MomentMix Augmentation with Length-Aware DETR for Temporally Robust Moment Retrieval

Supplementary Material

Contents

A. Additional Ablation Studies

ICA-Based Feature Diversity Analysis. To investigate the feature distributions of the QVHIGHLIGHTS *val* set, we project the features into a latent space constructed using PCA and FastICA. To ensure robustness, we filter out unstable components and retain only those showing high consistency across multiple initialization runs. Figure A1 presents the pairwise distributions of the components with the most distinct differences between short and non-short groups. The plot qualitatively demonstrates that non-short moments occupy a broader region in the latent space than short moments, as evidenced by wider kernel density estimates (KDEs) and larger covariance ellipses.

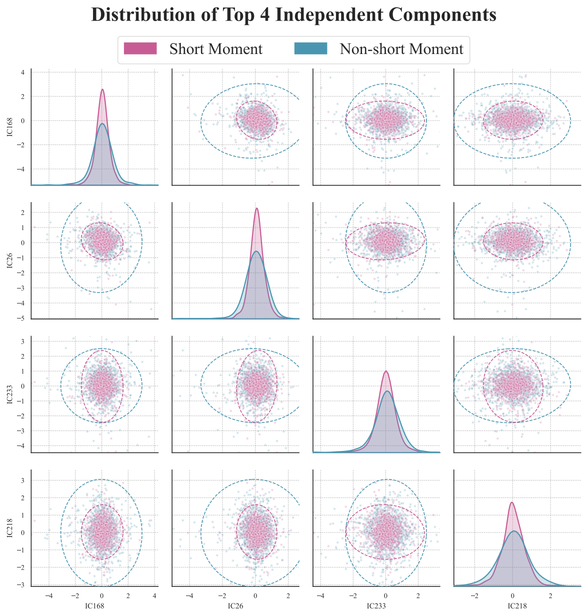


Figure A1. Pairwise scatter plots for the selected ICs. Diagonals show the 1D density (KDE), while off-diagonals show 2D distributions. The larger 95% covariance ellipses for the Non-short Moment group provide strong evidence of its greater feature diversity.

Effect of Length-Aware Decoder. We introduce the Length-Aware Decoder (LAD), a novel framework designed to improve moment center predictions by conditioning on moment length. LAD's core innovation is a length-wise bipartite matching scheme, which assigns queries to predefined length classes to generate length-expert queries. While LAD builds upon Anchor-DETR, Anchor-DETR's pattern embeddings alone are not length-aware. Furthermore, LAD's strategy differs from the *length class-wise one-to-one* matching of Group-DETR by employing a more precise *length class-wise one-to-one* matching tailored for moment retrieval, as shown in Fig. A2.

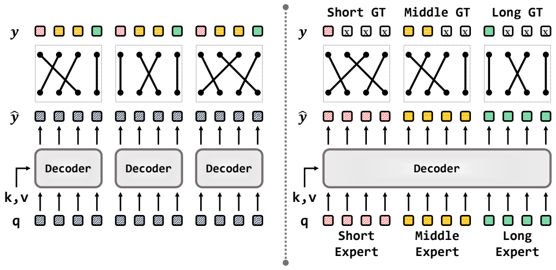


Figure A2. [Left] Group-DETR [4] employs one-to-many matching, where the same labels are utilized across all groups. [Right] Our length-wise matching is one-to-one, and it operates within each length class. By matching only the predictions and ground truths that belong to the same class, this approach enables the creation of length-wise expert queries.

The effectiveness of our approach is validated in Table A1 with two key findings. First, applying LAD's length-wise matching to the Anchor-DETR baseline yields significant gains in both overall and short-moment mAP, proving the value of explicit length conditioning. Second, in contrast to Group-DETR, which improves R@1 but suffer from significant drop in mAP, LAD achieves substantial improvements in both metrics. These results underscore that LAD's specialized mechanism effectively addresses performance degradation in challenging short moments, establishing it as a more robust and task-specific solution.

Method	Short		Middle		Long		Full	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Baseline	4.57	7.77	38.89	43.10	42.62	47.44	41.06	41.00
Anchor-DETR	4.48	7.43	39.53	43.36	45.4	49.09	42.46	41.55
Group-DETR	4.90	3.84	40.24	40.72	43.14	43.79	42.17	37.97
LAD	**8.76**	**11.01**	**40.55**	**45.53**	**43.69**	**50.76**	**43.65**	**44.48**

Table A1. Performance comparison on QVHIGHLIGHTS *val* set.

