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

Seojeong Park · Jiho Choi · Kyungjune Baek · Hyunjung Shim

{seojeong.park, jihochoi, kateshim}@kaist.ac.kr, kyungjune.baek@sejong.ac.kr

WACV 2026
TUCSON, AZ
3/6 - 3/10

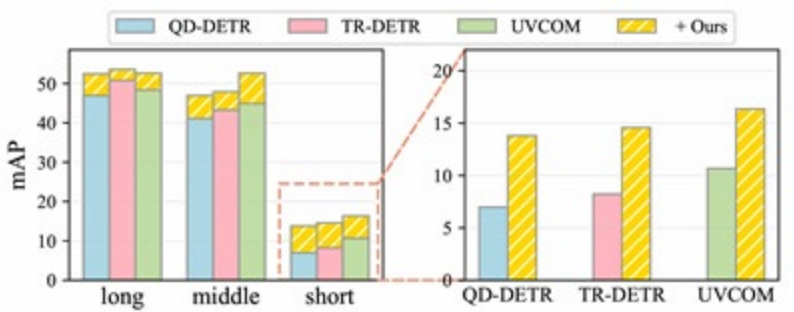
Topic Introduction & Motivation

Video Moment Retrieval

- Task
  - Localize moments within a video based on a given natural language query
  - **Input**: text query → **Output**: (start, end)
- Motivation
  - Improve user experience and search efficiency

Limitations of Previous Work

- DETR-based models suffer from a significant drop in performance <u>when handling short moments</u>



Moon, WonJun, et al. "Query-dependent video representation for moment retrieval and highlight detection." CVPR 2023.
Sun, Hao, et al. "Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection." AAAI 2024.
Xiao, Yicheng, et al. "Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection." CVPR 2024.

Key Observation

Why DETR-based Methods Underperform on Short Moments?

Data Perspective Analysis

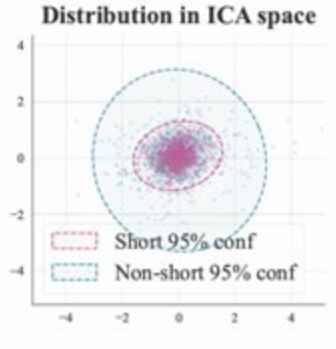
- Analyze visual features in ICA

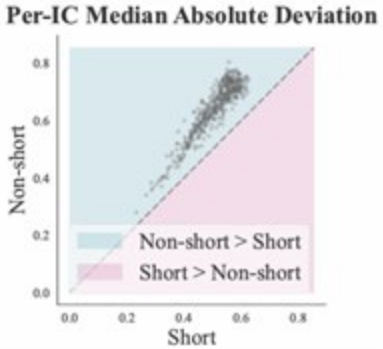
  → Short moments does not capture a wide range of visual features

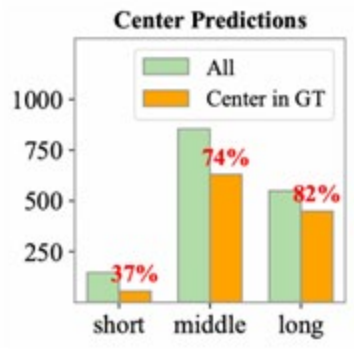
Model Perspective Analysis

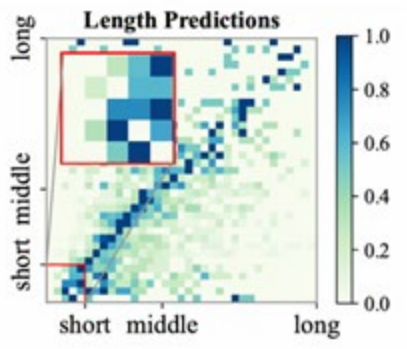
- Evaluated the prediction tendencies by assessing the center and length predictions

