YCU SUBMISSION FOR DCASE 2026 CHALLENGE TASK 6

Technical Report

Haruto Sugawara, Tomoko Nakamura, Ken Sakaguchi, Hikari Aida, Shunta Yuri

Yokohama City University, Yokohama, Japan
{y255614c, y255616e, y265618d, y265601a, y265634f}@yokohama-cu.ac.jp

ABSTRACT

This report describes our submission to DCASE 2026 Task 6 (Audio Moment Retrieval, AMR). Our system extends UVCOM, pre-trained on Clotho-Moment and fine-tuned on CASTELLA, with a set of improvement components investigated through ablation on the CASTELLA validation and test splits. Our primary system is a 19-member Weighted Box Fusion ensemble that combines models trained with Quality Focal Loss, delta-feature augmented inputs, CASTELLA-style audio augmentation, and two acoustic encoders (MS-CLAP and M2D-CLAP) at several temporal resolutions and learning-rate schedules. It achieves R1@0.7 = **39.57 %** on the CASTELLA test split.

1. INTRODUCTION

Audio Moment Retrieval (AMR) is the task of localizing temporal intervals in a long audio recording that match a free-form natural-language query. Unlike sound event detection, which targets short clips and a fixed label set, AMR handles long-form audio (typically over one minute) with arbitrary text queries. Formally, given an input audio recording $\mathbf{x}$ and a text query q, an AMR system predicts N candidate moments $\mathbf{y} = \{y_1, \ldots, y_N\}$ together with their confidence scores, where each $y_n = (t_{n,\text{start}}, t_{n,\text{end}})$ is a temporal interval.

DCASE 2026 Task 6 adopts AMR using the CASTELLA dataset [1], a human-annotated benchmark of 1,862 audio recordings (1,009 / 213 / 640 for train / validation / test) totaling 120 hours. The official baseline of this task is a DETR-style [10] transformer following the AMR framework of [2]. The official evaluation metric is Recall@1 at IoU 0.7 (R1@0.7), which counts a query as correct only when the top-1 predicted span has temporal $\text{IoU} \geq 0.7$ with the ground truth.

We adopt UVCOM [3], reported as the strongest DETR variant on CASTELLA [1], as our base system. Following the standard two-stage training recipe [1], we start from a UVCOM checkpoint pre-trained on the synthetic Clotho-Moment dataset [2], then fine-tune on CASTELLA. On top of this base, we investigate several improvement components — spanning loss functions, feature representations, data augmentation, post-processing, and ensembling — and combine those that improved performance under ensemble-level evaluation. Our primary system achieves R1@0.7 = **39.57 %** on the CASTELLA test split.

2. BASE SYSTEM

2.1. Dataset

CASTELLA [1] consists of 1,862 audio recordings collected from YouTube as a subset of AudioCaps and trimmed to between one and five minutes. The dataset is split into 1,009 / 213 / 640 recordings for train / validation / test, with 2,182 / 352 / 1,347 free-form text queries and 6,160 / 973 / 4,175 ground-truth temporal boundaries, respectively. Each audio recording is associated with multiple local captions (2.1 per recording on average) and a single global caption. Local captions describe salient audio moments together with their start and end timestamps at one-second resolution. The dataset is dominated by short moments: timestamps shorter than 10 s are the most frequent and constitute a known difficult case for AMR [1].

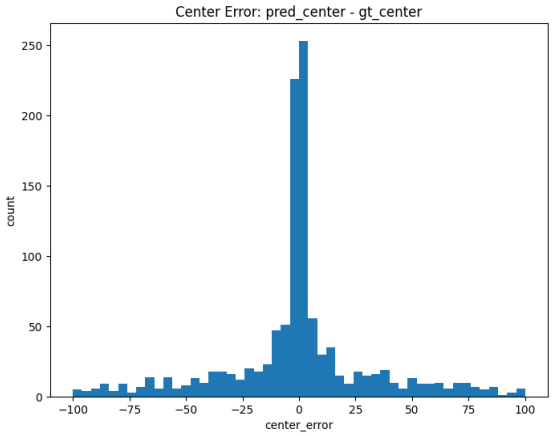

Figure 1: Distribution of center errors $t_{\text{pred,center}} - t_{\text{gt,center}}$ for the baseline's top-1 predictions on the CASTELLA test split.

