

Supplementary Material for Submission 2743

Saliency-Guided DETR for Moment Retrieval and Highlight Detection

A. SG-Attention: Computational Efficiency

SG-Attention only minimally increases computational cost compared to standard dot-product attention. In the conventional setup, the main cost arises from two matrix multiplications: QK^\top and $\text{AttnWeights} \times V$. Each multiplication requires $2L^2E$ multiply-add operations, leading to a total of $4L^2E$. By contrast, SG-Attention performs one additional elementwise multiplication of size L^2 , which is negligible when E is large. For instance, with $L = 76$ and $E = 256$, standard attention incurs approximately 5.915×10^6 operations, whereas SG-Attention adds only about 5.776×10^3 operations (less than 1% overhead). Consequently, this saliency mechanism imposes only a minor computational burden and does not impede real-time performance.

B. Comparative Study of Attention Mechanisms

We evaluate and compare three distinct attention mechanisms for fusing textual information into video feature representations: standard Cross-Attention, the Adaptive Cross-Attention proposed by CGDETR [1], and our novel Saliency-Guided Attention. The results for the baseline architecture, trained on the QVHighlights dataset with each of the three attention variants, are summarized in Tab. B.1.

C. Training Objectives

Our model is trained using losses categorized into three main groups.

Highlight Detection Task We employ margin ranking, rank contrastive, and cross-entropy (CE) losses on both local and global saliency scores, as defined in [1]. Additionally, we apply the CE loss to negative text-video pairs following [2] to suppress negative clip saliency. The total loss for the task is represented as:

$$\mathcal{L}_{hl} = \mathcal{L}_{\text{marg}} + \mathcal{L}_{\text{rctl}} + \mathcal{L}_{\text{bce.pos}} + \mathcal{L}_{\text{bce.neg}} \quad (\text{C.1})$$

Moment Retrieval Task For the MR task, we use CE loss, generalized IoU (GIoU) loss [3], and smooth L1 loss to train both DETR and ATSS detection heads. For the ATSS

head [6], we also incorporate Centerness Loss [5] to enhance localization precision. In the DETR head, CE loss is applied to IoU head predictions. Losses for auxiliary DETR queries are computed similarly to primary queries. The objectives for the task are:

$$\mathcal{L}_{\text{detr}} = \lambda_{\text{LI}} \mathcal{L}_{\text{LI}}(m, \bar{m}) + \lambda_{\text{gIoU}} \mathcal{L}_{\text{gIoU}}(m, \bar{m}) + \lambda_{\text{CE}} \mathcal{L}_{\text{CE}}(y, \bar{y}) + \lambda_{\text{IoU}} \mathcal{L}_{\text{CE}}(\text{IoU}, \bar{\text{IoU}}) \quad (\text{C.2})$$

$$\mathcal{L}_{\text{atss}} = \lambda_{\text{LI}} \mathcal{L}_{\text{LI}}(m, \bar{m}) + \lambda_{\text{gIoU}} \mathcal{L}_{\text{gIoU}}(m, \bar{m}) + \lambda_{\text{CE}} \mathcal{L}_{\text{CE}}(y, \bar{y}) + \lambda_{\text{Centrness}} \mathcal{L}_{\text{CE}}(c, \bar{c}) \quad (\text{C.3})$$

In this context, the ground truth values are represented as $m = (m_c, m_\sigma)$, y , c , and IoU , where m_c and m_σ denote the center and duration of the ground-truth moment, y represents the binary classification label, c is the target centerness score, and IoU refers to the target IoU score. Similarly, the predicted values are denoted as $\bar{m} = (\bar{m}_c, \bar{m}_\sigma)$, \bar{y} , \bar{c} , and $\bar{\text{IoU}}$, corresponding to the predicted moment, binary classification label, centerness score, and predicted IoU score, respectively.

Auxiliary Losses The third category includes auxiliary losses to enhance the model’s overall performance. First, alignment loss ensures consistency between moment and sentence token. Additionally, CE loss is applied to differentiate moment tokens from non-moment tokens within each video instance. Detailed descriptions of these losses can be found in [1]. The overall objective of the group is:

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{align}} \quad (\text{C.4})$$

Overall Objective The final objective function is the sum of all the aforementioned losses:

$$\mathcal{L}_{\text{obj}} = \lambda_{\text{atss}} \mathcal{L}_{\text{atss}} + \lambda_{\text{detr}} \mathcal{L}_{\text{detr}} + \lambda_{\text{hl}} \mathcal{L}_{\text{hl}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}} \quad (\text{C.5})$$

D. Qualitative Results and Analysis

In Fig. E.1, we present visualizations of predictions made by the models on the QVHighlights validation dataset. Compared to existing methods such as CG-DETR [1] and TR-DETR [4], SG-DETR consistently achieves more precise and coherent highlight detection results, as evidenced by improvements in both retrieval accuracy and highlight score distributions.

