Knowing Where to Focus: Event-aware Transformer for Video Grounding -Appendix-

Jinhyun Jang[1] Jungin Park[1] Jin Kim[1] Hyeongjun Kwon[1] Kwanghoon Sohn[1,2*]
[1]Yonsei University [2]Korea Institute of Science and Technology (KIST)
{jr000192, newrun, kimjin928, kwonjunn01, khsohn}@yonsei.ac.kr

In this document, we include supplementary materials for "Knowing Where to Focus: Event-aware Transformer for Video Grounding". We first provide more concrete implementation details of pseudo event timestamps generation (Sec. 1), and additional experimental results (Sec. 2), including ablation studies and qualitative results.

1. Pseudo event timestamps generation

We generate the pseudo event-level supervision (*i.e.*, pseudo event timestamps $\hat{\mathbf{P}}$ in Eq. (8)) to learn event reasoning. In this section, we describe the details of the pseudo event timestamps generation. While pseudo event timestamps generation is highly inspired by the prior work [2], which leverages the temporal self-similarity matrix (TSM), we detect pseudo events without any learnable parameters in an unsupervised manner.

Specifically, we first obtain the temporal self-similarity matrix $\mathbf{S} \in \mathbb{R}^{L_v \times L_v}$ by computing cosine similarity between video representations $\mathbf{h}_v$. Similar to [2], we define the contrastive kernel $\mathbf{Z} \in \mathbb{R}^{z \times z}$ with the kernel size $z = 5$ as follows:

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & 0 & -1 & -1 \\ 1 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \end{bmatrix} \tag{1}$$

Since the kernel is designed to imitate the boundary pattern in the TSM [2], we can obtain the boundary scores $\mathbf{b} \in \mathbb{R}^{L_v}$ by applying the convolution to the diagonal elements of the TSM. With the boundary scores $\mathbf{b}$, we remove the scores that are lower than the average boundary score $\bar{\mathbf{b}}$ and apply a sliding max filter with a size of 3 to filter out the consecutively distributed scores. The remaining indices are assumed to be the event boundary, and we define the pseudo event timestamps $\hat{\mathbf{P}}$ as the center coordinate and duration between each boundary index.

*Corresponding author

Table 1. Choice of attention mechanism for event reasoning on QVHighlights val split.

Methods	R1@0.5	R1@0.7	mAP	GFLOPs	Params
Cross-attention	$57.35_{\pm 1.4}$	$41.55_{\pm 1.2}$	$37.00_{\pm 1.0}$	0.49	10.1M
Slot attention	$61.36_{\pm 1.2}$	$45.79_{\pm 0.7}$	$41.74_{\pm 0.7}$	0.47	9.0M

Table 2. Performance with respect to the different number of iterations for slot attention on QVHighlights val split.

K	R1@0.5	R1@0.7	mAP	GFLOPs
1	$57.16_{\pm 1.6}$	$41.35_{\pm 0.9}$	$37.92_{\pm 0.7}$	0.467
2	$58.26_{\pm 1.2}$	$43.29_{\pm 0.6}$	$38.72_{\pm 0.8}$	0.469
3	$61.36_{\pm 1.2}$	$45.79_{\pm 0.7}$	$41.74_{\pm 0.7}$	0.472
4	$60.45_{\pm 1.3}$	$44.00_{\pm 0.7}$	$39.48_{\pm 0.6}$	0.474
5	$59.16_{\pm 1.2}$	$43.35_{\pm 0.6}$	$38.96_{\pm 0.6}$	0.476

2. Additional Experiments

In this section, we present additional component analysis on QVHighlights [4] (Sec. 2.1), ablation studies on Charades-STA [1] and ActivityNet Captions [3] (Sec. 2.2), and qualitative results for video grounding (Sec. 2.3).

2.1. Additional component analysis

We provide additional component analysis according to the choice of attention mechanism for event reasoning, the number of iterations K in slot attention, the number of transformer layers and qualitative analysis for the gated fusion transformer layer.

