5 presents and analyses the results, followed by the conclusion in Section 6.

2. DATASET

We used two datasets: Clotho-Moment [1] for pretraining and CASTELLA [3] for fine-tuning, as described below.

2.1. Clotho-Moment

Clotho-Moment is a large-scale synthetic dataset containing audio recordings with paired text queries and annotated temporal boundaries. It was constructed by overlaying audio clips from the Clotho [10] dataset as foreground events onto ambient audio from the Walking Tour [11] dataset as background, simulating realistic audio scenes with moments of interest. The dataset comprises 32,694, 4,918, and 6,649 samples for training, validation, and test split, respectively. We use the training and validation splits for pre-training the AM-DETR prior to fine-tuning on CASTELLA.

2.2. CASTELLA

CASTELLA is a human-annotated dataset of long audio recordings ranging from one to five minutes, each paired with free-format natural language captions and precise temporal boundary annotations. It provides 1,009, 213, and 640 recordings for training, validation, and test splits, respectively. We use the training and validation splits for fine-tuning and the test split to compute all final scores reported in Table 2.

3. METHODOLOGY

Our system follows the AM-DETR pipeline [1], replacing the baseline MS-CLAP encoder with a selection of stronger pretrained models and introducing complementary training improvements. The overall architecture is illustrated in Figure 1.

3.1. Audio pre-processing

We adopt the frame-level preprocessing strategy from [1], visualised in Figure 1 (a). Long recordings are segmented into non-overlapping one-second clips, and the per-clip features extracted by the audio encoder are stacked into a temporally ordered two-dimensional feature matrix. This yields one feature vector per clip: 60 to 300 vectors for recordings of one to five minutes, preserving the temporal structure of the recording rather than collapsing it into a single global embedding. To explicitly encode temporal position, we append a two-dimensional positional vector to each clip embedding, whose values increase linearly from 0 to 1 over the total number of frame-level clips.

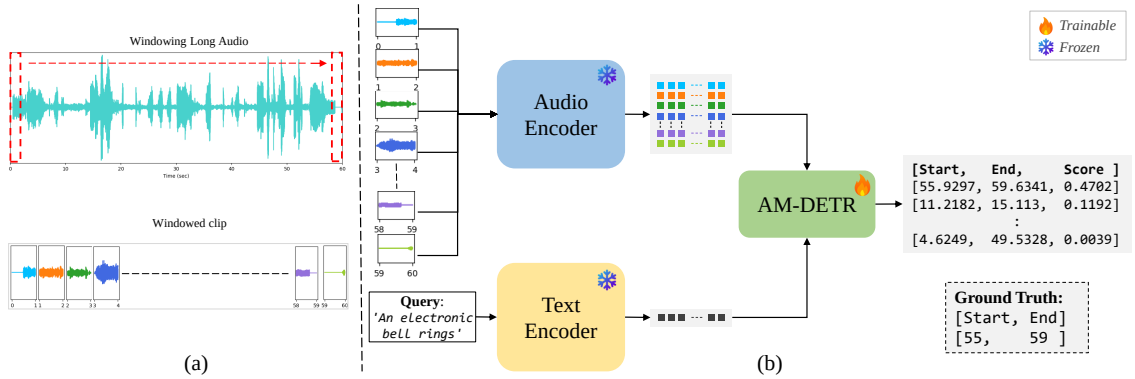

Fig. 1: (a) Audio preprocessing: Windowing long audio, (b) Architecture: Audio Moment DETR

3.2. Pretrained models

We evaluate the following pretrained encoders as alternatives to the MS-CLAP encoder used in the baseline.

M2D-CLAP [6]: It combines the self-supervised learning (SSL) Masked Modelling Duo (M2D) framework applicable to 2D structured data such as audio spectrograms with Contrastive language-audio pre-training (CLAP), producing a general-purpose audio-language representation that supports both transfer learning and zero-shot tasks. M2D-CLAP model comes with ∼89M parameters, and it operates on audio sampled at $16kHz$.

