BAM-DETR: Boundary-Aligned Moment Detection Transformer for Temporal Sentence Grounding in Videos – Supplementary Materials –

Pilhyeon Lee[1⋆] and Hyeran Byun[2]

[1] Department of Artificial Intelligence, Inha University
[2] Department of Computer Science, Yonsei University

1 Formulation of Saliency Losses

As mentioned in the main paper, we adopt saliency losses for effective multimodal alignment in the encoder as in the common practice [3, 5]. In specific, our total saliency-based loss is composed of three losses, *i.e.*, $\mathcal{L}_{\text{sal}} = \mathcal{L}_{\text{margin}} + \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{neg}}$. The margin-based loss $\mathcal{L}_{\text{margin}}$, defined in Eq. (2) of the main paper, aims to encourage the model to produce higher saliency scores for the clips relevant to the given sentence compared to less related clips. Meanwhile, the rank-aware contrastive loss $\mathcal{L}_{\text{cont}}$ is utilized to preserve the ground-truth clip ranking in predicted saliency scores. To be concrete, we first define the positive and negative sets based on an arbitrary reference score r, *i.e.*, clips whose saliency score labels are higher than r belongs to the positive set $\mathcal{B}_r^+$, and the remaining clips constitute the negative set $\mathcal{B}_r^-$. The rank-aware contrastive loss is then formulated using a set of reference scores $\mathcal{R}$ as follows.

$$\mathcal{L}_{\text{cont}} = -\sum_{\forall r \in \mathcal{R}} \log \frac{\sum_{\forall \hat{v} \in \mathcal{B}_r^+} \exp(S(\hat{v})/\tau)}{\sum_{\forall \hat{v} \in (\mathcal{B}_r^+ \cup \mathcal{B}_r^-)} \exp(S(\hat{v})/\tau)}, \tag{1}$$

where $S(\cdot)$ is a learnable saliency score predictor and τ is a temperature (set to 0.5). We define $\mathcal{R}$ to be the set of saliency score labels of positive clips within ground-truth moments.

The negative relation loss is based on the assumption that all video clips should exhibit low saliency scores when paired with unmatched (negative) sentences. Formally, the loss can be defined as follows.

$$\mathcal{L}_{\text{neg}} = -\sum_{\forall \hat{v}^{\text{neg}} \in \hat{\mathcal{V}}^{\text{neg}}} \log(1 - S(\hat{v}^{\text{neg}})), \tag{2}$$

where $\hat{\mathcal{V}}^{\text{neg}}$ denotes the memory features obtained by processing the video with a negative sentence through the encoder. In our implementation, a negative sentence is sampled from a different video-sentence pair in the mini-batch.

⋆ Correspondence to: Pilhyeon Lee <pilhyeon.lee@inha.ac.kr>

Table 1: Results on the QVHighlights validation split.

Method	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
Deformable DETR	60.52	49.35	62.53	46.41	44.73
BAM-DETR (Ours)	65.10	51.61	65.41	48.56	47.61

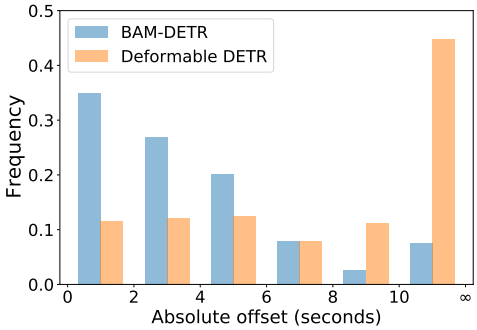


Fig. 1: Offset histogram on the QVHighlights validation split.

2 Comparison with Deformable DETR

The proposed boundary-focused attention layer incorporates deformable attention, which is first proposed in Deformable DETR [6]. It was originally designed for computationally efficient global attention with multi-scale features in object detection. In contrast, we employ deformable attention for local aggregation of neighbor features, aiding precise boundary prediction in temporal sentence grounding. To elucidate the discrepancy in their roles, we conduct a comparative experiment. For this study, we implement a 1D variant of single-scale Deformable DETR, tailored for temporal sentence grounding. We apply the proposed quality-based scoring to this model for a fair comparison. Its key differences with our BAM-DETR lie in the moment formulation (center-based *vs.* boundary-oriented) and the design of decoding layers (single-pathway *vs.* dual-pathway).