Effect of number of queries. Our method employs 40 queries, with 10 queries allocated to each length class. In comparison, QD-DETR and TR-DETR originally use 10 queries, while UVCOM uses 30. To ensure that the observed performance improvements with our method are not simply due to the increased number of queries, we retrained the baselines with 40 queries for a fair comparison. The results, presented in Table A2, clearly demonstrate that increasing the number of queries in the baselines does not guarantee performance gains and can even lead to performance drops, as observed with TR-DETR. This indicates that the performance gains achieved by our method are not trivial or merely due to an increased number of queries.

Method	Short		Middle		Long		Full	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
QD-DETR	4.45	8.34	39.54	43.54	43.89	47.80	41.90	41.24
QD-DETR†	5.48	8.88	40.17	43.95	40.70	44.17	41.39	40.29
QD-DETR+Ours	**11.07**	**15.27**	**43.12**	**48.53**	**44.39**	**52.65**	**46.13**	**47.70**
TR-DETR	5.80	9.91	44.01	46.95	47.35	51.70	46.32	45.10
TR-DETR†	3.66	7.32	39.44	43.01	**47.39**	50.07	42.91	41.83
TR-DETR+Ours	**9.91**	**15.17**	**47.11**	**51.83**	46.78	**53.53**	**49.15**	**49.80**
UVCOM	5.97	12.65	45.97	49.04	45.19	49.39	46.77	45.80
UVCOM†	5.48	10.39	40.17	47.82	40.70	47.79	41.39	43.87
UVCOM+Ours	**12.80**	**18.46**	**46.87**	**52.36**	**46.92**	**53.23**	**49.85**	**50.76**

Table A2. Results on QVHIGHLIGHTS *val* set. † indicates training with the number of queries the same as ours.

Effect of cut criteria in ForegroundMix. We propose ForegroundMix, which cuts a long foreground into shorter sub-foregrounds, shuffles them, and generates new short-moment data. We analyze the effect of ε_{cut}, which determines sub-foreground shortening relative to the original long foreground, with QD-DETR as the baseline.

As shown in Table A3, smaller values of ε_{cut} (more aggressive cutting and greater shortening) lead to improved performance on shorter moments. Regardless of the value, ε_{cut} consistently enhances overall performance. Since our primary objective is to improve short-moment performance, we adopt the smallest value, $\varepsilon_{\text{cut}} = 5$, as our default setting.

Method	**Short**		Middle		Long		Full	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Baseline	4.57	7.77	38.89	43.10	42.62	47.44	41.06	41.00
$\varepsilon_{\text{cut}} = 5$	7.86	12.21	41.42	45.28	43.45	47.69	43.84	43.32
$\varepsilon_{\text{cut}} = 10$	6.78	10.87	42.31	46.07	44.16	48.11	44.35	43.45
$\varepsilon_{\text{cut}} = 15$	5.45	8.68	41.35	44.78	44.34	48.37	43.46	42.48

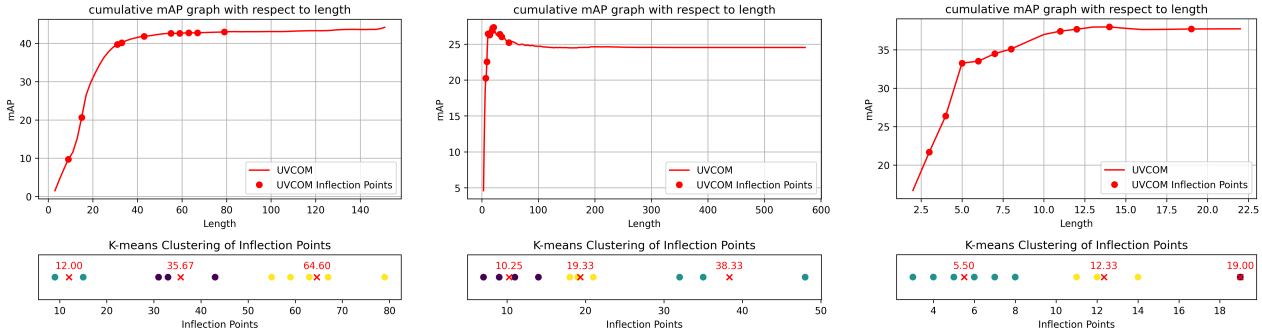
Table A3. Performance comparison on QVHIGHLIGHTS *val* set. ε_{cut} controls sub-foreground shortening in ForegroundMix. Across all values, ε_{cut} consistently improving overall performance, with smaller values excelling in short-moment enhancement.