  → Inaccuracies in center prediction are a significant source of overall error



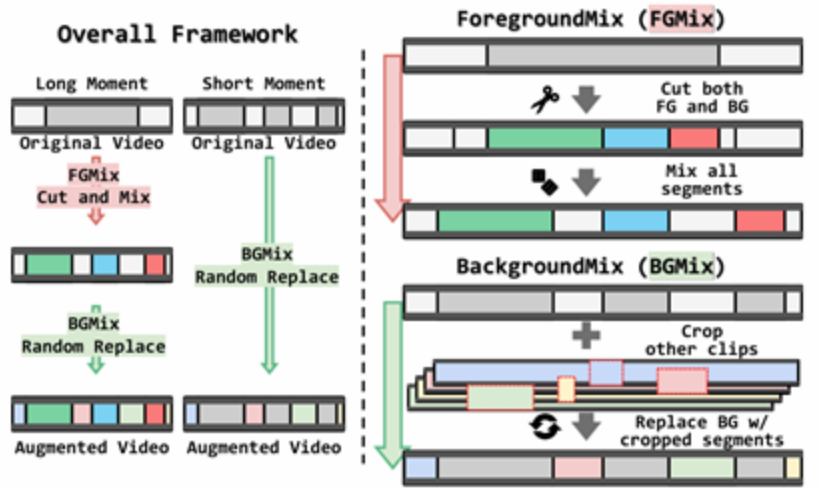






MomentMix: Augmentation for Short Moment

Data Perspective Solution



First Stage: ForegroundMix

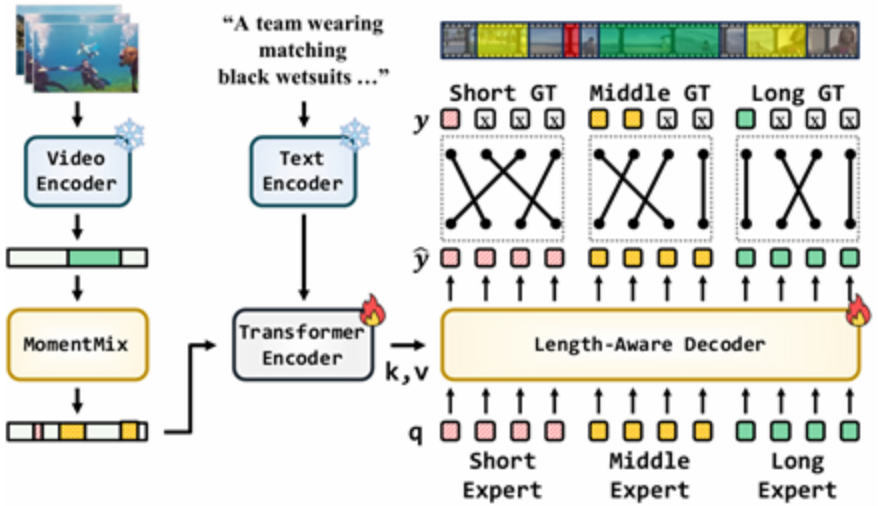
- Goal
  - Increase the visual diversity of foreground features in short moment
- Generate short moments by cutting and mixing longer foreground samples

Second Stage: BackgroundMix

- Goal
  - Improve the diversity of the visual background features to strengthen the association between foreground visual features and the text query
- Recombine backgrounds from various video clips

Length-Aware Decoder

Model Perspective Solution



Decoder Queries with Class-Pattern

- Predefine length classes (e.g., short, middle, long)
- Uniformly assign each decoder query to these length-specific classes

Length-wise Matching

- Modify the bipartite matching process
- Queries are matched with ground-truth moments within the same length class

→ Create "lengthwise expert" queries with length-wise matching

Main Results

Performance with respect to Moment Length on QVHighlights

- Significantly improves short-moment performance across all baselines

Method	Short		Middle		Long		Full	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
CG-DETR [25]	5.70	9.61	42.25	45.23	44.58	64.35	44.10	42.86
TaskWeave‡ [42]	2.91	6.31	38.93	44.00	47.58	52.67	42.45	43.33
R^2-Tuning‡ [23]	6.83	11.96	38.86	46.00	49.56	54.23	44.16	46.10
FlashVTG [2]	10.35	14.84	41.62	49.04	47.09	51.68	45.84	47.59
QD-DETR [26]	3.95	6.98	37.39	41.12	42.86	46.95	40.01	39.84
+Ours	10.33 (+6.38)	13.79 (+6.81)	41.32 (+3.93)	46.94 (+5.82)	45.40 (+2.54)	52.45 (+5.50)	45.02 (+5.01)	46.61 (+6.77)
TR-DETR [32]	4.95	8.22	40.08	43.27	**47.63**	50.80	43.70	42.62
+Ours	10.30 (+5.35)	14.57 (+6.35)	42.54 (+2.45)	47.90 (+4.63)	47.61 (-0.02)	**53.58** (+2.78)	46.59 (+2.89)	47.77 (+5.15)
UVCOM [36]	5.28	10.67	41.81	44.90	44.95	48.37	43.85	43.18
+Ours	**11.58** (+6.31)	**16.35** (+5.68)	**43.22** (+1.41)	**52.62** (+7.72)	46.48 (+1.53)	52.62 (+4.25)	**46.92** (+3.07)	**48.21** (+5.03)

Table 1. Performance gains of our method on the QVHIGHLIGHTS *test* set across different moment lengths. ‡ means test results from checkpoint provided by authors.

Overall Performance

- Our method yields significant improvements across all metrics, indicating enhanced overall performance across all baselines.