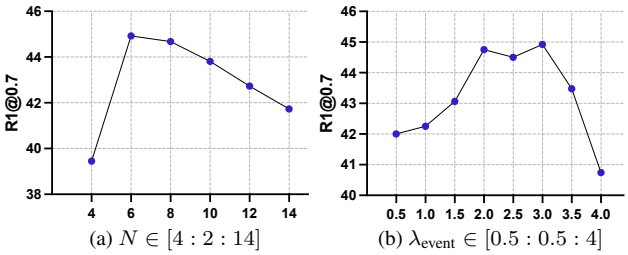
Slot attention vs. cross-attention. While we use the slot attention mechanism for event reasoning in the main paper, conventional cross-attention can be used as an alternative. The main difference between the slot and cross-attention is the attention normalization axis. In the cross-attention, the softmax normalization is applied over the input axis, making the attention values for each slot independent of each other. Contrary to this, the normalization along event slot direction as in the slot attention enables slots to compete and exchange information with each other to cover distinctive

Table 3. Comparison of models with different number of layers on QVHighlights val split. # layers indicate the number of transformer encoder-decoder layers used for the video grounding.

# layers	R1@0.5	R1@0.7	mAP	GFLOPs	Params
2	$60.90_{\pm 1.5}$	$44.06_{\pm 0.9}$	$38.91_{\pm 0.7}$	0.34	6.9M
3	$61.36_{\pm 1.2}$	$45.79_{\pm 0.7}$	$41.74_{\pm 0.7}$	0.47	9.0M
4	$61.68_{\pm 1.4}$	$45.90_{\pm 0.7}$	$41.78_{\pm 0.8}$	0.60	11.1M
5	$61.35_{\pm 1.4}$	$46.94_{\pm 0.8}$	$41.80_{\pm 0.6}$	0.73	13.2M

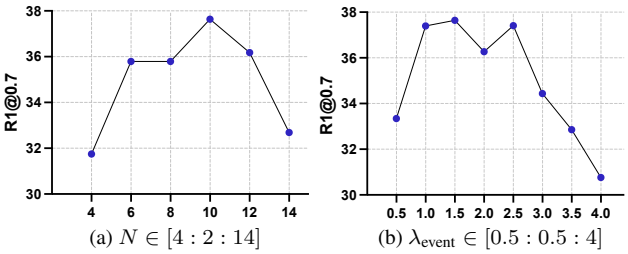
Table 4. Component ablation results for the proposed method on Charades-STA test split and ActivityNet Captions val_2 split.

Event reasoning	GF trans. layer	$\mathcal{L}_{\text{event}}$	Charades-STA		ANet Captions	
			R1@0.5	R1@0.7	R1@0.5	R1@0.7
			66.75	42.26	53.09	31.74
✓			66.91	42.67	54.44	33.87
✓		✓	67.24	43.85	55.09	35.21
✓	✓	✓	68.47	44.92	58.18	37.64



(a) $N \in [4:2:14]$ (b) $\lambda_{\text{event}} \in [0.5:0.5:4]$

Figure 1. Hyper-parameter analysis on Charades-STA test split.



(a) $N \in [4:2:14]$ (b) $\lambda_{\text{event}} \in [0.5:0.5:4]$

Figure 2. Hyper-parameter analysis on ActivityNet Captions val_2 split.

semantics in a given video. As shown in Tab. 1, we can obtain higher performance with the slot attention. In addition, the slot attention shows higher computational efficiency than the cross-attention in terms of GFLOPs and the number of parameters by reusing the parameters for every iteration.

Iteration K in slot attention. The number of iterations K in the slot attention determines how much each slot interacts with each other. To validate the effectiveness of the number of iterations K, we evaluate the performance, as shown in Tab. 2. The comparison between $K = 1, 2$ and 3 shows the larger number of K improves the performance with slightly lower computational efficiency (*i.e.*, GFLOPs). Meanwhile, larger values of K than 3 bring performance degradation. We speculate that a large number of iterations makes the model converges difficult, as analyzed in [5]. We set K to 3, which achieves a reasonable trade-off between training efficiency and performance.

Number of layers. We compare the performance according to the number of layers T in Tab. 3. Since a small number of layers (less than 3) insufficiently learn the video-sentence interaction, the result shows poor performance. While higher performance can be attained with more layers, the computational complexity also increases. Considering the overall performance and efficiency, we set T to 3.

2.2. Ablation study

We provide ablations on the key components of EaTR and hyper-parameters, including the number of moment queries N and the balancing parameter λ_{event}.

Component ablation. We study the impact of each component in EaTR on Charades-STA [1] and ActivityNet Captions [3] in Tab. 4. Each component introduces consistent improvement on both Charades-STA and ActivityNet Captions, where the full usage of components contributes 2.66% and 5.9% gain in terms of R1@0.7, respectively.

