Audio Moment Retrieval from Long Audio for DCASE 2026 task 6

Technical Report

Wei-Yu Chen, Chung-Li Lu

Advanced Technology Laboratory, Telecommunication Laboratories,
Chunghwa Telecom Co., Ltd., Taiwan
{weiweichen, chungli}@cht.com.tw

ABSTRACT

In this technical report, we briefly describe the system we designed for the Detection and Classification of Acoustic Scenes and Events (DCASE) 2026 Challenge Task 6: Audio Moment Retrieval from Long Audio. We first evaluated several existing models, including the official DCASE baseline QD-DETR, as well as QD-DETR, Moment-DETR, UVCom (with and without fine-tuning via the Lighthouse framework), and SpotSound-A. On the test set, fine-tuned UVCom achieved the best performance (R1@0.7 = 20.19%), followed by SpotSound-A (16.04%) and QD-DETR (15.96%).

As model sizes continue to grow, the time and computational cost of fine-tuning increases accordingly. We therefore explored an alternative approach: rather than fine-tuning, can we decompose a complex query into simpler sub-queries, predict each sub-query independently, and merge the predictions to recover the full answer? Using Gemma4, we decomposed original queries into sub-events, which were then individually localized by the best-performing UVCom and SpotSound models. Following the TFVTG framework, Gemma4 determined whether each sub-event occurred simultaneously or sequentially, and the predicted time windows were merged via union or intersection accordingly. UVCom achieved R1@0.7 = 20.19% on the test set; SpotSound achieved 16.04%. Results show that this direction remains highly challenging: all sub-event decomposition variants underperformed their respective baselines, primarily because the post-processing merge logic caused over-expansion of predicted windows, reducing overlap with ground-truth annotations.

***Index Terms*—**DCASE, audio temporal grounding, query decomposition, sub-event

1. INTRODUCTION

Audio Temporal Grounding aims to identify the corresponding time segment in a long audio recording given a natural language query. This task spans semantic understanding, audio event recognition, and temporal reasoning, making it a challenging problem. The DCASE 2026 Challenge Task 6 baseline uses the Castella (manually annotated) and Clotho-Moment (synthetic) datasets for training. We adopted the Castella dataset as our evaluation benchmark, which includes 1,347 test queries and 352 validation queries; the primary evaluation metric is R1@0.7.

We first evaluated the available models to establish baselines for subsequent experiments. The systems are grouped into three categories: (1) the official DCASE 2026 baseline using QD-DETR; (2) models integrated via the Lighthouse framework, including QD-DETR, Moment-DETR, and UVCom, with and without fine-tuning; and (3) SpotSound-A.

After establishing baselines, we further investigated whether query transformation—rather than fine-tuning—could achieve comparable improvements. This idea was also motivated by observations about the dataset itself: among the 352 validation queries, 169 (48%) are single-event and 183 (52%) are multi-event, an almost equal split. We thus explored whether complex multi-event queries could be decomposed into sub-queries, each localized by an existing model, with predictions subsequently merged. If successful, this approach could enable a modular pipeline where a language model handles query understanding and an audio model handles localization—without the need to retrain a single large unified model.

Concretely, Gemma4 served as the core component for splitting queries into sub-event sub-problems. Following the TFVTG framework, Gemma4 also determined the temporal relationship (simultaneous vs. sequential) between sub-events. Localization was then performed by the best-performing UVCom and SpotSound models, and final prediction windows were combined via union or intersection based on the sub-event type determined by Gemma4.

Experimental results show that the Gemma4-based query decomposition approach does not yet improve performance. The primary issue is that merging multiple sub-event prediction windows tends to over-expand the predicted range, reducing overlap with ground-truth annotations. Additionally, we experimented with re-ranking candidate windows: combining SpotSound with an event-distance ranking strategy yielded positive results on the validation set (22.73% → 24.15%), suggesting that improved ranking strategies remain a promising direction. Notably, the R1@0.7 metric leaves almost no margin for error on short windows of 1–2 seconds—even slight positional errors result in misses. This is a structural limitation of the evaluation metric for very short events, rather than a reflection of the model's actual localization ability.

2. PROPOSED METHODS

2.1. Baseline models

This study evaluated three categories of currently available systems as baselines.