B. Moment Length Class Selection

Defining length class. To define multiple length classes, we select corresponding length thresholds using the cumulative mAP graph with respect to length, as shown in Figure A3. We chose cumulative mAP because it effectively highlights lengths where the model underperforms. Initially, we compute the cumulative mAP for each moment length based on an existing moment retrieval baseline, UVCOM. Subsequently, we identify the inflection points on the graph and cluster them using K-means. These clustered points determine the length class thresholds.

Performance comparison based on the number of classes. The number of classes, $\mathcal{N}_c$, is determined by the value of k in K-means. To determine the optimal k, we experimented with different class numbers with QD-DETR as the baseline. As shown in Table A6, using four classes resulted in the highest length-awareness in the model.

Comparison of class-division strategies. We evaluate three substitutes for K-means—equal-interval, equal-count, and Fisher–Jenks variance minimization. As depicted in Tab. A7, every alternative yields substantial gains over the baseline, confirming robustness to how length classes are



(a) QVHIGHLIGHTS

(b) TACOS

(c) CHARADES-STA

Figure A3. We defined length class based on inflection points in the cumulative mAP graph with respect to length.

defined. Across all datasets, K-means consistently achieved the best performance, so we adopt it as the default.

Method	**Short**		Middle		Long		Full	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Baseline	4.57	7.77	38.89	43.10	42.62	47.44	41.06	41.00
$\mathcal{N}_c = 2$	6.95	9.01	38.19	44.95	43.36	49.80	41.56	42.99
$\mathcal{N}_c = 3$	7.20	9.64	38.54	44.70	41.29	49.13	41.08	43.03
$\mathcal{N}_c = 4$	8.76	11.01	40.55	45.53	43.69	50.76	43.65	44.48

Table A6. Performance comparison on QVHIGHLIGHTS *val* set. $\mathcal{N}_c$ indicates the number of length classes in LAD.

Method	QVHIGHLIGHTS		TACOS		CHARADES-STA	
	R1 avg.	mAP avg.	R1@0.5	R1@0.7	R1@0.5	R1@0.7
UVCOM	46.77	45.80	36.39	23.32	59.25	36.64
K-means	49.85	50.76	42.21	28.02	61.45	40.22
Equal-Interval	48.86	49.21	41.41	27.99	60.67	40.51
Equal-Count	48.39	50.42	41.54	27.09	58.47	38.66
Fisher-Jenks	47.65	49.58	41.01	26.84	60.19	40.86

Table A7. Performance Comparison of different strategies for moment length class definition. Across all datasets, all strategies improve performance over the baseline; K-means yields the best results overall.

C. Evaluation with Diverse Feature Types

We conducted experiments using various feature types to demonstrate that our methods—MomentMix augmentation and the Length-Aware Decoder—are robust and not limited to specific features.

Evaluation with additional audio features. Following prior work, we incorporated additional audio features extracted from PANNs [16] to evaluate our method's performance. As shown in Table A4, compared to the baseline UVCOM trained with the additional audio modality, our method significantly outperforms the baseline, indicating its effectiveness.

Evaluation with InternVideo2 features. To further validate the robustness of our method across different feature types, we utilized features from InternVideo2 [35], a recent foundational model for multimodal video understanding, for both video and text modalities. We re-trained the baseline UVCOM and our method using these richer and more powerful features. As shown in Table A5, despite the enhanced feature quality, the baseline still suffers from performance degradation in short moments. In contrast, our method significantly improves short-moment performance, achieving gains of 9.19% in R1 and 8.66% in mAP, along with overall performance improvements. These results demonstrate that our method effectively addresses the short-moment performance issues.