2.2. Distributed Features

The challenge distributes pre-extracted MS-CLAP [4] features for the CASTELLA audio recordings, following the feature-extraction protocol: audio is down-sampled to 32 kHz and processed with a sliding window of 1 s length and 1 s hop, producing 768-dim audio embeddings at 1 fps. Text queries are encoded by the same CLAP text branch into 768-dim embeddings.

2.3. Baseline Architecture

The official baseline of DCASE 2026 Task 6 is QD-DETR [13], a DETR-style [10] transformer that constructs query-dependent representations of the audio input for moment retrieval. Following the recipe of [1], the model is trained in two stages: pre-training on Clotho-Moment [2] and fine-tuning on CASTELLA. The default loss combines L1 + GIoU span regression and cross-entropy classification, together with auxiliary saliency / contrastive losses.

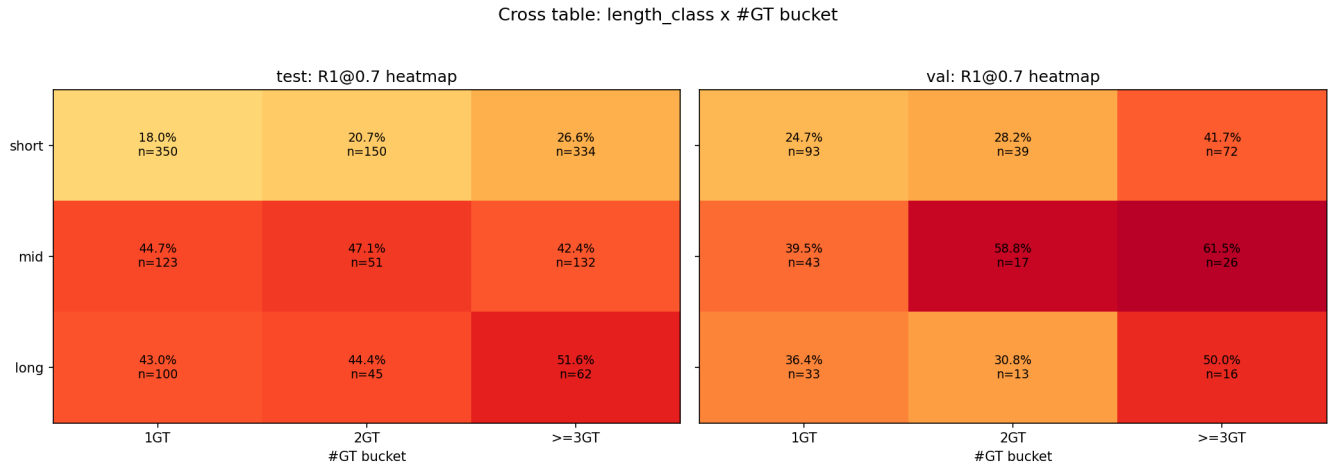

Figure 2: R1@0.7 broken down by ground-truth moment length (short: $\leq$ 10 s, middle: 10–30 s, long: $>$ 30 s) and by the number of ground-truth moments per query (1 GT, 2 GT, $\geq$ 3 GT) on the CASTELLA test (left) and validation (right) splits. Short moments and single-ground-truth queries are the dominant failure cases.

3. PRELIMINARY ANALYSIS

Before designing improvements, we analyzed baseline predictions on the CASTELLA test split to identify dominant failure modes.

The center is approximately correct. Figure 1 shows the distribution of center errors $t_{\text{pred,center}} - t_{\text{gt,center}}$ for the baseline's top-1 predictions. The errors are sharply concentrated around zero, indicating that the baseline localizes the approximate position of the target moment well in most cases. The remaining error therefore lies primarily in the *precision of the span boundaries*: the predicted center is roughly correct, but the predicted interval is not tight enough around the ground truth to satisfy the R1@0.7 threshold.