LAION-CLAP [7]: The model is trained with the contrastive learning paradigm between the audio and text embeddings in pairs on the large-scale `LAION-Audio-630K` dataset, one of the largest publicly available audio-caption datasets. LAION-CLAP model has ∼158M parameters and is designed to accept audio sampled at $48kHz$

BEATs [8]: It is an iteratively trained audio SSL model that jointly optimises an acoustic tokenizer and a Transformer encoder for rich semantic audio representations, achieving strong performance on audio classification benchmarks. All variants of the BEATs transformer encoder come in a ∼90M-parameter version; we used the `BEATs_iter3+ (AS2M)` version of trained weights from BEATs.

T5 [9]: It is a sequence-to-sequence language model. We use only the encoder block of the `T5-Base` variant (∼109M parameters) from the T5 family to extract contextual text embeddings from natural language queries.

3.3. Architecture

The proposed architecture (Source code available [1]) is presented in Figure 1 (b) and consists of two major components.

3.3.1. Encoders: Audio and text encoders are kept frozen throughout all experiments; they are used solely for offline feature extraction (Extracted features are available for both Clotho-Moment [2] and CASTELLA [3]). Audio features are extracted independently by M2D-CLAP, LAION-CLAP, and BEATs. Text features are extracted independently by M2D-CLAP, LAION-CLAP and T5. In ensemble configurations, features from multiple encoders are concatenated along the feature dimension before being passed to the downstream head, requiring only architectural changes to the audio and text projection blocks' first-layer input dimension as per the concatenated length of features in AM-DETR, Table 1 lists AM-DETR's parameters for respective encoder(s).

[1]Code: https://github.com/Snehitc/AMR-encoder-exploration
[2]Features (Clotho-Moment): https://zenodo.org/records/20770460
[3]Features (CASTELLA): https://zenodo.org/records/20772071

Table 1: AM-DETR's trainable parameters for respective encoder combination.

ID	Encoder(s)	Trainable Param (AM-DETR)
M1	MS-CLAP	7.18 M
M2	M2D-CLAP	7.18 M
M3	LAION-CLAP	7.05 M
M4	BEATs, T5	7.18 M
M5	(M2D, LAION)-CLAP	7.45 M
M6	M2D-CLAP, BEATs, T5	7.58 M
M7	LAION-CLAP, BEATs, T5	7.45 M
M8	(M2D, LAION)-CLAP, BEATs, T5	7.84 M

3.3.2. Downstream Head: We adopted the AM-DETR from the baseline [1] as the downstream moment retrieval head model. AM-DETR is inspired by QD-DETR [2] and builds upon the DETR [5] architecture. It encodes the cross-modal interaction between the audio feature sequence and the text query embedding via a cross-attention Transformer, then passes the fused representation through a Transformer decoder that predicts K candidate moment proposals as (center, width) pairs together with confidence score. The model is trained with a combination of L1 loss, generalized IoU (gIoU) loss and score cross-entropy loss.

4. EXPERIMENTS

For all experiments, encoders are kept frozen, and features are extracted offline to reduce training overhead. For each audio recording, the preprocessing step processes a 2D feature matrix of shape $T \times D$, where $T \in [60, 300]$ is the number of one-second clips and D is the encoder output dimension. This matrix is saved to disk and used directly as input to AM-DETR, avoiding redundant re-computing across training runs.

4.1. Training settings

During training, 15% of the clip-level audio feature vectors are randomly masked to improve robustness to partial or noisy audio inputs. All experiments share the following hyperparameters: Adam optimiser with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}; training is set to 200 epochs. These settings are consistent across both pre-training and fine-tuning stages.

4.2. Two Stage Training

Following the training strategy of the DCASE Challenge Task 6 baseline [12], we employ a two-stage training procedure.