Attention Type		MR					HD	
		R1		mAP			\geq Very Good	
		@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
(a)	CA	70.1 \pm 0.3	53.7 \pm 0.5	70.0 \pm 0.1	49.3 \pm 0.2	48.3 \pm 0.4	41.9 \pm 0.1	68.6 \pm 0.8
(b)	ACA	67.7 \pm 0.6	51.5 \pm 0.7	68.22 \pm 0.2	47.7 \pm 0.8	47.1 \pm 0.7	42.0 \pm 0.1	69.41 \pm 0.5
(c)	SGCA	71.1 \pm 0.5	56.5 \pm 0.8	71.0 \pm 0.4	52.2 \pm 0.7	50.1 \pm 0.6	42.2 \pm 0.1	70.1 \pm 0.6

Table B.1. Ablation study of attention types on QVHighlights val split. CA, ACA, and SGCA stand for Cross Attention, Adaptive Cross-Attention, and Saliency-Guided Cross-Attention, respectively. Results are reported as mean \pm standard deviation, averaged over three runs.

E. Visualization of the Pretrain Annotation Framework

To clearly illustrate how the InterVid-MR pretraining dataset was created, we present key statistics related to the annotation generation process in Fig. E.2. Although the annotation process is relatively straightforward, the visualization confirms that the resulting annotations consistently achieve high quality.

References

- [1] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023. 1
- [2] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023. 1
- [3] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 1
- [4] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. *arXiv preprint arXiv:2401.02309*, 2024. 1
- [5] Z Tian, C Shen, H Chen, and T He. Fcos: Fully convolutional one-stage object detection. arxiv 2019. *arXiv preprint arXiv:1904.01355*, 2019. 1
- [6] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 1

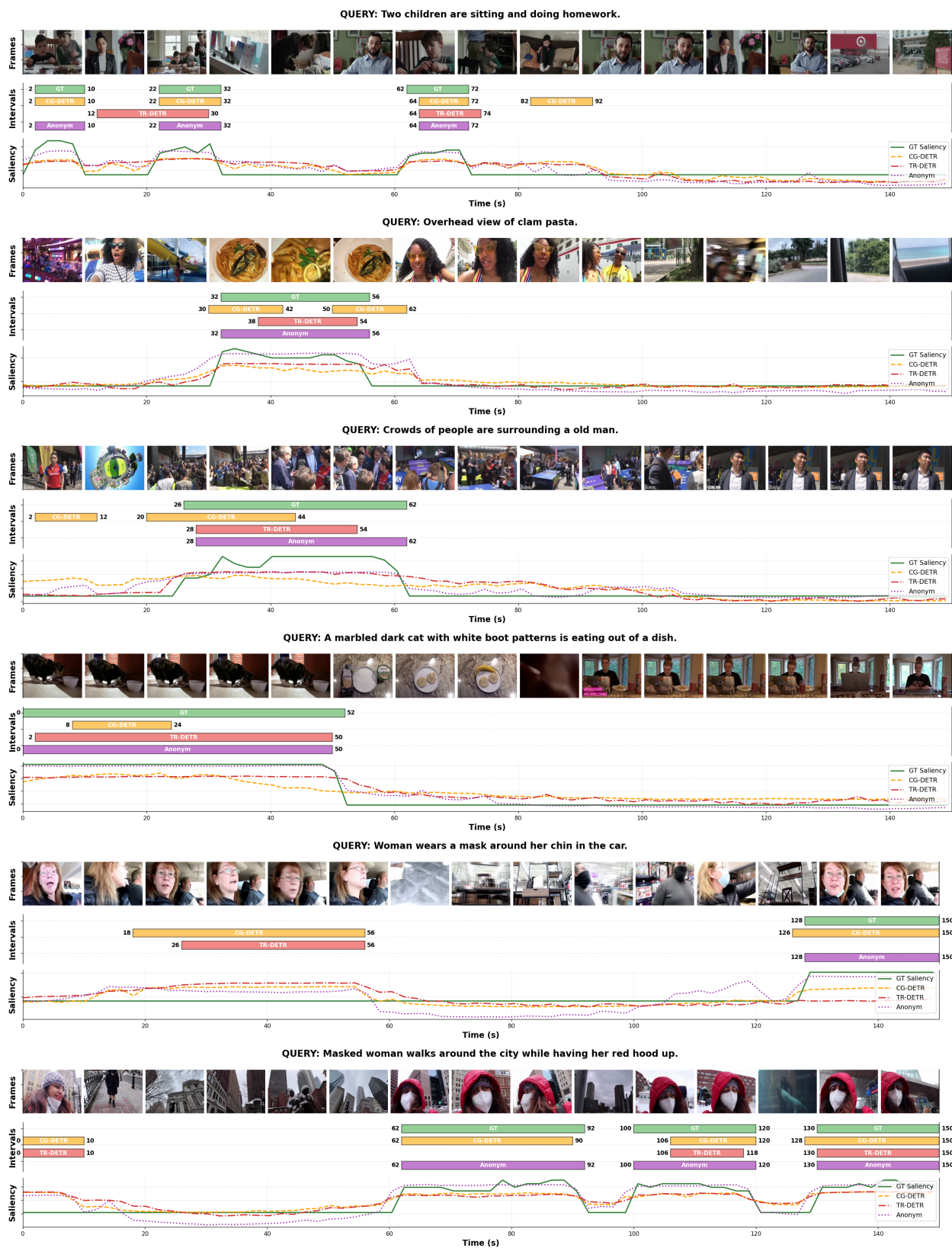


Figure E.1. Qualitative results



Figure E.2. Design of the Pretrain Annotation