

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

Jinhyun Jang¹ Jungin Park¹ Jin Kim¹ Hyeongjun Kwon¹ Kwanghoon Sohn^{1,2,*}
¹Yonsei University ²Korea Institute of Science and Technology (KIST)
 {jr000192, newrun, kimjin928, kwonjunn01, khsohn}@yonsei.ac.kr

Abstract

Recent DETR-based video grounding models have made the model directly predict moment timestamps without any hand-crafted components, such as a pre-defined proposal or non-maximum suppression, by learning moment queries. However, their input-agnostic moment queries inevitably overlook an intrinsic temporal structure of a video, providing limited positional information. In this paper, we formulate an event-aware dynamic moment query to enable the model to take the input-specific content and positional information of the video into account. To this end, we present two levels of reasoning: 1) Event reasoning that captures distinctive event units constituting a given video using a slot attention mechanism; and 2) moment reasoning that fuses the moment queries with a given sentence through a gated fusion transformer layer and learns interactions between the moment queries and video-sentence representations to predict moment timestamps. Extensive experiments demonstrate the effectiveness and efficiency of the event-aware dynamic moment queries, outperforming state-of-the-art approaches on several video grounding benchmarks. The code is publicly available at <https://github.com/jinhyunj/EaTR>.

1. Introduction

Over the decade, online video platforms have been explosively developed, with the number of videos uploaded every day growing exponentially. Accordingly, the amount of work for video search (e.g. video summarization [52, 22], video retrieval [53, 71], text-to-video retrieval [6, 54]) has been explored to enable users to efficiently browse the information they want. While they have presented an efficient way to search videos by considering the whole content, providing a user-defined moment in a video is a different desire. As an alternative to this way, video grounding [1, 31, 41, 46] has been explored in recent years.

*Corresponding author

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF2021R1A2C2006703).

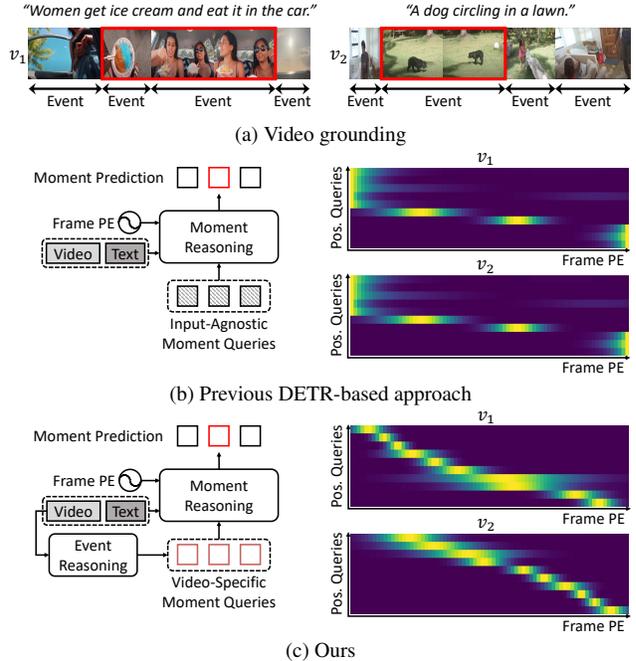


Figure 1. (a) Video grounding aims to localize timestamps of a moment referring to a given sentence. Each video is composed of its own set of event units with varying lengths. (b) Previous DETR-based methods learn input-agnostic moment queries, providing fixed referential search areas. (c) The proposed method learns event-aware moment queries, providing reliable referential search areas according to the given video.

Video grounding (also called natural language video moment localization) aims to localize timestamps of a moment referring to a given natural language sentence in a video, as shown in Fig. 1a. A key to identifying sentence-relevant moments is to 1) align video-language information; and 2) precisely reason temporal area. Most works have accomplished this by aligning sentence and heuristically pre-defined temporal proposals (e.g. sliding windows [1, 16, 37], temporal anchors [5, 77, 87]) or directly learning sentence-frame interactions [44, 7, 79]. However, they highly rely on the quality of hand-crafted components (e.g. proposals, non-maximum suppression) to achieve a promising result.

The recent success of detection transformer (DETR) [3]

has inspired approaches to integrate transformers [66] into video grounding framework [26, 2, 70, 14]. They learn a referential search area with a set of trainable embeddings, called *moment queries*, as an alternative to the heuristically designed proposals. Each moment query probes and aggregates video-sentence representations through the cross-attention mechanism where the final moment queries are used to predict the timestamps of the sentence-relevant moments. While they have achieved outperforming performance over the CNN-based approaches, the design choices of moment queries are still underexplored, exhibiting several drawbacks. Specifically, since moment queries are learned to contain general positional information, they produce a fixed input-agnostic search area during inference, as shown in Fig. 1b. However, a video is a complex visual stream that consists of multiple semantic units (*i.e.*, events) with varying lengths [21, 61, 59, 23]. In addition, the moment queries are controlled with equal contributions to aggregate video-sentence representations in the decoder layers, missing salient information and resulting in slow convergence.