To analyze the behavior of deformable attention, we look into the absolute values of predicted offsets, *i.e.*, how distant features are referenced during the attention process. These offsets can indicate whether the attention is responsible for global or local interaction. Fig. 1 presents a visual comparison between the normalized histograms of predicted offsets from two comparative methods. In our BAM-DETR, the deformable attention primarily concentrates on the neighbor features near the boundaries, *e.g.*, over 80% of offsets are shorter than 5 seconds. Conversely, in the case of Deformable DETR, the deformable attention strives to aggregate global information, *e.g.*, about 45% of offsets are longer than 10 seconds. These results clearly confirm the different roles of deformable attention in the two models. In addition, we compare the grounding performance in Table 1, where our BAM-DETR substantially outperforms Deformable-DETR. This underscores the importance of our boundary-oriented moment modeling as well as the design of dual-pathway decoding layers.

Table 2: Ablation study on the loss functions on QVHighlights.

$\mathcal{L}_{\text{loc}}$	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{qual}}$	$\mathcal{L}_{\text{sal}}$	$\mathcal{L}_{\text{regul}}$	R1		mAP		
					@0.5	@0.7	@0.5	@0.75	Avg.
✓	✓				56.77	41.03	58.63	39.25	39.12
✓	✓		✓		60.58	46.65	62.09	43.83	42.94
✓	✓		✓	✓	63.61	50.26	63.01	44.98	44.16
✓		✓			59.23	46.13	60.52	44.80	43.48
✓		✓	✓		63.23	50.00	64.03	47.42	46.64
✓		✓	✓	✓	65.10	51.61	65.41	48.56	47.61

3 Details of Boundary Alignment Evaluation

We provide more details regarding the experimental setup of boundary alignment evaluation performed in Fig. 4a of the main paper. Inspired by the trimap evaluation of DeepLab [1], we propose a novel metric of boundary hit rate under varying band widths to evaluate the degree of boundary alignment. In detail, we expand boundary points of the n-th ground truth $\{t_{s_n}, t_{e_n}\}$ with a given *band width* of l_w to form boundary zones. We can denote the starting and ending zones by $Z_{s_n} = [t_{s_n} - 0.5l_w, t_{s_n} + 0.5l_w]$ and $Z_{e_n} = [t_{e_n} - 0.5l_w, t_{e_n} + 0.5l_w]$, respectively. Then, for the m-th proposal $\{\hat{t}_{s_m}, \hat{t}_{e_m}\}$, we check whether both of its boundaries fall in the corresponding zones. We iterate this process for all combinations of ground truths and predictions, and mark a video as correct if any pair is positive. Formally, the binary variable of h of a video is defined as:

$$h = \max_{\forall n,m} \Big[\text{Hit}^s(n,m) \cdot \text{Hit}^e(n,m) \Big],$$
$$\text{where} \quad \text{Hit}^z(n,m) = \mathbb{1}\big[|\hat{t}_{z_m} - t_{z_n}| \leq 0.5l_w\big], \quad z \in \{s, e\}.$$

Note that we measure the hit rate over the whole validation set.

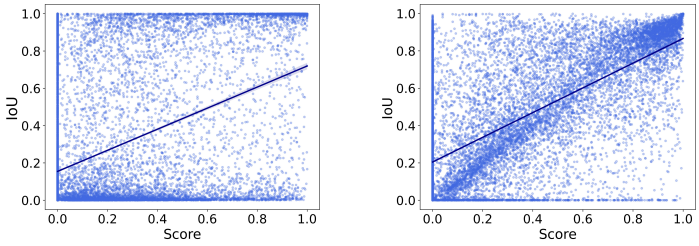
4 More Analyses

Ablation study on loss functions. Our model employs several loss functions for training. We conduct an ablative experiment to diagnose their effects. Table 2 summarizes the results, where the upper part adopts the typical classification-based scoring whereas the lower one leverages our proposed quality-based scoring. We first examine the benefit of saliency losses. Consistent with the recent findings [3], we observe that the saliency losses effectively guide the cross-modal alignment in the encoder, leading to notable performance improvements. Then we investigate the importance of our regularization loss designed for boundary-sensitive feature construction (*cf*., Eq. (5) of the main paper). It can be observed that regardless of the choice of scoring methods, the boundary regularization leads to significant performance boosts. Putting together the results in Table 6b of the main paper, it becomes clear that boundary-sensitive features are essential for precise boundary updating. Lastly, the comparison between the two separate parts validates the efficacy of our quality-based scoring, especially in terms of mAPs.