The first category is the official DCASE 2026 Task 6 baseline using QD-DETR. QD-DETR extends Moment-DETR by adding

query-conditioned cross-attention, enabling more precise alignment between query semantics and temporal segments.
The second category consists of models integrated via the Lighthouse framework, including QD-DETR, Moment-DETR, and UVCom. Lighthouse is a unified temporal grounding framework that integrates multiple models and datasets with a unified inference API. We used CLAP as the audio and text feature extractor, and compared performance with and without domain-specific fine-tuning on the Castella dataset.
All models above were evaluated without any additional training beyond what is described.

2.2. Gemma sub-event decomposition

For complex queries containing multiple events—such as "A woman is talking while mixing something in a dish" or "Someone opens and closes a door"—we used Gemma4 to decompose the original query into sub-event descriptions. Each sub-event was independently localized by UVCom or SpotSound, producing individual prediction windows. Following the TFVTG framework, Gemma4 then determined whether the sub-events occurred simultaneously or sequentially, and the predictions were merged via union or intersection accordingly.

2.2.1. Sub-Event Combination and Re-Ranking

Each sub-event model outputs multiple candidate time windows. Rather than directly computing union or intersection when merging sub-events, we enumerate all possible combinations of candidates to generate multiple merged candidates. We tested three re-ranking strategies that reorder candidates without changing window boundaries: (1) Event-distance order: for simultaneous events, higher overlap is preferred; for sequential events, shorter intervals are preferred. (2) Confidence score: averaged across sub-event confidence scores (since SpotSound-A does not output confidence scores, we assigned scores from 1.0 to 0.5 based on prediction rank). (3) Combined score: a weighted combination of event-distance order and confidence score.

3. EXPERIMENTS

3.1. Baseline results

Table 1 presents the R1@0.7 results for all models on the validation set. Overall, fine-tuned models consistently outperform their pre-trained counterparts, confirming that domain-specific fine-tuning significantly benefits audio temporal grounding.
For QD-DETR, the fine-tuned version (R1@0.7 = 17.33%) is approximately 10 percentage points higher than the pre-trained version (7.67%). Moment-DETR and UVCom show similar trends, demonstrating that domain fine-tuning is effective across different architectures. Among all systems, fine-tuned UVCom achieves the highest performance, while SpotSound demonstrates stronger generalization. Compared to UVCom, SpotSound is considerably larger—requiring approximately 20 GB of memory—placing higher demands on hardware.

Notably, SpotSound achieves R1@0.7 = 22.73 without any fine-tuning, approaching the performance of fine-tuned UVCom (R1@0.7 = 28.12), demonstrating strong out-of-the-box generalization.

Model	R1@0.3	R1@0.5	R1@0.7
QD-DETR (baseline)	38.35	27.56	16.19
QD-DETR (pre-trained)	27.56	16.19	7.67
QD-DETR (fine-tuned)	45.74	33.24	17.33
Moment-DETR(pre-trained)	24.72	14.77	7.67
Moment-DETR(fine-tuned)	32.39	21.88	10.51
UVCom (pre-trained)	29.55	17.05	9.66
UVCom(fine-tuned)	50.85	41.76	28.12
SpotSound	47.16	34.66	22.73

Table 1: The single model on validation set.

3.2. Single- vs. multi-event performance

Table 2 shows R1@0.7 results on the validation set broken down by single-event (169 queries) and multi-event (183 queries) query type. SpotSound shows a notable drop on multi-event queries (−6.37 percentage points), while fine-tuned UVCom shows the opposite trend, performing better on multi-event queries (+7.43 percentage points). SpotSound handles single-event queries reasonably well but is clearly weaker on multi-event queries. In contrast, fine-tuned UVCom is better suited for complex events. The single- vs. multi-event classification used throughout this report was automatically determined by Gemma4 based on query text.

Model	Single (169)	Multi (183)	Gap
UVCom(fine-tuned)	24.26	31.69	+7.43
SpotSound	26.04	19.67	−6.37

Table 2: R1@0.7 by query type on validation set.

3.3. Sub-Event Combination and Re-Ranking Results

Table 3 shows that all three re-ranking strategies are uniformly ineffective for UVCom—UVCom's original ranking is already optimal, and alternative scoring dimensions actually reduce hit rates. Conversely, the same three strategies all yield positive effects for SpotSound, with event-distance order achieving the largest improvement (22.73% → 24.15%).