In this paper, we propose a novel Event-aware Video Grounding TRansformer (EaTR) that formulates a video as a set of event units and treats video-specific event units as dynamic moment queries. Our EaTR performs two different levels of reasoning: 1) Event reasoning that identifies the event units comprising the video and produces the content and positional queries; and 2) Moment reasoning that fuses the moment queries with the given sentence, and interacts with the video-sentence representations to predict the final timestamps for video grounding. Specifically, the randomly initialized learnable event slots identify the distinctive event units from the given video using a slot attention mechanism [42]. The identified event units are then used as the moment queries in moment-level reasoning. While the moment queries in EaTR provide the input-specific referential search area as shown in Fig. 1c, the interaction between the video-sentence representations and the sentence-relevant moment queries should be properly captured to predict an accurate moment timestamps. To this end, we introduce a gated fusion transformer layer to effectively minimize the impact of the sentence-irrelevant moment queries and capture the most informative referential search area. In the gated fusion transformer layer, we fuse the moment queries and sentence representation according to their similarity in order to adaptively aggregate sentence information to the informative moment queries. The fused moment queries interact with the video-sentence representations in the transformer decoder to make the final decision for video grounding.

Extensive experiments on several video grounding benchmarks [26, 16, 24] demonstrate the effectiveness of the event-aware video grounding framework, achieving a new state-of-the-art performance over the previous methods [26, 41, 90]. In summary, our key contributions are as follows: (i) We

present a novel Event-aware Video Grounding Transformer (EaTR) that enhances the temporal reasoning capability of the moment queries by learning the video-specific event information. (ii) We introduce effective event reasoning and the gated fusion that highlight the distinctive events in a given video and sentence. (iii) We conduct extensive experiments to validate the effectiveness of the proposed method, and outperform state-of-the-art approaches on three video grounding benchmarks, including QVHighlights [26], Charades-STA [16], and ActivityNet Captions [24].

2. Related Work

Video grounding. A standard framework of localizing a moment corresponding to a given sentence can be categorized into two different paradigms. (i) Proposal-driven approaches first generate several candidate proposals and rank them based on their similarity with a sentence. Most works utilize pre-defined proposals such as sliding windows [1, 16, 37, 38, 18, 86] or temporal anchors [5, 77, 87, 81, 34]. Others proposed to generate high-quality proposals by exploring every possible pairs of start-end points [85, 32, 72] or with sentence guidance [60, 73, 8, 31]. (ii) Proposal-free approaches directly predict the target moment via learning video-sentence interactions. Several works attempt to solve the problem by formulating attention mechanisms [78, 19, 83, 58, 47], making dense predictions [44, 7, 79], combining complementary visual features (*e.g.* object regions, motion features) [80, 9, 36], and reducing the dataset bias [50, 84, 75, 49, 30, 20]. Although the two paradigms have achieved impressive results, they are limited in their use of hand-crafted components (*e.g.* pre-defined proposals, non-maximum suppression) and redundancy (*e.g.* large number of candidate proposals).

To simplify the whole process into an end-to-end manner, recent works [26, 2, 70, 41, 14, 74, 46] adopted DETR-based architecture into the video grounding task. Despite the progress, the ineffective use of learnable moment queries limits the model capability on temporal reasoning. UMT [41] attempted to improve the query design by conditioning them on extra modality (*e.g.* audio, optical flow). Instead of relying on additional input, we propose to utilize a video itself as an interpretable positional guidance.

DETR and its variants. The adoption of transformers [66] to object detection (DETR) [3] has streamlined the whole pipeline by removing the need of hand-crafted components while improving the performance. Despite its success, DETR has its own issue of slow training convergence. Several studies [91, 63, 76, 17, 45, 10, 68, 82, 39, 28] attribute the issue to the naive design of object query and its operation on cross-attention module; object queries require a long training time to accurately learn where and what to focus at the cross-attention module. The seminal work discovered

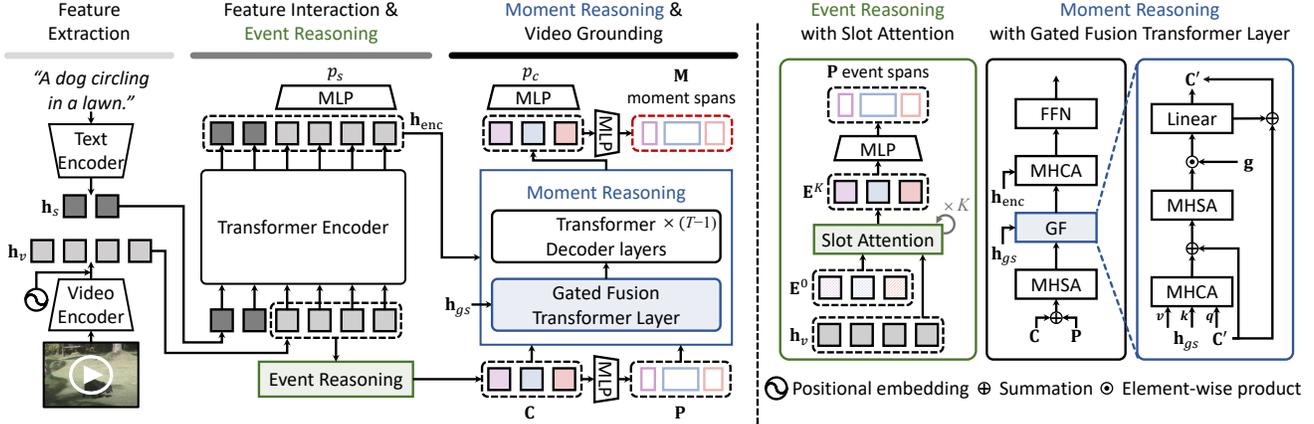


Figure 2. An overview of the proposed EaTR. The procedure consists of three steps: 1) feature extraction, 2) feature interaction and event reasoning, and 3) moment reasoning for video grounding. The event reasoning is done through a slot attention mechanism with learnable event slots (E^0). The outputs of the event reasoning network (E^K, P) serve as an initial moment queries (C, P) of the moment reasoning network. The initial moment queries are fused with the sentence feature (h_{gs}) through a gated fusion transformer layer in order to better focus on the sentence-relevant queries. The queries then interact with the video-sentence representations (h_{enc}) for making the final prediction on the moment timestamps (M).

the importance of spatial priors for convergence speed and performance by reformulating object queries as 2D center coordinates (*i.e.*, x, y) [45, 68] or 4D box coordinates (*i.e.*, x, y, w, h) [91, 17, 39, 28]. Others imposed spatial-constraint at the cross-attention module such as sparse sampling [91], Gaussian prior [17], or conditional weight [45, 39].