Number of moment queries. We depict the impact of the number of moment queries N on Charades-STA [1] and ActivityNet Captions [3] in Fig. 1a and Fig. 2a. For Charades-STA, a small N achieves better results than the large N where the optimal result is obtained with $N = 6$. In contrast, for ActivityNet Captions, the overall tendency is similar to the results of QVHighlights [4] where the optimal result is obtained with $N = 10$. The main difference between Charades-STA and the other two datasets lies in the granularity of videos: Charades-STA mostly contains fine-grained videos (*i.e.*, visually similar with subtle changes) consisting of few events whereas the other two datasets (*i.e.*, QVHighlights and ActivityNet Captions) contain coarse videos (*i.e.*, visually distinct with significant changes) consisting of numerous events. Due to the difference in the granularity of the video, a small number of N is enough for Charades-STA while a large number of N enables the model to better capture the numerous events in videos for QVHighlights and ActivityNet Captions. Thus, we set $N = 6$ for Charades-STA and $N = 10$ for ActivityNet Captions.

Effect of λ_{event}. The sensitivity of $\mathcal{L}_{\text{event}}$ on Charades-STA [1] and ActivityNet Captions [3] are in Fig. 1b and Fig. 2b. The event localization loss introduces an improvement with $2 \leq \lambda_{\text{event}} \leq 3$ for Charades-STA and with $1 \leq \lambda_{\text{event}} \leq 2.5$ for ActivityNet Captions. The values of λ_{event} smaller than 1 or larger than 3.5 degrades the performance which is a similar tendency across all three datasets.

Figure 3. Qualitative results of our EaTR on QVHighlights val split.

2.3. Qualitative results

We provide the qualitative results on QVHighlights [4] and Charades-STA [1] in Fig. 3 and Fig. 4 to validate the superiority of EaTR on the fine- and coarse-grained videos, respectively. We depict the cross-attention weight from the last decoder layer computed between the video-sentence representations and the moment queries that make the final prediction with the highest confidence score. Note that we only depicted the attention map corresponding to the video frames for clear analysis. As shown in Fig. 3, our EaTR correctly localizes the timestamp corresponding to the sentence regardless of the length of the target moment. In addition, we provide additional results for a single video labeled with two different sentences in Fig. 4. As shown in the figure, different moment queries are activated according to the given sentence and make the correct final prediction, demonstrating the effectiveness of the event-aware video grounding framework.

Failure cases. Since our EaTR generates the event-aware moment queries based on the visual contents of videos, the model is hard to provide informative referential search area when a video has visually similar frames. As shown in Fig. 5, our EaTR fails to localize the given sentence on the fine-grained videos composed of visually similar frames.

References

[1] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017.

[2] Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim. Uboco: Unsupervised boundary contrastive learning for generic event boundary detection. In *CVPR*, 2022.

[3] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

[4] Jie Lei, Tamara L Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *NeurIPS*, 2021.

[5] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 2020.

