The Devil is in the Spurious Correlations: Boosting Moment Retrieval with Dynamic Learning

Supplementary Material

S1. Details of Network and Objectives

In this section, we present our network and loss functions in detail.

Transformer encoder-decoder with prediction heads. We follow the architectural principles outlined in [3, 4] for the design of our transformer and prediction heads, with modifications introduced in the encoder. Specifically, we integrate our proposed *Temporal Dynamic Tokenizer* into the encoder to address spurious correlations effectively.

Given a spurious pair $p'_i = \{\tilde{V}'_i, \tilde{V}'_{k_i}\}$, this module processes all video pairs uniformly. As introduced in Section 3.2, we use T to denote the temporal dynamics. To incorporate these dynamics, we employ two transformer encoder layers with cross attention that facilitate bidirectional interactions: (1) between the temporal dynamics and the text query, and (2) between the video content and the text query. Once the temporal dynamics and video content are individually aligned with the text, we apply a weighted element-wise addition to combine their outputs. This rejected representation is subsequently processed through a standard transformer layer to refine the contextual understanding. Given a spurious pair $p_i = \{\tilde{V}_i, \tilde{V}_{k_i}\}$, our approach generates a refined spurious pair $p'_i = \{\tilde{V}'_i, \tilde{V}'_{k_i}\}$ that incorporates these temporal and semantic enhancements.

Loss Functions. We compute the loss between the predicted output $\hat{y}$ and its corresponding ground truth y (m_i) for $\tilde{V}''_i$, as well as between $\hat{y}'$ and its ground truth y' (m'_i) for $\tilde{V}''_{k_i}$. The predictions are matched with their targets, and the loss is calculated using L1 loss, generalized IoU (gIoU) loss, and cross-entropy loss, respectively, as described in [3].

S2. Sensitiveness Analysis

S2.1. Video Synthesizer for Dynamic Context

In Section 3.1, we construct a new sample $\tilde{V}_{k_i}$ with dynamic context from spurious pair $p_i = \{V_i, V_{k_i}\}$ as follows,

$$\tilde{V}_{k_i} = \alpha \cdot V_i + (1 - \alpha) \cdot V_{k_i}, \tag{10}$$

where α represents the sampling ratio of V_i while $1 - \alpha$ corresponds to V_{k_i}.

We examine the impact of the sampling ratio α on the quality of the synthesized samples. In detail, we adapt α ranging from 0.1 to 0.9 with a step size of 0.2.

As illustrated in Table S1, when the sampling ratio α increases, the synthesized video incorporates more tokens from the videos containing the target moments with corresponding dynamic contexts, thus improving the performance of moment retrieval. The performance starts to decline from $\alpha = 0.9$, due to the lack of dynamics of the contexts. Specifically, when $\alpha = 1.0$, the synthesized video is identical to the original video. This ablation study on α demonstrates the effectiveness of our *Video Synthesizer for Dynamic Context* in improving model performance by balancing contextual information and target moment focus. Besides, even with various sampling ratios α, our method still achieves promising results, which demonstrate the robustness of the proposed method.

α	MR-R1		MR-mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
0.0	11.61	3.35	23.93	7.5	10.09
0.3	65.10	51.94	65.77	48.13	47.55
0.5	64.77	51.10	66.79	49.08	47.95
0.7	65.88	53.67	66.43	49.86	49.05
0.9	64.19	51.23	66.29	48.88	47.94

Table S1. Sensitiveness analysis of sampling ratio α on QVHighlights *val* split.

β	MR-R1		MR-mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
0.1	65.10	51.94	67.37	50.12	48.87
0.3	64.77	51.48	66.50	49.76	48.48
0.5	65.15	51.26	66.24	48.44	47.81
0.7	65.88	53.67	66.43	49.86	49.05
0.9	62.97	50.19	65.81	48.76	47.83

Table S2. Sensitiveness analysis of sampling ratio β on QVHighlights *val* split.

S2.2. Dynamics Enhancement

In section 3.2, the model learns from both dynamic and video information via cross-attention machines. To emphasize the learned dynamic information, we inject text-guided dynamic representation T' into video $\tilde{V}_i$ as follows,

$$\tilde{V}'_i = \beta \cdot \tilde{V}_i + (1 - \beta) \cdot V_i, \tag{11}$$

where β represents the injection ratio of the video information we used, while $1 - \beta$ corresponds to temporal information T'. We also examine the impact of the injection ratio β on the quality of the injected videos. In detail, we adapt β

Method	QVHighlights val					Charades-STA test	
	R1@0.5	R1@0.7	mAP@0.5	mAP@0.75	mAP	R1@0.5	R1@0.7
QD-DETR	68.58	52.13	67.87	45.94	45.40	66.63	42.78
CG-DETR	70.27	55.62	69.17	52.62	50.93	69.11	46.13
BAM-DETR	69.72	55.13	69.38	52.89	51.13	68.49	48.33
TD-DETR **(ours)**	**71.29**	**57.23**	**72.99**	**54.94**	**53.23**	**73.49**	**53.01**

Table S3. Comparison of models performance on QVHighlights *val* split using InternVideo2 feature representations.