Short Moments and Single-GT Queries Are Harder. Figure 2 breaks R1@0.7 down by the length of the ground-truth moment (short: $\leq$ 10,s, middle: 10–30,s, long: $>$ 30,s) and by the number of ground-truth moments per query (1,GT, 2,GT, $\geq$ 3,GT). On the test split, short moments are consistently difficult, with R1@0.7 remaining below $27 compared with 40--52$agrees with the observation of [1] that short moments are a dominant failure case for AMR. The heatmap also reveals that, within each length class, queries containing a single ground-truth moment are generally harder than those with multiple ground-truth moments (e.g., $18.0 versus 26.6$that, when multiple acceptable targets exist, the top-1 prediction needs to overlap with only one of them, providing additional tolerance to localization errors.

Implications. These observations point to several challenges that may be limiting performance, including imprecise boundary estimation and the difficulty of localizing short moments. Potential directions for addressing these issues include capturing finer temporal details, improving span localization quality, and encouraging prediction confidence to better reflect localization accuracy. The stronger performance on queries with multiple ground-truth moments further suggests that reducing localization error itself may be an important factor for improving R1@0.7. he improvements described in Section 4 were primarily motivated by these observations, although some components were also explored for complementary reasons.

4. IMPROVEMENT COMPONENTS

We describe our improvements over the QD-DETR baseline as a series of independent components.

4.1. Backbone Architecture: UVCOM

In place of the QD-DETR baseline, we adopt UVCOM [3], the strongest DETR variant reported on CASTELLA [1], as our base architecture. UVCOM extends the DETR framework with a Comprehensive Integration Module (CIM) that performs intra- and inter-modality interaction at multiple granularities, followed by a DETR decoder with $N_q = 30$ object queries. This multi-granularity design is well-suited for AMR, where free-form text queries must be matched against audio events that vary widely in both timescale and content type. We use the implementation provided by the Lighthouse library [11] and fine-tune from the Clotho-Moment pre-trained UVCOM checkpoint distributed with Lighthouse; pre-training the backbone ourselves did not transfer as well as the official checkpoint. The remaining improvement components below are applied on top of this UVCOM base.

4.2. M2D-CLAP Feature Replacement

We replace the MS-CLAP input features with M2D-CLAP [5], which combines masked-modeling pre-training with CLAP-style audio-language alignment. M2D-CLAP features are extracted from the raw audio at 50 fps natively and mean-pooled to 1 fps, 3 fps, 5 fps, and 6.25 fps variants; models trained on these variants enter the final ensemble alongside MS-CLAP-based models.

In a matched-control comparison in which *only* the audio feature was changed, the M2D-CLAP model clearly outperformed its MS-CLAP counterpart. The interaction between temporal resolution and optimization is more nuanced than we first assumed: at the default learning-rate schedule, a single high-resolution / small-batch cell (5 fps) was under-trained and did not beat the 1 fps model, but at matched schedule the 1 fps $\rightarrow$ 3 fps step improved test R1@0.7 by roughly +1 point on average, and our strongest single models are high-resolution (3 fps / 5 fps / 6.25 fps) once paired with a

warm-restart learning-rate schedule suited to that regime. We therefore attribute the improvement to the representation itself (masked-modeling pre-training capturing acoustic detail complementary to contrastive-only MS-CLAP) and, plausibly, to finer temporal resolution under an adequate schedule — though resolution and schedule remain entangled in the high-resolution runs. M2D-CLAP further adds *encoder diversity* to the ensemble: M2D-based members exhibit error profiles distinct from MS-CLAP members, and adding them to the MS-CLAP ensemble improved test R1@0.7 by +4.9 points (5 fps members) plus a further +1.4 points (1 fps members).