Generic event boundary detection. Generic Event Boundary Detection (GEBD) [61] is a recently introduced video understanding task that aims to identify every instant that human perceive as event boundaries. The event boundary includes a change of subject, action, and environment. Recent works [23, 64] explored frame-wise similarity, namely Temporal Self-similarity Matrix (TSM), to better perceive temporal variations. Specifically, UBoCo [23] focused on producing boundary-sensitive features (*i.e.*, distinctive TSM) in an unsupervised manner with a recursive TSM parsing mechanism. Having a common goal of identifying the events without supervision, we adopt their contrastive kernel for generating pseudo event information. However, instead of directly adopting the whole process, we newly design the event reasoning network suitable for the DETR-based video grounding network.

Slot attention. Slot attention [42] is a recently proposed iterative attention mechanism that aims to learn the object-centric representation. The randomly initialized *slots* are introduced to interact with the input features and group the pixels belonging to the same object. Our event reasoning network employs the grouping property of the slot attention mechanism to aggregate visually similar frames into multiple events and train the network with the event localization loss.

3. Proposed Method

3.1. Background and motivation

Video grounding aims to localize the timestamps of moments referring to a given sentence \mathcal{S} in an untrimmed video \mathcal{V} . Recent DETR-based methods [26, 2, 74, 46] formulate learnable query embeddings Q (*i.e.*, moment query) that represent a set of learnable referential search areas, and predict the target moments using transformer [66] in an end-to-end manner. The moment query can be decomposed into two parts according to their physical roles: a content query C and a positional query P . Each query is responsible for aggregating video-sentence representations based on a semantic similarity (with the content query) and a positional similarity (with the positional query). The previous works define the initial content and positional queries as zero embeddings and learnable embeddings respectively, and progressively aggregate video-sentence information by going through the transformer decoder. While they have successfully demonstrated the effectiveness of the DETR-based architecture, the design of the input-agnostic moment query makes the search area ambiguous and training difficult.

To address this problem, we propose an event-aware video grounding transformer (EaTR) where a video is treated as a set of event units. In our framework, we formulate the dynamic moment queries to provide precise referential search area by identifying the event units from the given video, and fuse the dynamic moment queries with sentence information. In the following section, we describe feature extraction (Sec. 3.2), event reasoning (Sec. 3.3), and moment reasoning (Sec. 3.4). The overall architecture is shown in Fig. 2.

3.2. Feature extraction and interaction

Given a video \mathcal{V} of length L_v and a sentence \mathcal{S} of length L_s , we first encode the video and sentence representations using corresponding pretrained backbone networks (e.g. I3D [4], CLIP [56]) as follows:

$$\mathbf{h}_v = f_v(\mathcal{V}) + PE \in \mathbb{R}^{L_v \times d}, \quad \mathbf{h}_s = f_s(\mathcal{S}) \in \mathbb{R}^{L_s \times d}, \quad (1)$$

where \mathbf{h}_v and \mathbf{h}_s are the video and sentence representations, PE is a set of positional embeddings for video frames, $f_v(\cdot)$ and $f_s(\cdot)$ are the pretrained backbone networks for each modality, respectively. Similar to Moment-DETR [26], the video and sentence representations interact with each other through a stack of T transformer encoder layers to obtain the video-sentence representations \mathbf{h}_{enc} :

$$\mathbf{h}_{\text{enc}} = f_{\text{enc}}(\mathbf{h}_v || \mathbf{h}_s), \quad (2)$$

where $f_{\text{enc}}(\cdot)$ is the transformer encoder and $||$ is the concatenate operation. Following [26], we feed the partial representations corresponding to the video in \mathbf{h}_{enc} into a linear layer to predict the saliency score $p_s \in \mathbb{R}^{L_v}$ that represents the similarity between the video frames (or clip) and words in the sentence. To enlarge the saliency score gap between the sentence-relevant frames and other frames, we employ the saliency loss [26]:

$$\mathcal{L}_{\text{sal}} = \max(0, \alpha + \bar{p}_{s,\text{out}} - \bar{p}_{s,\text{in}}), \quad (3)$$

where $\bar{p}_{s,\text{in}}$ and $\bar{p}_{s,\text{out}}$ are the average saliency scores of randomly sampled frames within and outside of the ground truth time interval, respectively, and α is a margin.

3.3. Event reasoning

The main problem of the previous moment queries is that they provide an ambiguous referential search area due to the design of the input-agnostic moment queries. To handle the problem, we propose to identify the distinctive event units from the given video and utilize event information for initializing the dynamic moment queries.