4.2.1. Pre-Training: AM-DETR is first pre-trained on the training and validation splits of Clotho-Moment [1]. The synthetic nature and large scale of this dataset allow the model to learn robust moment localization priors before exposure to the more challenging real-world recordings in CASTELLA.

Table 2: Experiment results on CASTELLA test split.

ID	Audio Encoder	Text Encoder	CASTELLA only					Two Stage Training PT: Clotho-Moment, FT: CASTELLA				
			Recall1 ↑		mAP ↑			Recall1 ↑		mAP ↑		
			@0.5	@0.7	avg	@0.5	@0.75	@0.5	@0.7	avg	@0.5	@0.75
M1	Baseline: MSCLAP	Baseline: MSCLAP	23.16	10.32	9.11	20.34	6.96	25.61	13.59	12.06	23.60	10.72
M2	M2D-CLAP	M2D-CLAP	26.73	12.32	10.39	22.98	8.81	**43.88**	25.39	19.40	**36.89**	17.50
M3	LAION-CLAP	LAION-CLAP	19.45	7.42	6.86	16.72	4.97	38.08	21.97	16.38	31.47	15.09
M4	BEATs	T5	18.63	8.02	7.01	15.98	5.28	34.45	19.97	15.64	29.77	14.18
Ensemble												
M5	(M2D, LAION)-CLAP	(M2D, LAION)-CLAP	30.44	13.81	11.47	25.23	9.20	43.21	**26.43**	**19.49**	35.94	**17.96**
M6	M2D-CLAP, BEATs	M2D-CLAP, T5	28.14	11.73	10.80	24.22	8.71	42.09	25.17	19.00	36.47	16.69
M7	LAION-CLAP, BEATs	LAION-CLAP, T5	19.97	7.28	7.15	17.51	5.37	40.61	22.79	17.47	33.47	15.39
M8	(M2D, LAION)-CLAP, BEATs	(M2D, LAION)-CLAP, T5	**39.20**	**18.63**	**15.79**	**34.46**	**13.16**	41.65	24.57	19.01	35.37	18.13

4.2.2. Fine-tuning: The pre-trained model is subsequently fine-tuned on the training and validation splits of CASTELLA [3]. The temporal positional encoding a linearly spaced two-dimensional vector appended to each clip embedding, is included consistently in both stages, ensuring that the model's positional assumptions remain compatible across the domain shift from synthetic to real-world audio.

4.3. CASTELLA only

To isolate the benefit of pre-training, we also train each encoder configuration directly on CASTELLA without the Clotho-Moment pre-training stage. This controls for encoder quality independently of the training strategy and allows a direct comparison of the encoder contributions under a fixed, simpler training regime.

5. RESULTS

All the results are reported on the CASTELLA test split and sumarized in Table 2. The primary ranking metric for DCASE Task 6 is Recall1@0.7 (R1@0.7); We additionally report R1@0.5 and mean Average Precision (mAP) at thresholds of 0.5 and 0.75. For visualisation, we also report the R1@0.7 for all the models in Figure 2.

5.1. Single encoders

Across both training settings, M2D-CLAP (M2) is consistently the strongest single-encoder configuration. With two-stage training, M2 achieves an R1@0.7 of 25.39, a gain of 11.8 points over the two-stage MS-CLAP baseline (M1: 13.59). LAION-CLAP (M3) and BEATs+T5 (M4) trail M2D-CLAP but both exceed the baseline under two-stage training, indicating that the choice of the pretrained encoder is a primary factor in system performance.

5.2. Ensemble configurations

Ensembling consistently improves R1@0.7 in the two-stage setting. The M2D-CLAP + LAION-CLAP ensemble (M5) achieves the best overall R1@0.7 of 26.43 with two-stage training, and also outperforms all systems on mAP (avg) and mAP@0.75 metrics. However, extending the ensemble to include BEATs and T5 (M8) does not further improve two-stage performance, suggesting that BEATs and T5 encode information that is already well-covered by the two CLAP-based models in this setting.