D. More Qualitative Results

We provide comparisons with other models across a broader range of samples. Predictions with confidence scores exceeding 0.7 are visualized with alpha = 0.5 on the QVHIGHLIGHTS *val* set; "Ours" denotes UVCOM augmented with our method.

As shown in Figure A4, Existing models frequently struggle to effectively distinguish between foreground and background, resulting in inaccurate predictions or missed detections of short moments. In contrast, our model excels in accurately and robustly predicting short moments. our approach yields consistently higher accuracy on short moments.

We also include representative failure cases that are difficult irrespective of length, primarily due to visual–text alignment challenges. Examples are strong window reflections that obscure content and screen-in-screen scenes that induce visual ambiguity as shown in Figure A5.

Method	Short		Middle		Long		Full					
	R1	mAP	R1	mAP	R1	mAP	R1			mAP		
	Avg.		Avg.		Avg.		@0.5	@0.7	Avg.	@0.5	@0.75	Avg.
UVCOM‡	4.45	11.00	44.18	48.03	43.89	48.84	64.26	49.42	44.76	64.92	45.29	44.70
+ Ours	13.38 (+8.93)	17.77 (+6.77)	45.51 (+1.33)	51.12 (+3.09)	45.84 (+1.95)	53.51 (+4.67)	66.71 (+2.45)	52.97 (+3.55)	48.77 (+4.01)	68.04 (+3.12)	51.52 (+6.23)	50.10 (+5.40)

Table A4. Performance comparison on the QVHIGHLIGHTS *val* set using additional audio modality. ‡ means the result reproduced from the original repository.

Method	Short		Middle		Long		Full					
	R1	mAP	R1	mAP	R1	mAP	R1			mAP		
	Avg.		Avg.		Avg.		@0.5	@0.7	Avg.	@0.5	@0.75	Avg.
UVCOM‡	5.64	10.83	48.96	51.12	49.16	51.71	70.13	54.97	49.99	67.95	47.88	47.56
+ Ours	14.83 (+9.19)	19.49 (+8.66)	49.23 (+0.27)	54.60 (+3.48)	49.56 (+0.40)	55.67 (+3.96)	71.10 (+0.97)	58.00 (+3.03)	52.85 (+2.86)	71.45 (+3.50)	54.84 (+6.96)	53.25 (+5.69)

Table A5. Performance comparison on the QVHIGHLIGHTS *val* set using the InternVideo2$_{s2}$-6B features for both video and text modalities. ‡ means the result reproduced from the original repository.

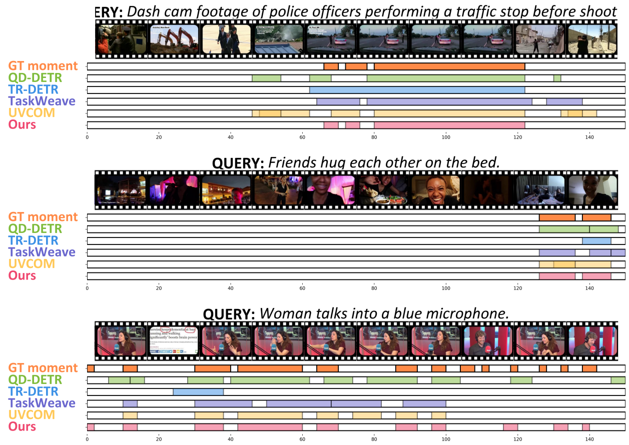


Figure A4. Qualitative success on short moments with sharper boundaries and fewer spurious hits.

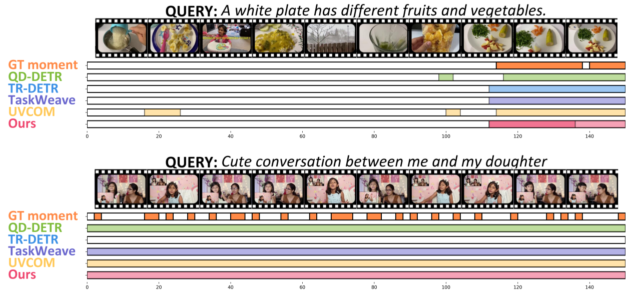


Figure A5. Qualitative failure driven by difficult visual–text alignment.