Specifically, we derive N event units from the video representation \mathbf{h}_v using a set of N learnable event slots $\mathbf{E} \in \mathbb{R}^{N \times d}$ and the slot attention mechanism [42]. Formally, the event slots iteratively interact with the video representations \mathbf{h}_v for K iterations to group the visually similar frames and obtain the final event units \mathbf{E}^K . In k -th iteration, we first embed \mathbf{h}_v and \mathbf{E}^k using layer normalization followed by linear projections, such that:

$$\mathbf{h}'_v = (\text{LN}(\mathbf{h}_v))\mathbf{W}_1, \quad \mathbf{E}'^{k-1} = (\text{LN}(\mathbf{E}^{k-1}))\mathbf{W}_2, \quad (4)$$

where \mathbf{h}'_v and \mathbf{E}'^{k-1} are the embedded video representations and event slots, $\text{LN}(\cdot)$ is layer normalization, \mathbf{W}_1 , and \mathbf{W}_2 are the linear projection matrices. The k -th interaction matrix

between \mathbf{h}'_v and \mathbf{E}'^{k-1} can be computed as:

$$\mathbf{A}^k = \text{Softmax} \left(\frac{(\mathbf{h}'_v)(\mathbf{E}'^{k-1})^\top}{\sqrt{d}} \right) \in \mathbb{R}^{L_v \times N}, \quad (5)$$

where $\text{Softmax}(\cdot)$ is the softmax function along the event slot direction. With the interaction matrix \mathbf{A}^k , the k -th event slots are updated by following equation:

$$\mathbf{U} = (\hat{\mathbf{A}}^k)^\top (\mathbf{h}_v) \mathbf{W}_3 + \mathbf{E}^{k-1}, \quad \text{where } \hat{\mathbf{A}}^k_{l,n} = \frac{\mathbf{A}^k_{l,n}}{\sum_{L_v} \mathbf{A}^k_{l,n}},$$

$$\mathbf{E}^k = (\text{LN}(\mathbf{U}))\mathbf{W}_4 + \mathbf{U}, \quad (6)$$

with additional linear projection matrices \mathbf{W}_3 and \mathbf{W}_4 . Different from the conventional slot attention [42], we replace the GRU layer with residual summation to avoid inefficient computations. Since each event slot $\mathbf{e}_n^K \in \mathbf{E}^K$ contains visual information corresponding to the individual event of the video, we use \mathbf{E}^K as the initial content queries \mathbf{C} . In addition, we project \mathbf{E}^K to the 2-dimensional embedding space to derive the initial positional queries \mathbf{P} that represent the center and duration of the referential search area for each moment query as:

$$\mathbf{P} = \mathbf{E}^K \mathbf{W}_p = \{(c_n, w_n)\}_{n=1}^N, \quad (7)$$

where c_n and w_n are the center and width of the referential time span for the n -th content query. While cross-attention can be used as an alternative to slot attention, slot attention achieves superior performance with higher efficiency by enforcing the slots to compete each other and reusing linear projection matrices (e.g. \mathbf{W}_1) for every iteration.

To guarantee the moment queries contain the event units, we learn event reasoning by generating the pseudo event timestamps of the video based on the temporal self-similarity matrix [51, 12, 48, 23]. Specifically, we employ the recent contrastive kernel [23] that computes the event boundary scores by convolving with the diagonal elements of TSM. By thresholding and sampling the boundary scores, we are able to obtain the timestamps of the event units $\hat{\mathbf{P}}$, where each element $\hat{\mathbf{P}}_i \in [0, 1]^2$ defines the normalized center coordinate and duration of an event. With the pseudo event timestamps, we formulate an event localization loss between positional queries and each corresponding pseudo event timestamp. Since the order of the predicted event sets is arbitrary, we find an optimal assignment between the positional queries and pseudo event spans via the Hungarian matching algorithm [25, 3]. The optimal assignment $\hat{\sigma}$ is determined based on the similarity of the pseudo event spans $\hat{\mathbf{P}}$ and predicted event spans \mathbf{P} as:

$$\hat{\sigma} = \arg \min_{\sigma \in \mathfrak{S}_N} \sum_i^N \mathcal{C}(\hat{\mathbf{P}}_i, \mathbf{P}_{\sigma(i)}), \quad (8)$$

$$\mathcal{C}(\hat{\mathbf{P}}_i, \mathbf{P}_{\sigma(i)}) = \lambda_{l_1} \|\hat{\mathbf{P}}_i - \mathbf{P}_{\sigma(i)}\|_1 + \lambda_{\text{iou}} \mathcal{L}_{\text{iou}}(\hat{\mathbf{P}}_i, \mathbf{P}_{\sigma(i)}),$$

where \mathcal{L}_{iou} is generalized temporal IoU [57]. λ_{l_1} and λ_{iou} are the balancing parameters. Given the optimal assignment, the event localization loss is defined as:

$$\mathcal{L}_{\text{event}} = \sum_i^N \mathcal{C}(\hat{\mathbf{P}}_i, \mathbf{P}_{\hat{\sigma}(i)}). \quad (9)$$

3.4. Moment reasoning

In moment reasoning, we aggregate the video-sentence representations to the moment queries through a stack of T transformer decoder layers to predict the final moment timestamps. Although the initial dynamic moment queries contain the video-specific referential search areas, the sentence-relevant moment queries should be enhanced while filtering out irrelevant queries to produce more reliable search areas. To this end, we propose a gated fusion (GF) transformer layer that enhances the sentence-relevant moment queries and suppresses the other queries by fusing the moment queries with a global sentence representation. Our GF transformer layer is heavily inspired by [69, 35], but transformed into a form suitable to the transformer architecture.