Model	All (352)	Single (169)	Multi (183)
UVCom(fine-tuned)	28.12	24.26	31.69
+Event-distance order	19.89	17.75	21.86
+Confidence score	20.17	17.75	22.40
+Combined score	18.18	15.38	20.77
SpotSound	22.73	26.04	19.67
+Event-distance order	24.15	25.44	22.95
+Confidence score	23.86	25.44	22.40
+Combined score	23.58	25.44	21.86

Table 3: R1@0.7 for sub-event combination and re-ranking strategies on the validation set

3.4. Results by Ground-Truth Window Length

Table 4 groups results by the maximum ground-truth window length for each query. For UVCom, the 1–2 second interval is the only range that benefits from re-ranking (from 7.14% up to 9.52–11.90%), but at the cost of drops in all intervals ≥3 seconds, leaving the overall effect negative. For SpotSound, the 3–6 second and 7–13 second ranges show the largest improvements (+5.68 and +7.32 percentage points, respectively), while the 1–2 second and >13 second ranges show slight declines; the overall positive result is driven by the 3–13 second queries. Both models achieve very low baseline scores in the 1–2 second range, primarily because IoU@0.7 leaves almost no tolerance for positional error with short windows—a structural constraint of the evaluation metric.

Model	1–2s (42)	3–6s (88)	7–13s (82)	>13s (140)
UVCom(fine-tuned)	7.14	17.05	36.59	36.43
+Event-distance order	9.52	12.50	24.39	25.00
+Confidence score	9.52	13.64	24.39	25.00
+Combined score	18.18	11.90	12.50	23.17
SpotSound	9.52	12.50	24.39	32.14
+Event-distance order	7.14	18.18	31.71	28.57
+Confidence score	7.14	20.45	30.49	27.14
+Combined score	7.14	19.32	29.27	27.86

Table 4: R1@0.7 by maximum ground-truth window length on the validation set.

3.5. Final Submitted System Results

Table 5 presents the results of the four systems submitted to the test set. System 1 is fine-tuned UVCom (R1@0.7 = 20.19%), the best-performing submission. System 2 is the original SpotSound (16.04%). System 3 is SpotSound with Gemma4 sub-event decomposition post-processing (13.81%)—underperforming the baseline, reflecting that the Gemma-based window merging issues observed on the validation set persist on the test set. System 4 is SpotSound with the secondary-score re-ranking strategy (18.19%), which outperforms original SpotSound and is consistent with the positive trends seen on the validation set.

System	R1@0.7
QD-DETR (baseline)	10.32
System 1	20.19
System 2	16.04
System 3	13.81
System 4	18.19

Table 5: R1@0.7 on test set.

4. CONCLUSION

Experimental results show that fine-tuned UVCom is the most effective approach, while SpotSound offers stronger generalization. UVCom is considerably smaller than SpotSound—which requires approximately 20 GB of memory to run.

Using Gemma4 to decompose queries offers marginal improvements for SpotSound's R1@0.7, but the combination of Gemma4 and SpotSound is very slow, meaning most experiments could only be completed on the validation set. Only four systems completed test set evaluation and were submitted.

Additionally, for very short events of 1–2 seconds, R1@0.7 demands nearly perfect prediction accuracy—even slight positional deviations result in misses. This causes consistently low scores in this interval for all models.

5. REFERENCES

[1] A. Martin et al., "The Castella dataset for audio temporal grounding," in Proc. DCASE Workshop, 2026.

[2] SpotSound Team, "SpotSound-A: Audio-based sound event detection system,"Internal Technical Document, Tech. Rep., 2026.

[3] J. Yang et al., "UVCom: Unified video-audio clip moment retrieval," in Proc.ACM Multimedia, 2024.

[4] Gemma Team, "Gemma: Open models based on Gemini research and technol-ogy," 2024, arXiv:2403.08295.

[5] J. Lei et al., "TVQA: Localized, compositional video question answering," in Proc. EMNLP, 2018.

[6] K. Miyazaki et al., "Language-based audio retrieval task in DCASE 2022 chal-lenge," in Proc. DCASE Workshop, 2022.

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

[8] S. Moon et al., "Query-dependent video representation for moment retrieval and highlight detection," in Proc. CVPR, 2023.

[9] J. Lei et al., "Detecting moments and highlights in videos via natural language queries," in Proc. NeurIPS, 2021.

[10] B. Elizalde et al., "CLAP: Learning audio concepts from natural language super-vision," in Proc. ICASSP, 2023.