GISP@HEU'S SUBMISSION FOR DCASE 2026 TASK 6: FREQUENCY-AWARE CROSS-MODAL FUSION FOR AUDIO MOMENT RETRIEVAL

Technical Report

Feiyang Xiao[1†], *Li'ang Luo*[1†], *Kejia Zhang*[1], *Qiaoxi Zhu*[2], *Guangjun He*[3], *Pengming Feng*[3]
Wenwu Wang[4], *and Jian Guan*[1*]

[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]University of Technology Sydney, Ultimo, Australia
[3]State Key Laboratory of Space Information System and Integrated Application (SISIA), Beijing, China
[4]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK

ABSTRACT

This technical report presents GISP@HEU's submission for DCASE 2026 Task 6 audio moment retrieval, which aims to retrieve corresponding segments in long audio recordings based on the content-semantic correlation between audio and text queries. In our submission, we describe four systems built upon the UVCOM framework, focusing on improving the cross-modal fusion process between audio and text features.

***Index Terms*—** Audio moment retrieval, audio-text representation, frequency attention

1. INTRODUCTION

Audio Moment Retrieval (AMR) is a recently proposed audio-language task that aims to retrieve temporal moments in long audio recordings according to a natural-language query [1]. Unlike conventional audio retrieval, which retrieves short audio clips from a database, AMR requires precise localization of relevant segments within untrimmed audio streams [2]. The fundamental mechanism enabling this retrieval lies in measuring semantic consistency: by mapping both audio and text into a shared embedding space, the system can identify and localize audio segments that exhibit high semantic alignment with the query [3, 4]. The task has attracted increasing attention due to its potential applications in multimedia indexing, surveillance analysis, audio archive search, and content understanding [5]. Recent benchmark datasets such as Clotho-Moment [1] and CASTELLA [2] have facilitated the development and evaluation of AMR systems by providing temporally annotated audio-language pairs.

Existing AMR approaches are largely inspired by video moment retrieval (VMR) [2, 6]. Early methods, such as Audio Moment DEtection TRansformer (AM-DETR), adopt DETR-based architectures to model temporal dependencies and cross-modal interactions between audio and text [1]. More recently, Unified Video COMprehension (UVCOM) framework has achieved strong retrieval performance by introducing a unified video comprehension framework that effectively models multimodal representations and temporal grounding [7]. The success of these approaches highlights the importance of robust cross-modal fusion mechanisms.

†These authors contributed equally to this work.
*Corresponding author.

In this work, we develop three UVCOM-based systems that enhance cross-modal fusion from different perspectives. System 1 incorporates frequency attention and an efficient discriminative feed-forward network (EDFFN) to strengthen frequency-aware representation learning. System 2 investigates the effectiveness of frequency attention alone. System 3 introduces multi-channel frequency decomposition to capture complementary frequency information during audio-text interaction. In addition, system 4 is an ensemble system by combining the outputs of the three individual models. The resulting four systems constitute our submissions to DCASE 2026 Task 6.

2. SUBMISSION SYSTEMS

2.1. System 1: EDFFN-based

System 1 extends the original UVCOM framework by redesigning the cross-modal fusion module. A frequency attention mechanism is introduced to explicitly model frequency-dependent correlations between audio and textual representations, enabling the model to emphasize acoustically informative frequency regions. Meanwhile, an efficient discriminative feed-forward network (EDFFN) is employed in this operation to improve feature discrimination and representation capacity after multi-modal interaction. The design aims to achieve more accurate audio-text alignment and temporal localization.

2.2. System 2: Frequency-based

System 2 focuses on evaluating the contribution of the frequency attention mechanism independently. The original UVCOM cross-modal fusion module is expended with only frequency attention. By selectively enhancing frequency-aware interactions between audio and text features, this system investigates whether frequency-domain attention alone can improve retrieval performance without introducing additional feed-forward enhancements.

2.3. System 3: Multi-Channel Frequency Decomposition-based

System 3 introduces a multi-channel frequency decomposition strategy into the UVCOM framework. Audio features are decomposed into multiple frequency channels, allowing the model to learn complementary representations from different frequency bands. The

Table 1: Performance comparison between our systems and the official baseline on the test set of CASTELLA dataset [2].

Method	R1@50%	R1@70%	mAP-avg	mAP@50%	mAP@70%
Official baseline [2]	25.61	13.59	12.06	23.60	10.72
System 1	40.83	28.21	20.55	35.11	20.00
System 2	**41.13**	30.44	22.32	35.47	22.25
System 3	40.31	28.36	21.50	35.23	21.15
System 4	41.05	**31.55**	**24.63**	**37.79**	**24.01**

decomposed features are subsequently integrated within the cross-modal fusion process, enabling finer-grained modeling of audio events and improving the correspondence between audio content and textual queries.

2.4. System 4: Ensemble System

System 4 is an ensemble model that combines the predictions of Systems 1-3. Since the three individual systems capture complementary aspects of frequency-aware audio-text interactions, their outputs are aggregated to obtain more robust retrieval results. The ensemble strategy reduces the variance of individual models and leverages their complementary strengths, leading to improved overall retrieval performance.

3. EXPERIMENTAL RESULTS

Following the open source experimental setting in DCASE 2026 Task 6 official website, we compare our systems with the official baseline on the test set of CASTELLA dataset [2]. Results are shown in Table 1.

Table 1 compares the proposed systems with the official baseline on the CASTELLA test set. All proposed systems consistently outperform the baseline across all evaluation metrics, demonstrating the effectiveness of frequency-aware cross-modal fusion for audio moment retrieval. Among the three single systems, System 2 achieves the best performance, obtaining 41.13% R1@50%, 30.44% R1@70%, and 22.32% mAP-avg, which correspond to improvements of 15.52, 16.85, and 10.26 percentage points over the official baseline, respectively. Furthermore, the ensemble model (System 4) delivers the strongest overall results, achieving 31.55% R1@70%, 24.63% mAP-avg, 37.79% mAP@50%, and 24.01% mAP@70%. These findings indicate that frequency-aware modeling effectively enhances audio-text alignment, while combining multiple complementary systems further improves retrieval performance and robustness.

4. CONCLUSION

In this technical report, we presented four systems for DCASE 2026 Task 6 Audio Moment Retrieval. Built upon the UVCOM framework, the proposed systems enhance cross-modal fusion through frequency-aware modeling. Experimental results on the CASTELLA test set demonstrate that all proposed systems consistently outperform the official baseline, showing improvement by the use of frequency-domain information for audio-text alignment and temporal localization.

5. ACKNOWLEDGMENT

This work was partly supported by the Heilongjiang Provincial Natural Science Foundation of China under Grant No. BS2025F009.

6. 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] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long audio dataset with captions and temporal boundaries," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2026, pp. 336–340.

[3] F. Xiao, J. Guan, Q. Zhu, X. Liu, W. Wang, S. Qi, K. Zhang, J. Sun, and W. Wang, "A reference-free metric for language-queried audio source separation using contrastive language-audio pretraining," in *Proc. Detect. Classif. Acoust. Scenes Events Workshop (DCASE)*, Tokyo, Japan, October 2024, pp. 191–195.

[4] F. Xiao, X. Feng, T. Ye, K. Zhang, H. Lan, G. He, P. Feng, Q. Zhu, and J. Guan, "ME-CLAPScore: Modeling semantic alignment and mismatch for audio-text relevance assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2026.

[5] 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.

[6] 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. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 23 023–23 033.

[7] 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," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 18 709–18 719.