Table 3: Generalizability evaluation of quality-based scoring on the QVHighlights validation split.

Method	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
Moment-DETR‡ [3]	53.23	34.00	54.80	29.02	30.58
+ quality-based scoring	56.77	38.65	55.09	35.30	34.98
QD-DETR‡ [5]	62.90	46.77	62.66	41.51	41.24
+ quality-based scoring	64.26	50.32	63.79	46.03	44.50
EaTR‡ [2]	57.74	42.71	59.40	39.34	39.06
+ quality-based scoring	59.42	45.61	60.24	42.29	41.61

‡All models are reproduced by official codebase



(a) Classification-based scoring **(b)** Quality-based scoring

Fig. 2: Correlation between scores and IoUs with ground truths: (a) the classification scores show a moderate correlation (Pearson's r of 0.44); (b) the quality scores exhibit a stronger correlation (Pearson's r of 0.67).

Comparison between scoring methods. We present the quality-based scoring method to replace the conventional classification-based one. To compare two scoring methods, we draw scatter plots of scores *vs.* IoUs with ground truths using all predictions on the QVHighlights validation set. Fig. 2a shows that classification scores correlate with IoUs to an extent. On the other hand, we observe in Fig. 2b that our quality-based scoring shows a much stronger correlation with IoUs. These results validate its efficacy in estimating the localization qualities of proposals, indicating that it is more appropriate for proposal ranking.

Generalizablity of the quality-based scoring. By design, our quality-based scoring method is generalizable to any query-based approach. To investigate this property, we conduct experiments by adopting the quality-based scoring on top of three representative models: Moment-DETR [3], QD-DETR [5], and EaTR [2]. The results are shown in Table 3, where the proposed scoring method brings consistent improvements over different baselines. Noticeably, we can observe the pronounced gains at high IoU thresholds, which indicates better alignment of proposals with the ground truths. This corroborates our claim that moment proposals ought to be ranked based on their localization qualities rather than the degree of matching.

Efficiency comparison. We perform an efficiency comparison with previous state-of-the-art methods in terms of computational costs (# of FLOPs) and

Table 4: Efficiency comparison results on the QVHighlights validation split.

Method	R1		mAP			FLOPs	Params
	@0.5	@0.7	@0.5	@0.75	Avg.		
QD-DETR‡ [5]	62.90	46.77	62.66	41.51	41.24	0.59G	7.7M
UniVTG†‡ [4]	59.74	40.90	58.61	36.76	36.13	0.98G	43.4M
EaTR‡ [2]	60.90	46.13	62.01	42.17	41.43	0.47G	9.1M
BAM-DETR^{slim}	63.94	50.19	64.51	48.51	47.03	0.43G	7.2M
BAM-DETR	65.10	51.61	65.41	48.56	47.61	0.65G	9.5M

†The hidden dimension is four times larger than that of competitors
‡All models are reproduced by official checkpoints

memory (# of Parameters). The comparison results on the QVHighlights validation set are shown in Table 4. We can observe that our BAM-DETR has a comparable model size with EaTR [2]. In terms of localization performance, it outperforms all the existing approaches by large margins, especially under strict evaluation metrics, which is consistent with the test split results (*cf*., Table 2 of the main paper). To make a fairer comparison, we also implement a small variant of our model equipped with slimmer encoding layers, namely BAM-DETR^{slim}. In detail, we halve the hidden dimension of the encoder and reduce the number of fully-connected layers within each attention block. As a result, BAM-DETR^{slim} can achieve better efficiency with a cost of slightly sacrificing localization performance. Nevertheless, it is shown that BAM-DETR^{slim} suffices to largely surpass the existing approaches even with fewer parameters and FLOPs. These results confirm the effectiveness of the proposed method.