Enhanced moment query. Each positional query $\mathbf{p}_n = (c_n, w_n)$ is first extended to d -dimensional space through the sinusoidal positional encoding (PE) [45, 39], concatenation, and MLP layer as:

$$\mathbf{p}_n \leftarrow \text{MLP}(\text{Concat}(\text{PE}(c_n), \text{PE}(w_n))). \quad (10)$$

By expanding the positional queries to d -dimensional space, the multi-head self-attention (MHSA) and cross-attention (MHCA) layers can take the content and positional queries as inputs at the same level. Before fusing the moment queries and global sentence representation, we feed the sum of the content and positional queries into the MHSA layer, such that the enhanced moment query \mathbf{C}' can be obtained by:

$$\mathbf{C}' = \text{MHSA}(\mathbf{C} \oplus \mathbf{P}) \in \mathbb{R}^{N \times d}, \quad (11)$$

where \oplus is an element-wise summation.

Gated fusion (GF) transformer layer. The GF transformer layer takes \mathbf{C}' and the global sentence representation \mathbf{h}_{gs} which is obtained by applying max pooling on \mathbf{h}_s . We treat \mathbf{C}' as the query and \mathbf{h}_{gs} as the key and value of the MHCA layer to aggregate the query-relevant sentence information, such that the aggregated sentence representations $\hat{\mathbf{C}}$ is derived as:

$$\hat{\mathbf{C}} = \text{MHCA}(\mathbf{C}', \mathbf{h}_{gs}, \mathbf{h}_{gs}) \in \mathbb{R}^{N \times d}. \quad (12)$$

The similarity between the n -th moment query \mathbf{c}'_n and aggregated sentence representation $\hat{\mathbf{c}}_n$ is then used as a gate to suppress the irrelevant queries. In other words, the gate for the sentence-relevant query has a high value, otherwise

represents a low value. The gate for the n -th moment query is computed as a single scalar by:

$$g_n = \text{Sigmoid}(\mathbf{c}'_n \cdot \hat{\mathbf{c}}_n^\top). \quad (13)$$

The gated fusion is then formulated with \mathbf{C}' , $\hat{\mathbf{C}}$, and the gate \mathbf{g} by the following equation:

$$\mathbf{C}' \leftarrow \text{Linear}(\mathbf{g} \odot \text{MHSA}(\mathbf{C}' \oplus \hat{\mathbf{C}})) + \mathbf{C}', \quad (14)$$

where Linear is a single linear layer and \odot is the element-wise multiplication. The enhanced moment queries then interact with the video-sentence representations \mathbf{h}_{enc} through the modulated MHCA [39, 46] and are fed to a feed-forward network (FFN).

Moment prediction. After the GF transformer layer, the moment queries go through the remaining $(T - 1)$ transformer decoder layers where each positional query is updated layer-wise with its offset predicted from the corresponding content query [91, 39]. The output of the transformer decoder \mathbf{h}_{dec} is fed to FFN for predicting the moment span \mathbf{M} . We also utilize a linear layer to predict the confidence score $p_c \in \mathbb{R}^N$ corresponding to each moment query. To learn the moment localization, set prediction loss based on bipartite matching is applied [3]. Given the ground truth moment timestamps $\hat{\mathbf{M}}_i \in [0, 1]^2$ consisting of the normalized center coordinate and width, the optimal assignment is determined based on the timestamp similarities and the corresponding confidence scores using the Hungarian algorithm as:

$$\begin{aligned} \hat{\sigma}' &= \arg \min_{\sigma' \in \mathfrak{S}_N} \sum_i^N \left[-\lambda_c p_{c, \sigma'(i)} + \mathcal{C}(\hat{\mathbf{M}}_i, \mathbf{M}_{\sigma'(i)}) \right], \\ \mathcal{C}(\hat{\mathbf{M}}_i, \mathbf{M}_{\sigma'(i)}) &= \lambda_{l_1} \|\hat{\mathbf{M}}_i - \mathbf{M}_{\sigma'(i)}\|_1 \\ &\quad + \lambda_{\text{iou}} \mathcal{L}_{\text{iou}}(\hat{\mathbf{M}}_i, \mathbf{M}_{\sigma'(i)}), \end{aligned} \quad (15)$$

where λ_{l_1} , λ_{iou} and λ_c are the balancing parameters. With the optimal assignment $\hat{\sigma}'$, the moment localization loss [2, 26] is defined as:

$$\mathcal{L}_{\text{moment}} = \sum_i^N \left[-\lambda_c \log p_{c, \hat{\sigma}'(i)} + \mathcal{C}(\hat{\mathbf{M}}_i, \mathbf{M}_{\hat{\sigma}'(i)}) \right]. \quad (16)$$

Overall objectives. The overall objective is defined as:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{moment}} + \lambda_{\text{sal}} \mathcal{L}_{\text{sal}} + \lambda_{\text{event}} \mathcal{L}_{\text{event}}, \quad (17)$$

where λ_{sal} and λ_{event} are the balancing parameters.

4. Experiments

4.1. Datasets and evaluation protocols

We evaluate the proposed method on three standard video grounding benchmarks, including the QVHighlights [26], Charades-STA [16], and ActivityNet Captions [24] datasets.

Table 1. Experimental results on QVHighlights val split. HD represents highlight detection. We repeat the experiment with 5 different seeds and report the mean performance and standard deviation. † indicates the model with additional audio input.