Method	MR						HD	
	R1			mAP			≥ Very Good	
	@0.5	@0.7	Avg.	@0.5	@0.75	Avg.	mAP	HIT@1
M-DETR [18]	52.89	33.02	-	54.82	29.40	30.73	35.69	55.60
UMT † [22]	56.23	41.18	-	53.83	37.01	36.12	38.18	59.99
EaTR [13]	57.98	42.41	-	59.95	39.29	39.00	-	-
UniVTG [20]	58.86	40.86	-	57.60	35.59	35.47	38.20	60.96
CG-DETR [4]	65.43	48.38	44.10	64.51	42.77	42.86	40.33	66.21
MomentDiff [19]	57.42	39.66	-	54.02	35.73	35.95	-	-
TaskWeave‡ [42]	61.87	46.24	42.45	63.75	43.63	43.33	37.87	59.08
BAM-DETR [17]	62.71	48.64	-	64.57	46.33	45.36	-	-
R^2-Tuning‡ [23]	**66.08**	48.90	44.16	**68.09**	47.65	46.10	39.18	64.20
FlashVTG [2]	**66.08**	50.00	45.84	67.99	48.70	47.59	41.07	66.15
QD-DETR [26]	61.22	44.49	40.01	62.31	39.45	39.84	39.01	62.13
+Ours	63.62 (+2.40)	48.77 (+4.28)	45.02 (+5.01)	65.50 (+3.19)	47.78 (+8.33)	46.61 (+6.77)	40.35 (+1.34)	64.46 (+2.33)
TR-DETR [32]	64.66	48.96	43.70	63.98	43.73	42.62	39.91	63.42
+Ours	65.63 (+0.97)	**51.23** (+2.27)	46.59 (+2.89)	66.89 (+2.91)	49.04 (+5.31)	47.77 (+5.15)	**41.54** (+1.63)	66.02 (+2.60)
UVCOM [36]	63.55	47.47	43.85	63.37	42.67	43.18	39.74	64.20
+Ours	65.37 (+1.82)	50.71 (+3.24)	**46.92** (+3.07)	66.65 (+3.28)	**49.22** (+6.55)	**48.21** (+5.03)	40.91 (+1.17)	**66.54** (+2.34)

Table 2. Performance comparison on QVHIGHLIGHTS *test* set. † indicates training with additional audio features. ‡ means test results from checkpoint provided by authors.

Method	CHARADES-STA		TACOS		CHARADES-STA‡	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7	R1@0.5	R1@0.7
SAP [7]	-	-	-	-	27.42	13.36
SM-RL [33]	-	-	-	-	24.36	11.17
MAN [39]	-	-	-	-	41.24	20.54
2D-TAN [43]	46.02	27.50	27.99	12.92	40.94	22.85
VSLNet [41]	42.69	24.14	23.54	13.15	-	-
M-DETR [18]	53.63	31.37	24.67	11.97	-	-
QD-DETR [26]	57.31	32.55	-	-	52.77	31.13
UniVTG [20]	58.01	35.65	34.97	17.35		
TR-DETR [32]	57.61	33.52	-	-	53.47	30.81
BAM-DETR [17]	59.83	39.83	41.54	26.77	-	-
R^2-Tuning [23]	-	-	38.72	25.12	-	-
FlashVTG [2]	60.11	38.01	41.76	24.74	54.25	**37.42**
UVCOM [36]	59.25	36.64	36.39	23.32	54.57	34.13
+Ours	**61.45** (+2.20)	**40.22** (+3.58)	**42.21** (+5.82)	**28.02** (+4.70)	**56.16** (+1.59)	36.10 (+1.97)

Table 3. Results on CHARADES-STA and TACOS *test* set. ‡ indicates training with VGG features and GloVe features.

Main Results

Component Analysis

- To examine the impact of MomentMix and the Length-Aware Decoder

→ While each component individually improves performance, their combined application leads to even greater improvements.

		Short		Middle		Long		Full	
MMix	*LAD*	R1	mAP	R1	mAP	R1	mAP	R1	mAP
✘	✘	4.57	7.77	38.89	43.10	42.62	47.44	41.06	41.00
✔	✘	6.48	11.44	42.96	46.88	44.15	49.57	44.66	44.68
✘	✔	8.76	11.01	40.55	45.53	43.69	50.76	43.65	44.48
✔	✔	**11.07**	**15.27**	**43.12**	**48.53**	**44.39**	**52.65**	**46.13**	**47.70**

Table 4. Performance comparison with baseline(QD-DETR) on QVHIGHLIGHTS *val* set. MMix, and LAD indicate MomentMix and Length-Aware Decoder, respectively.

Evaluation in Few-shot Scenarios

- To validate the effectiveness of MomentMix as a data augmentation technique, we conducted experiments using 50%, 20%, and 10% of the training data

→ MomentMix effectively generates new training samples by enhancing feature diversity

Method	R1			mAP		
	@0.5	@0.7	Avg.	@0.5	@0.75	Avg.
100% train data	61.39	46.18	41.06	61.68	41.57	41.00
50% train data	57.23	40.26	36.10	57.51	35.63	35.98
+ MomentMix	63.16 (+5.93)	47.74 (+7.48)	43.36 (+7.26)	61.91 (+4.40)	41.90 (+6.27)	41.73 (+5.75)
20% train data	46.84	30.45	26.58	48.27	25.35	26.88
+ MomentMix	52.45 (+5.61)	37.68 (+7.23)	33.69 (+7.11)	52.66 (+4.39)	34.25 (+8.90)	33.72 (+6.84)
10% train data	32.45	16.84	15.90	37.10	15.37	18.17
+ MomentMix	43.10 (+10.65)	28.71 (+11.87)	25.61 (+9.71)	44.97 (+7.87)	26.12 (+10.75)	26.62 (+8.45)

Table 6. Results on the QVHIGHLIGHTS *val* set using 50%, 20%, and 10% of the original training data.