5 Further Qualitative Results

We perform further qualitative comparisons with previous query-based methods in Fig. 3 and Fig. 4. The comparison results across various scenarios demonstrate the superiority of our BAM-DETR over the strong competitors.

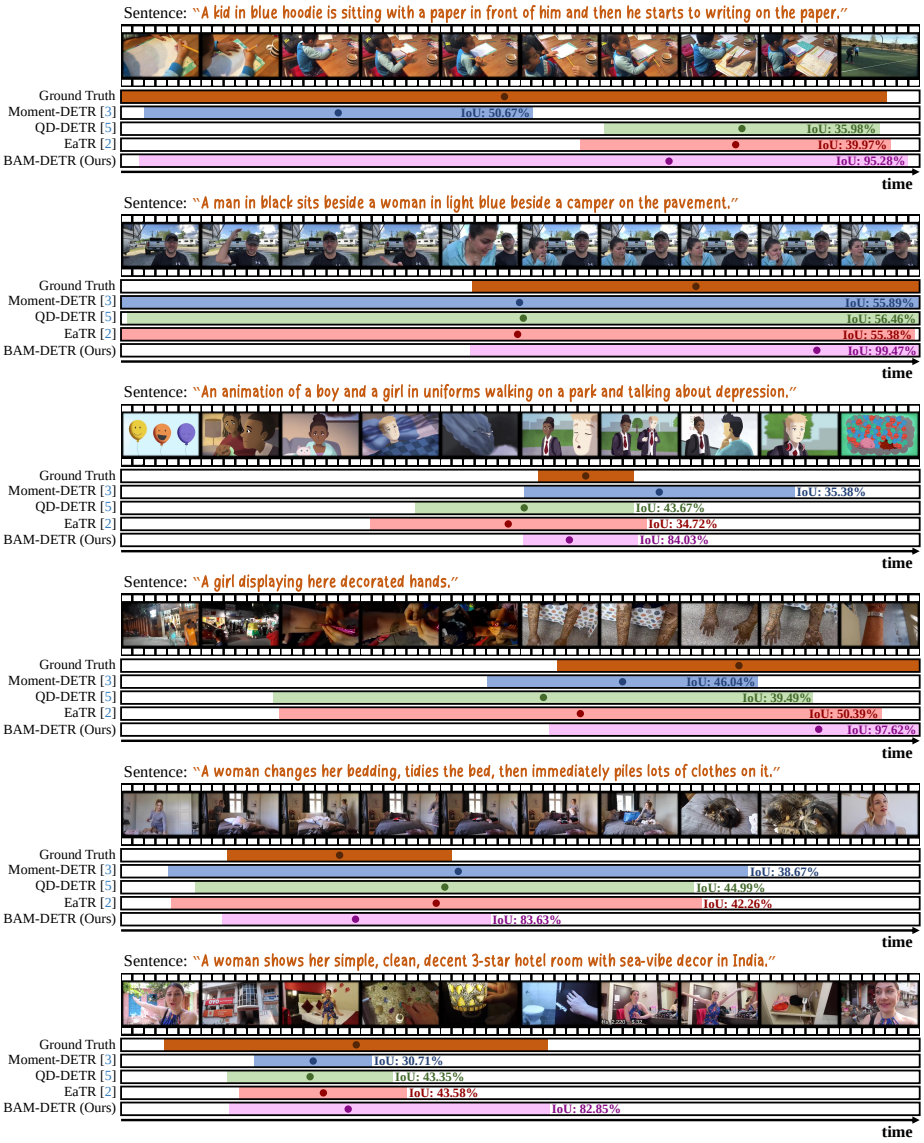


Fig. 3: Qualitative comparison on the QVHighlights validation split.

Fig. 4: Qualitative comparison on the QVHighlights validation split.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. PAMI **40**(4), 834–848 (2017) 3
2. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: ICCV. pp. 13846–13856 (2023) 4, 5
3. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. In: Neurips. vol. 34, pp. 11846–11858 (2021) 1, 3, 4
4. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: ICCV. pp. 2794–2804 (2023) 5
5. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: CVPR. pp. 23023–23033 (2023) 1, 4, 5
6. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2021) 2