Methods	Video Grounding					HD		GFLOPs	Params
	R1		mAP			≥ Very Good			
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1		
BeautyThumb [62]	-	-	-	-	-	14.36	20.88	-	-
DVSE [40]	-	-	-	-	-	18.75	21.79	-	-
MCN [1]	11.41	2.72	24.94	8.22	10.67	-	-	-	-
CAL [13]	25.49	11.54	23.40	7.65	9.89	-	-	-	-
CLIP [56]	16.88	5.19	18.11	7.00	7.67	31.30	61.04	-	-
XML [27]	41.83	30.35	44.63	31.73	32.14	34.49	55.25	-	-
XML+ [27]	46.69	33.46	47.89	34.67	34.90	35.38	55.06	-	-
Moment-DETR [26]	52.89 \pm 2.3	33.02 \pm 1.7	54.82 \pm 1.7	29.40 \pm 1.7	30.73 \pm 1.4	35.69 \pm 0.5	55.60 \pm 1.6	0.28	4.8M
UMT† [41]	56.23	41.18	53.83	37.01	36.12	38.18	59.99	0.63	14.9M
MH-DETR [74]	60.05	42.48	60.75	38.13	38.38	38.22	60.51	0.34	8.2M
QD-DETR [46]	62.40 \pm 1.1	44.98 \pm 0.8	62.52 \pm 0.6	39.88 \pm 0.7	39.86 \pm 0.6	38.94 \pm 0.4	62.40 \pm 1.4	0.60	7.6M
Ours	61.36 \pm 1.2	45.79 \pm 0.7	61.86 \pm 0.6	41.91 \pm 0.6	41.74 \pm 0.7	37.15 \pm 0.5	58.65 \pm 1.4	0.47	9.0M

Table 2. Experimental results on Charades-STA test split with I3D features and ActivityNet Captions val_2 split with C3D features.

Methods	Charades-STA		ActivityNet Captions	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7
BPNNet [72]	50.75	31.64	42.07	24.69
DRN [79]	53.09	31.75	45.45	24.36
FIAN [55]	58.55	37.72	47.90	29.81
LGI [47]	59.46	35.48	41.51	23.07
DeNet [89]	59.70	38.52	43.79	-
CPN [88]	59.77	36.67	45.10	28.10
CSMGAN [34]	60.04	37.34	49.11	29.15
SSCS [11]	60.75	36.19	46.67	27.56
CBLN [32]	61.13	38.22	48.12	27.60
IA-Net [35]	61.29	37.91	48.57	27.95
APGN [31]	62.58	38.86	48.92	28.64
MGSL-Net [30]	63.98	41.03	51.87	31.42
SMIN [67]	64.06	40.75	48.46	30.34
SLP [29]	64.35	40.43	52.89	32.04
D-TSG [33]	65.05	42.77	54.29	33.64
SSRN [90]	65.59	42.65	54.49	33.15
Ours	68.47	44.92	58.18	37.64

- **QVHighlights** is the recently proposed dataset that supports both video grounding and highlight detection that select representative clips in a given video. The dataset contains 10,148 videos with 18,367 moments and 10,310 sentences. The dataset also provides annotations of 5-scale saliency scores (from very bad to very good) within the annotated moment for highlight detection. Since the official test splits do not have the ground truth, we test on the official validation set.
- **Charades-STA** includes 16,128 moment-sentence pairs, where the average duration of the full video and annotated moment are 30 and 8.1 seconds long, respectively. The official splits provide 12,408 and 3,720 pairs

for train and test split.

- **ActivityNet Captions** contains 15K videos with 72K sentences, where the average duration of the full video and annotated moment are 117.6 and 36.2 seconds, respectively. The train, validation_1, and validation_2 splits include 37,417, 17,505, and 17,031 moment-sentence pairs. Following the previous works [79, 67], we utilize val_1 for the validation and val_2 for testing.

Evaluation metrics. We adopt Recall1@IoU m following the previous works [47, 26, 9]; The percentage of top-1 predicted moment having IoU larger than threshold m with the ground truth moment. We report results with $m = \{0.5, 0.7\}$. For QVHighlights, we report mean average precision (mAP) with IoU threshold 0.5, 0.75, and the average mAP over IoU thresholds [0.5: 0.05: 0.95]. Although highlight detection is not the task of our interest, we report the results using mAP and HIT@1 for QVHighlights.

4.2. Implementation details

Feature representations. For the QVHighlights dataset, we leverage the pre-trained SlowFast [15] and CLIP [56] features to extract the video features, following [26, 41, 74, 46] for fair comparisons. The features are pre-extracted every 2 seconds. For the Charades-STA and ActivityNet Captions datasets, we extract the video features from the most commonly used pre-trained I3D [4] and C3D [65] backbones, respectively. Each feature vector captures 16 consecutive frames with 50% overlap. We uniformly sample 200 feature vectors from each video for ActivityNet Captions dataset. For the sentence features, we leverage the token-level CLIP text features.

Training settings. We set the number of layers in the transformer encoder and decoder as $T = 3$. Following [26], the

Table 3. Component ablation results for the proposed method on QVHighlights val split.