4.3. Quality Focal Loss

The base classification loss is binary cross-entropy on foreground / background. We replace it with Quality Focal Loss (QFL) [6] with focal exponent $\beta = 2$:

$$\mathcal{L}_{\text{QFL}} \;=\; -\,|\sigma - y|^{\beta}\,[(1-y)\log(1-\sigma) + y\log\sigma]\,, \qquad (1)$$

where σ is the predicted foreground probability and $y \in [0, 1]$ is the actual IoU between matched prediction and ground truth used as a soft quality target. QFL encourages classification confidence to align with localization accuracy. This is particularly relevant for the AMR R1@0.7 metric, which only counts the top-1 prediction: a ranking signal that already reflects IoU quality directly increases the chance that the top-1 prediction satisfies the IoU ≥ 0.7 threshold, without requiring any additional post-processing or auxiliary head. QFL-trained seeds form the core of our final ensemble.

4.4. CASTELLA-style Audio Augmentation

We construct additional training examples by recombining pairs of CASTELLA training recordings. Because raw waveforms are not distributed for the training features, both operations act directly on the frame-level feature sequences: (i) **mixing**, where the feature sequences of two randomly paired training recordings are convexly combined frame-by-frame with a Beta-distributed weight, keeping the original query while retaining the ground-truth boundaries of *both* source recordings as positive moments; and (ii) **cut-and-paste**, where the annotated moment segment of one training recording is inserted into a non-moment region of another randomly drawn recording, keeping the donor's query and re-mapping the ground-truth boundaries to the new time coordinates.

The two operations expand the effective training distribution in ways that mirror challenging conditions in AMR deployment: mixing simulates a target moment partially masked by an unrelated acoustic event, while cut-and-paste exposes the model to abrupt scene transitions and varied temporal layouts, both reducing the model's reliance on the position bias of the original recordings. The augmentation produced large gains in early development and is retained ($p = 0.5$) in every member of the final ensemble. A variant that removes the cross-query label noise of the mixing operation improved single-model validation scores but *lost* to the original design under ensemble evaluation, so the original design was kept.

4.5. Delta-feature Branch

We compute first-order temporal differences $\Delta f_t = f_t - f_{t-1}$ over the CLAP feature sequence to expose short-term acoustic changes (onsets / offsets) that may aid boundary localization. Rather than concatenating $[f_t;\, \Delta f_t]$ into the existing audio projection — which perturbs the projection's input statistics and destroys the identity-resume property of the pre-trained checkpoint — the delta sequence is processed by a *separate, zero-initialized* LayerNorm + Linear branch whose output is added to the original audio projection. The branch is attached after the first epoch, so training starts exactly from the pre-trained behavior. This design improved test R1@0.7 by +1.63 / +2.82 / +3.04 points over the matched no-delta control on three seeds (mean +2.50, 3/3 positive), and the three delta seeds are members of the final ensemble.

4.6. Weighted Box Fusion Ensemble

We fuse the span predictions of multiple trained models using a 1D adaptation of Weighted Box Fusion [9], producing a single ranked list of moments per query. The final ensemble contains 19 members spanning six training configurations — QFL-only (4 seeds), QFL + delta-feature (3 seeds), and four M2D-CLAP variants: 5 fps and 1 fps under the constant-schedule recipe and 3 fps and 5 fps under the warm-restart schedule (3 seeds each) — with fusion weights derived from validation R1@0.7 and a union fallback that guarantees a prediction for every query. Diversity across configurations, in particular across the two acoustic encoders and across temporal resolutions, was the dominant source of ensemble gain, exceeding what additional seeds of a single configuration provide. We also found that single-model validation improvements do not reliably predict a model's value as an ensemble member; all training-side changes were therefore accepted or rejected at the ensemble level. Member selection is performed at the granularity of training *configurations* (all seeds of a recipe kept or dropped together), not individual seeds: a validation-locked search over seed-level subsets gained +6 points on validation yet failed to beat the configuration-level ensemble on the test split, whereas adding the two warm-restart configurations as whole blocks improved test R1@0.7 by +2.6 points (36.97 $\rightarrow$ 39.57), with every length bucket non-degraded.Our strongest submission extends this 19-member ensemble with six 6.25 fps warm-restart snapshots.