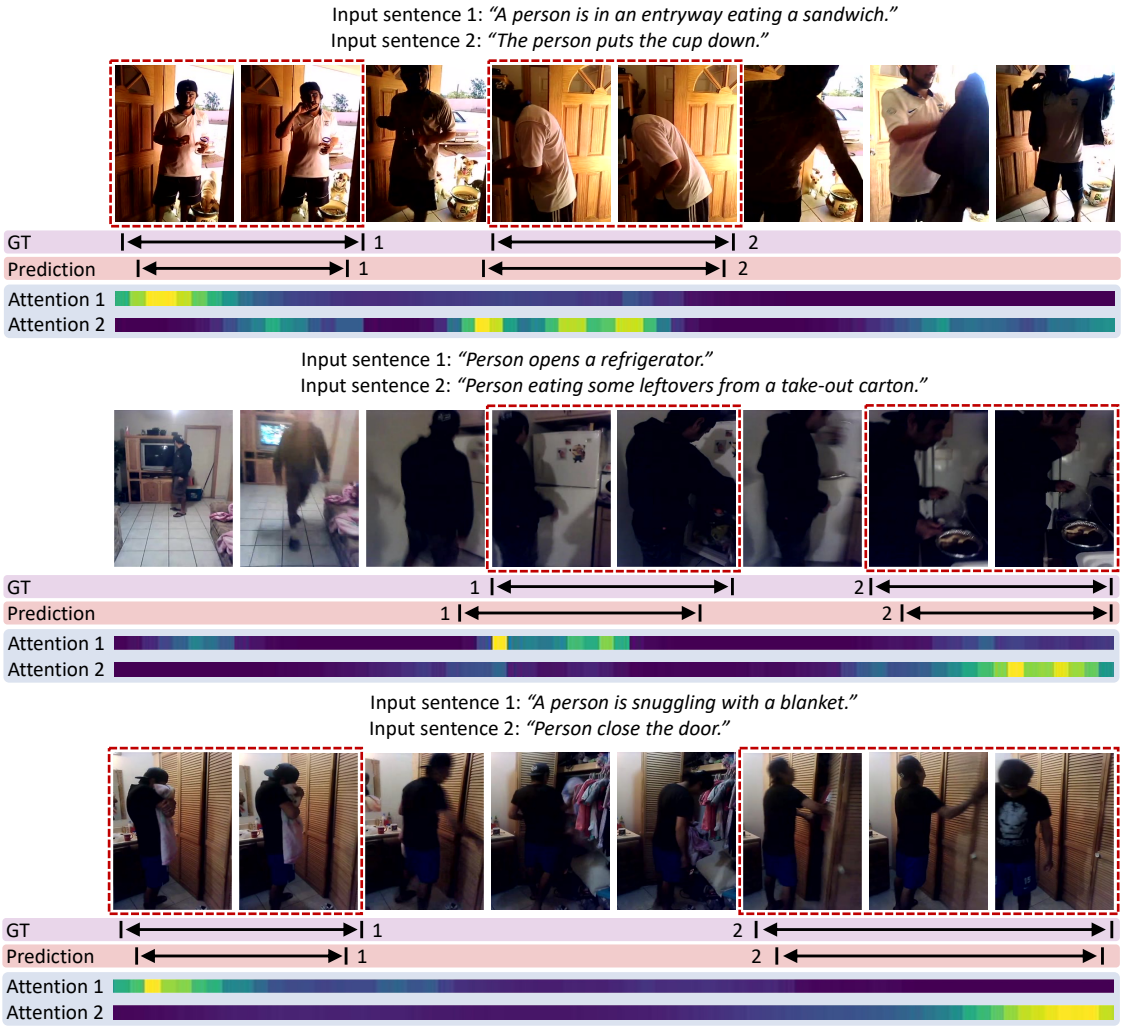


Figure 4. Qualitative results of our EaTR on Charades-STA test split.

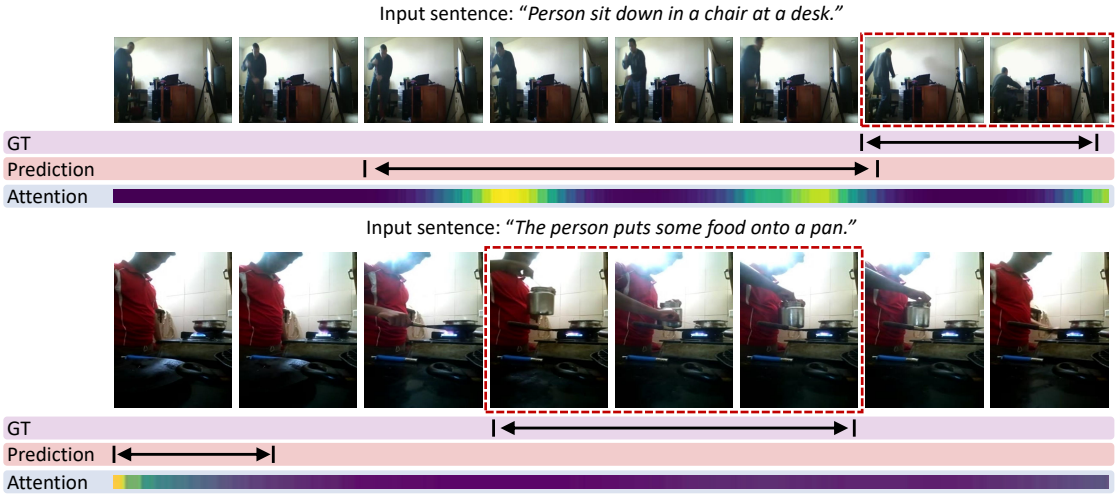


Figure 5. Failure cases of our EaTR on Charades-STA test split.