Event reasoning	GF trans. layer	$\mathcal{L}_{\text{event}}$	R1@0.5	R1@0.7	mAP
			55.10 \pm 1.4	40.03 \pm 1.0	35.02 \pm 0.9
✓			57.71 \pm 1.4	42.32 \pm 0.9	37.42 \pm 0.7
✓		✓	59.10 \pm 1.3	43.32 \pm 0.9	38.98 \pm 0.6
✓	✓	✓	61.36 \pm 1.2	45.79 \pm 0.7	41.74 \pm 0.7

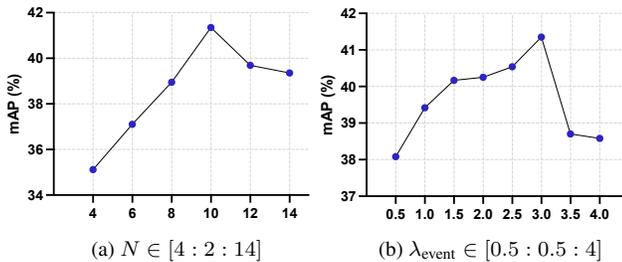


Figure 3. Hyper-parameter analysis on QVHighlights val split.

balancing parameters for the total loss function are set to $\lambda_{l_1} = 10$, $\lambda_{\text{iou}} = 1$, $\lambda_c = 4$, and $\alpha = 0.2$. λ_{sal} is set as 1 for QVHighlights, 4 for Charades-STA, and ActivityNet Captions, respectively. For all the models, we set the hidden dimensions to 256 and the number of attention heads to 8. We train all the models with batch size 32 for 200 epochs using AdamW [43] with weight decay $1e-4$. The initial learning rate is set to $1e-4$ for QVHighlights, and $2e-4$ for Charades-STA and ActivityNet Captions, respectively. All the experiments are implemented with Pytorch v1.12.1 with a single NVIDIA RTX A6000 GPU.

4.3. Comparison with state-of-the-art

In this section, we present the performance comparisons with the state-of-the-art methods [72, 79, 55, 47, 89, 88, 34, 11, 32, 35, 31, 30, 67, 29, 33, 90, 62, 40, 1, 13, 56, 27, 26, 41] to demonstrate the effectiveness of the proposed method. Note that we compare the models with the same feature representations for a fair comparison.

Results on QVHighlights. We provide the performance comparisons between the proposed EaTR and the concurrent DETR-based approaches [26, 41, 74, 46] on QVHighlights [26]. As shown in Tab. 1, our EaTR establishes new state-of-the-art performances on several metrics for video grounding, demonstrating the effectiveness of the event-aware dynamic moment queries. For highlight detection, our EaTR outperforms Moment-DETR by 1.46% and 3.05% in terms of mAP and HIT@1 while showing lower performance than the other approaches [41, 74, 46]. We speculate that the ability to detect highlights mainly depends on learning cross-modal interactions rather than improving the temporal reasoning ability.

Table 4. Component ablation results for the fusion method in the GF transformer layer on QVHighlights val split.

Method	R1@0.5	R1@0.7	mAP
Add	58.68 \pm 1.4	43.77 \pm 0.9	38.07 \pm 0.7
Concat	58.06 \pm 1.1	43.10 \pm 0.7	38.86 \pm 0.7
MHCA	59.74 \pm 1.2	44.71 \pm 0.6	39.69 \pm 0.6
GF	61.36 \pm 1.2	45.79 \pm 0.7	41.74 \pm 0.7

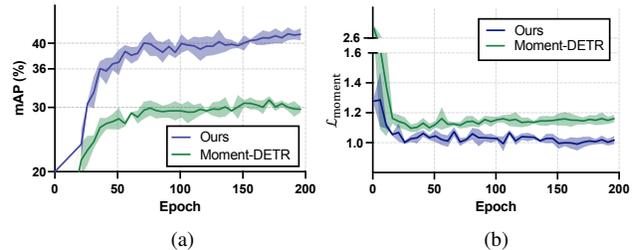


Figure 4. Comparison of training convergence of Moment-DETR [26] and Ours on QVhighlight val split. We plot the (a) mAP (%) curves and (b) $\mathcal{L}_{\text{moment}}$ curves. We repeat the experiment with 5 different seeds and present the mean performance and range.

We additionally compare GFLOPs and the number of model parameters to evaluate the computational efficiency. While our EaTR requires more parameters than [26, 74, 46], we attain an outstanding performance with fewer computations (*i.e.*, GFLOPs) compared to [41, 46].

Results on Charades-STA and ActivityNet Captions.

We report the results evaluated on Charades-STA [16] and ActivityNet Captions [24] in Tab. 2. The main difference between the previous works [31, 79, 30] and our EaTR is the use of hand-crafted components, such as temporal proposals or post-processing steps. Typically, the proposal-driven approaches [32, 31, 34] utilize an excessive number of candidate proposals generated with heuristics, while the proposal-free approaches [79, 55, 30] make dense frame-wise predictions to achieve promising results. These methods require pre-processing steps for defining the candidate proposals or post-processing steps (*e.g.* non-maximum suppression) for reducing the redundant predictions. Contrary to this, our EaTR utilizes only a set of moment queries (typically less than 10 queries) to predict accurate moments without requiring any hand-crafted components. Despite this training efficiency, our EaTR outperforms state-of-the-art methods, achieving 2.88% and 3.69% performance improvements in terms of R1@0.5 on each dataset.