5.3. CASTELLA-only results

In the CASTELLA-only setting, M8 (all four encoders) is the strongest configuration with an R1@0.7 of 18.63, while M2D-CLAP alone achieve only 12.32. This reversal, where a broader encoder ensemble is more beneficial without pre-training, suggests that the diverse representation from BEATs and T5 helps compensate for the limited CASTELLA training data, while pre-training on Clotho-Moment reduces the marginal benefit of that diversity.

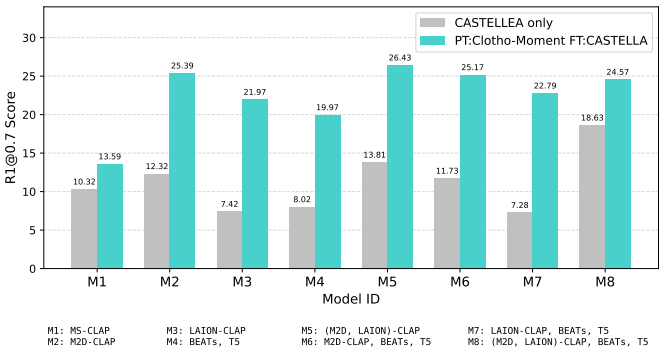

Fig. 2: Model Performance on R1@0.7 Score: CASTELLA only vs PT: Clotho-Moment FT: CASTELLA

5.4. Effect of pre-training

Pre-training on Clotho-Moment provides a substantial and consistent improvement for every encoder configuration. For M2D-CLAP alone, two-stage training more than doubles R1@0.7 from 12.32 to 25.39. The improvement is similarly pronounced for LAION-CLAP and ensemble variants, confirming that exposure to the large synthetic Clotho-Moment dataset is critical for generalising to real-world long audio in CASTELLA.

6. CONCLUSION

We presented a systematic investigation of pretrained audio and text encoders within the AM-DETR framework for Audio Moment Retrieval, submitted to DCASE 2026 Task 6. Our experiments demonstrate that the choice of the pretrained encoder significantly impacts retrieval performance: M2D-CLAP emerges as the strongest single encoder, while an ensemble of M2D-CLAP and LAION-CLAP achieves the best overall result with an R1@0.7 score of 26.43 on the CASTELLA test split, nearly doubling the strong two-stage baseline of 13.59. We further show that two-stage training with Clotho-Moment pre-training is consistently beneficial across all encoder configurations, and that frame masking during training provides additional robustness. Notably, the relative benefit of including diverse encoders such as BEATs and T5 is more pronounced when pre-training is absent, suggesting a complementarity between data augmentation via pre-training and feature augmentation via ensemble diversity. Future work may explore fine-grained encoder fusion strategies and lightweight adapter-based fine-tuning of frozen encoders.

REFERENCES

[1] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Language-based audio moment retrieval," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2025, pp. 1–5.

[2] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-dependent video representation for moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 023–23 033.

[3] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long audio dataset with captions and temporal boundaries," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 15 352–15 356.

[4] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2024, pp. 336–340.

[5] 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*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 213–229.

[6] D. Niizumi, D. Takeuchi, M. Yasuda, B. T. Nguyen, Y. Ohishi, and N. Harada, "M2D-CLAP: Exploring General-purpose Audio-Language Representations Beyond CLAP," *IEEE Access*, vol. 13, pp. 163 313–163 330, 2025.

[7] Y. Wu*, K. Chen*, T. Zhang*, Y. Hui*, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[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, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193. [Online]. Available: https://proceedings.mlr.press/v202/chen23ag.html

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[10] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[11] S. Venkataramanan, M. N. Rizve, J. Carreira, Y. M. Asano, and Y. Avrithis, "Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=Yen1lGns2o

[12] DCASE Challenge: Task 6 (Website), "Audio moment retrieval from long audio," https://dcase.community/challenge2026/task-audio-moment-retrieval-from-long-audio.