5. EXPERIMENTS

5.1. Setup

We follow the two-stage training recipe of [1]: Stage 1 uses the Clotho-Moment pre-trained UVCOM checkpoint distributed with the Lighthouse library [11]; Stage 2 fine-tunes this checkpoint on CASTELLA. Fine-tuning uses AdamW [12] (learning rate 1×10^{-4}, weight decay 1×10^{-4}) at full fp32 precision on a single NVIDIA RTX 5090 GPU, with audio augmentation (Section 4.4) applied at probability 0.5 and checkpoints selected by validation mAP in all runs. Two learning-rate schedules are used. Most members follow the *constant schedule*: 250 epochs with a step decay ($\times 0.1$ at epoch 200), at batch size 80 for MS-CLAP members and 12 for M2D-CLAP members. The two *warm-restart* M2D configurations (3 fps batch size 12, and 5 fps batch size 8) instead hold the learning rate at 1×10^{-4} through epoch 180 and then apply an SGDR cosine warm-restart schedule to epoch 260.We evaluate on the CASTELLA test split (1,347 queries) using the official protocol.

5.2. Results

Table 1 reports the comparison. Row 1 is the official QD-DETR baseline reported in the CASTELLA paper; rows 2–5 are our four challenge submissions, all produced by the same un-rescaled Weighted Box Fusion with union fallback and differing only in member composition. Details of each submission are Section 5.3.

All four submissions substantially outperform the official QD-DETR baseline across every metric. The primary submission (#1) achieves R1@0.7 = 39.57, improving over the baseline by 23.37

Table 1: Our four challenge submissions on the CASTELLA test split (%). All use the same un-rescaled Weighted Box Fusion with union fallback and differ only in member composition. Best per column in bold.

Method	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	mAP
QD-DETR (baseline)	30.6	16.2	26.5	13.7	12.2
#1 (19-member, clean configuration-level best)	51.8	39.57	49.0	**29.8**	29.9
#2 (16-member, diversity hedge)	51.3	38.38	48.4	29.4	29.4
#3 (25-member, +6.25 fps, strongest)†	**53.5**	**40.68**	**49.9**	29.7	**30.3**
#4 (19-member, selection-transfer probe)	52.5	38.01	49.1	28.0	29.1

† #3 is our highest-scoring submission but sits essentially at our training-free fusion ceiling, and its 6.25 fps members did not pass our stricter clean-merge gate; #1 is therefore reported as the clean configuration-level best.

points, while the diversity-focused variant (#2) reaches a comparable 38.38 despite using fewer ensemble members. Introducing additional encoder and temporal-resolution diversity (#3) further improves R1@0.7 to 40.68 , yielding the strongest overall performance among all submissions.The selection-transfer probe (#4), despite having the same ensemble size as #1, reaches only R1@0.7 = 38.01. This suggests that member composition contributes more to final performance than ensemble size alone, and that carefully chosen diversity is a key factor behind the observed gains.

Impactful improvements. Based on the experiments conducted, the most impactful improvements were M2D-CLAP feature replacement (the M2D members lift the ensemble from R1@0.7 = 30.66 to 39.57), the UVCOM backbone adoption, CASTELLA-style audio augmentation, the delta-feature branch (+2.5 points per seed), and Quality Focal Loss. Two of these address characteristics of the CASTELLA dataset itself: the augmentation mitigates the limited size of the training set (1,009 recordings) by recombining pairs of training features, and M2D-CLAP provides a stronger and complementary acoustic representation — the gain stems from the representation and from encoder diversity in the ensemble, with finer temporal resolution contributing once the learning-rate schedule is adapted to the high-resolution / small-batch regime (Section 4.2); short moments nonetheless remain the dominant unsolved failure mode (19-member ensemble short-moment R1@0.7 = 8.0, up from 7.2 for the 13-member system). The remaining components address general AMR challenges: UVCOM's multi-granularity interaction, and QFL's alignment of confidence with localization quality for the top-1-only metric.