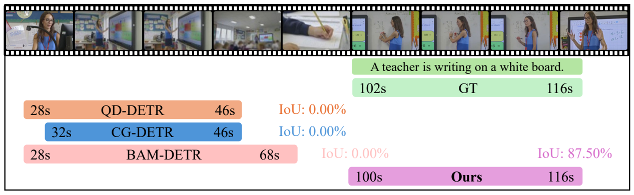

Figure S1. Model prediction on query "A teacher is writing on a whiteboard.". Baselines tend to predict the teacher writing on a screen instead of the target moment which indicates baseline models fail to distinguish between "screen" and "whiteboard".

ranging from 0.3 to 0.9 with a step size of 0.2 and evaluate $\beta = 0$ as an extra experiment. As illustrated in Table S2, when the injection ratio β decreases, the video is injected with more temporal information, thus improving the performance of moment retrieval. The performance achieves the highest performance when $\beta = 0.7$, which indicates the benefits of dynamic enhancement. Specifically, when $\beta = 1.0$, no dynamic information is injected into the video, thus the performance drops a lot in contrast to those with dynamics representation. Note that when $\beta = 0.0$, the model relies solely on temporal dynamic information, which leads to poor predictions due to the absence of any object-related cues. This ablation study on β validates the effectiveness of our *Temporal Dynamics Enhancement* in boosting moment retrieval by encouraging our model to align text queries with temporal-dynamic representations. Besides, even with various sampling ratios β, our method still achieves promising results, which demonstrate the robustness of the proposed method.

Method	QVHighlights val			Charades-STA	
	R1@0.7	mAP@0.75	mAP	R1@0.5	R1@0.7
baseline	46.66	41.82	41.22	57.31	32.55
w/ random	51.29	47.82	47.56	58.66	37.98
w/ dissimilarity	47.23	46.83	45.98	59.25	38.28
w/ similarity	**53.67**	**49.86**	**49.05**	**60.89**	**40.35**

Table S4. Comparisons across different sampling strategies. Mix indicates a mixed selection between similarity and dissimilarity.

Method	Para (M)	GFLOPs
BAM-DETR	14.50	60.32
Ours	13.89	69.52

Table S5. Comparisons on computational cost.

S2.3. Computational costs

Tab. S5, shows that our method achieves strong performance improvements while using fewer parameters and a minor increase in GFLOPs. Besides, our method dynamically synthesizes videos only during training, so there is no additional storage involved.

S3. Ablation Analysis on Video Sampling Strategy

In Section 1.1, we select a video that is contextually similar to V to ensure both the challenge and rationality of the synthesized video. As shown in Table S4, we compare our similarity-based selection strategy with a random sampling approach on QVHighlights and Charades-STA. The *w/ random* selection still outperforms the QD-DETR baseline but falls short of *w/ similarity*, demonstrating the effectiveness of our approach in generating meaningful and challenging synthetic video contexts.

S4. Additional Results of Predicted Results

S4.1. More Prediction Examples.

More visualization results of predictions and baselines comparison from our proposed *TD-DETR* model are presented in Figure S1 and Figure S2.

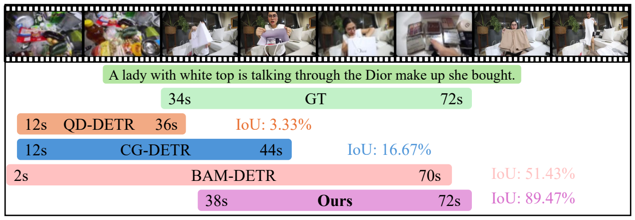

Figure S2. Model prediction on query "A lady with white top is talking through the Dior make-up she bought.". Baselines tend to predict the woman with some food and clothes instead of the target moment which indicates baseline models fail to distinguish between "make-up" and "clothes".

S4.2. New Validation Split on Spurious Correlation

Except for Spurious R@1 and Spurious mAP, we introduce a new validation split based on the QVHighlight validation set to further evaluate spurious correlations. Specifically, similar to Section 1.1, we replace the contextual frames of

Method	Standard R1 ↑			Standard mAP ↑		
	@0.5	@0.7	mIOU	@0.5	@0.75	Avg.
QD-DETR	58.29	39.29	52.76	57.25	34.86	34.78
CG-DETR	62.03	43.77	56.57	59.9	38.3	38.48
BAM-DETR	59.74	41.87	54.95	60.05	39.5	39.24
TD-DETR	**65.77**	**46.94**	**59.43**	**64.5**	**42.13**	**42.21**

Table S6. Performance comparison on our dynamic context validation split.

a video with clips from another video, creating a more dynamic and diverse context. This modification aims to disrupt excessive contextual associations and better assess the model's robustness against spurious correlations. The modified validation split will be released publicly with our code.

All illustrated in Table S6, our proposed *TD-DETR* still achieves state-of-the-art performance among all baselines on such a dynamic context validation.

S5. Generalization across Different Feature Representations

With the rapid advancement of large multi-modal models in video understanding, InterVideo2—a video foundation model introduced by [6]—has demonstrated strong capabilities in moment retrieval. Beyond SlowFast [2], we further evaluate our model's generalization across different feature representations. All illustrate in Table S3, the proposed *TD-DETR* also achieve state-of-the-art performance. Our TD-DETR outperforms the previous state-of-the-art model by a substantial margin, with improvements of up to 3.81% in R@1@0.7 and 3.88% in mAP@0.75 on QVHighlights *val* split and 7.30% in R@1@0.5 and 9.68% in R1@0.7 on Charades-STA *test* split.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 3

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 3

[3] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, pages 11846–11858, 2021. 3, 1

[4] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*, pages 23023–23033, 2023. 3, 1

[5] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019. 3

[6] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, pages 396–416. Springer, 2024. 3