5.3. Challenge Submission

Following the DCASE 2026 submission rules (up to four system outputs per task), our submissions are generated by a dedicated evaluation pipeline. The official evaluation set (177 queries over 100 recordings, ground truth withheld) is processed along two feature paths: MS-CLAP members consume the organizer-distributed evaluation features, while M2D-CLAP members consume features we extract from the official evaluation audio with the organizers' permission, using the identical extraction protocol as in training. Per-member predictions are fused with the same WBF configuration, weights, and union fallback as the internal system, and the fused top-1 span (floating-point seconds, clamped to the clip duration) is converted to the official five-field submission format.

Under the four-submission budget, we submit four ensembles drawn from this single system family. All four use the same union fallback and the same (un-rescaled) Weighted Box Fusion rule, and differ only in member composition and selection policy.

- **#1** (clean configuration-level best; 2nd strongest): the 19-member ensemble described in Section 4.6; CASTELLA test R1@0.7 = 39.57.
- **#2** (diversity hedge): a 16-member ensemble selected to maximize per-query error decorrelation from system #1 (lowest correctness-vector correlation and member-set overlap with the strongest system); CASTELLA test R1@0.7 = 38.38.
- **#3** (strongest, high temporal resolution): the 19-member system #1 augmented with six 6.25,fps warm-restart snapshots, totaling 25 members — adding the finest temporal resolution we trained; CASTELLA test R1@0.7 = **40.68** — the highest of the four, improving every length bucket over system #1, though it sits essentially at our training-free fusion ceiling, and its 6.25,fps members did not pass our stricter clean-merge gate.
- **#4** (selection-transfer probe): the ensemble of the 19 single models with the highest *individual* CASTELLA test R1@0.7. This entry is deliberately selection-biased to stress-test on the withheld official evaluation whether test-split selection transfers, as a deliberate counterpart to our finding that validation-level selection does not; CASTELLA test R1@0.7 = 38.01.

Because official evaluation labels are withheld, CASTELLA test performance is used only to order the four submissions; the challenge does not require designating a primary system. Systems 1–3 are produced by a fixed, label-free fusion rule, whereas system 4 is intentionally selection-biased as a transfer-of-overfitting probe.

6. CONCLUSION

We presented our submission to DCASE 2026 Task 6 (Audio Moment Retrieval). Building on the UVCOM backbone, we incorporated improvements targeting the dominant failure modes identified in our preliminary analysis: imprecise span boundaries and the difficulty of short moments. The combination of Quality Focal Loss, CASTELLA-style audio augmentation, a delta-feature input branch, M2D-CLAP feature replacement at multiple temporal resolutions and schedules, and Weighted Box Fusion ensembling improves R1@0.7 from 16.2 % (QD-DETR baseline) to 39.57 % (our primary 19-member ensemble) on the CASTELLA test split, a gain of 23.4 points. Our strongest submission, a 25-member ensemble with additional 6.25 fps warm-restart members, further reaches $R1@0.7 = 40.68\,\%$. Our four challenge submissions are drawn from this single system family; the remaining two entries hedge for encoder/resolution diversity and probe whether test-split selection transfers to the withheld evaluation. Notably, we observed throughout this study that single-model validation gains do not reliably predict ensemble-member value; our member acceptance therefore operated at the ensemble level.

7. REFERENCES

[1] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long audio dataset with captions and temporal boundaries," *arXiv:2511.15131*, 2026.

[2] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Language-based audio moment retrieval," in *Proc. ICASSP*, 2025, pp. 1–5.

[3] 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. CVPR*, 2024, pp. 18709–18719.

[4] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. ICASSP*, 2024, pp. 336–340.

[5] 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.

[6] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. NeurIPS*, 2020.

[7] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. ECCV*, 2018, pp. 784–799.

[8] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. ICCV*, 2019, pp. 9627–9636.

[9] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, pp. 104117, 2021.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020, pp. 213–229.

[11] T. Nishimura, S. Nakada, H. Munakata, and T. Komatsu, "Lighthouse: A user-friendly library for reproducible video moment retrieval and highlight detection," in *Proc. EMNLP: System Demonstrations*, 2024.

[12] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[13] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-dependent video representation for moment retrieval and highlight detection," in *Proc. CVPR*, 2023, pp. 23023–23033.