4.4. Ablation study and discussion

To investigate the impact corresponding to key components of the proposed method, we conduct ablation studies on the validation set of QVHighlights [26]. In addition, we provide visualization examples to discuss how the dynamic moment queries of the proposed method work. The ablation

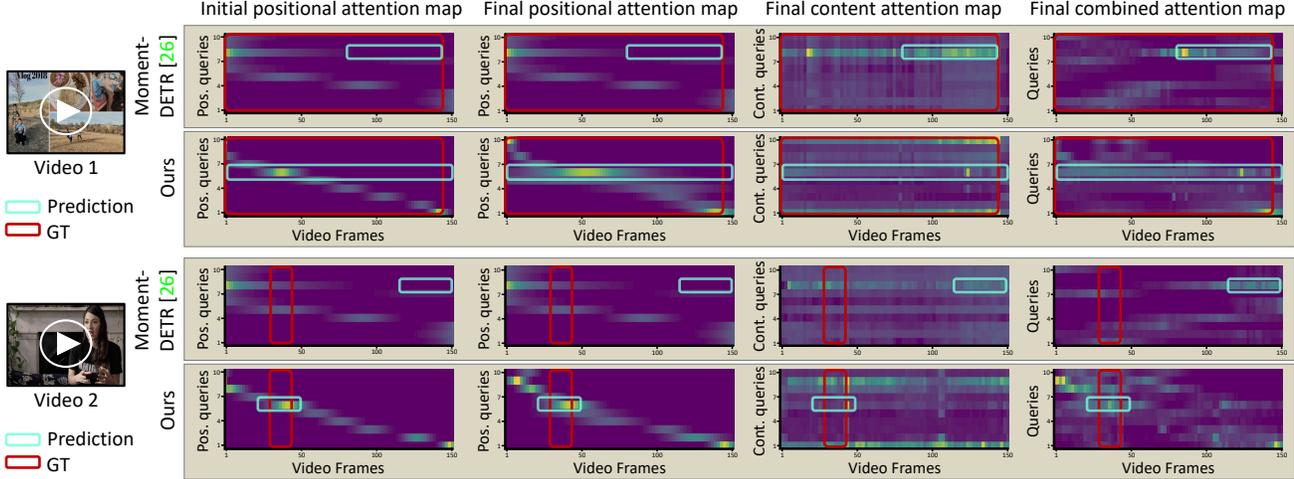


Figure 5. Visualization of the positional attention weights, the content attention weights, and the combined attention weights of the cross-attention module of Moment-DETR [26] and Ours. Each attention map is scaled by the corresponding confidence score of each query and sorted in order for clear analysis.

studies for the other two datasets and additional qualitative results are provided in the supplementary material.

Component ablation. We first investigate the effectiveness of each component in our EaTR. As shown in Tab. 3, we report the impact according to event reasoning, the gated fusion transformer layer, and the event localization loss $\mathcal{L}_{\text{event}}$. Sequentially applying the event reasoning and event localization loss contributes 2.61% and 4.0% to performance improvement, and using all components with gated fusion layer improves performance by 6.26% in terms of R1@0.5.

Number of moment queries. We depict the performance in terms of mAP according to the number of moment queries N in Fig. 3a. Since the number of moment queries determines the granularity of the referential search area, the performance is gradually improved as N increases. Meanwhile, a large number of N makes it difficult for the model to capture long events, resulting in performance degradation. We set N to 10 which represents the best performance.

Effect of λ_{event} . The balancing parameter λ_{event} controls the impact of a newly introduced event loss $\mathcal{L}_{\text{event}}$ in the total training objective. To study the sensitivity of $\mathcal{L}_{\text{event}}$, we report the performance according to λ_{event} in Fig. 3b. The result shows the performance monotonically increases with $1 \leq \lambda_{\text{event}} \leq 3$. The values of λ_{event} smaller than 1 or larger than 3 derive the poor performance, validating the impact of the proposed event localization loss.

Fusion method in the GF layer. We conduct an additional ablation study for the fusion method in the GF layer, including the addition, concatenation, and multi-head cross-attention (MHCA) fusions. As shown in Tab. 4, we observe

poor performance with the first two fusion methods. While the main goal of the fusion layer is to enhance the sentence-relevant moment queries and suppress the irrelevant ones, the addition and concatenation fusions fail to emphasize the informative moment queries, making the interaction with the video-sentence representations less effective. To verify the effectiveness of the gated fusion, we present the performance with the MHCA fusion, which skips Eq. (13) and (14). Although the MHCA provides a slight improvement over simple fusions (*i.e.*, addition and concatenation), it is still insufficient to capture distinctive queries, leading to lower performance than the GF layer.

Convergence analysis. We argued that providing reliable referential search areas accomplish a high training efficiency. To validate this, we compare the training convergence of our EaTR and Moment-DETR [26] which uses input-agnostic moment queries. As shown in Fig. 4, our EaTR converges much faster and achieves higher performance than Moment-DETR, demonstrating the importance of the precise referential search area when training video grounding models.

Attention visualization. We validate the impact of our dynamic moment queries on the final prediction. As shown in Fig. 5, we visualize the attention between (first column) the initial positional queries and frame positional embeddings, (second column) the final positional queries and frame positional embeddings, (third column) the final content queries and video-sentence representations, and (last column) the whole moment queries and video-sentence representations with frame positional embeddings. To compare the dynamic moment query of our EaTR and the input-agnostic moment query of Moment-DETR, we depict attention maps for two different videos. In each attention map, the horizontal and

vertical axes represent the frame and query indices, respectively. The initial positional queries of Moment-DETR provide a fixed input-agnostic search area regardless of the input video. Therefore, the model heavily relies on the quality of the video-sentence interactions, ignoring the highlighted positional attention and leading to the wrong prediction. Meanwhile, our EaTR provides different search areas according to the video, making correct predictions with a balanced contribution of the content and positional queries.

5. Conclusion and Future Work

In this paper, we have introduced a novel Event-aware Video Grounding Transformer, termed EaTR, that performs event and moment reasoning for video grounding. In event reasoning, we identify the event units comprising a given video with the slot attention mechanism. The event units are treated as the initial dynamic moment queries that provide the video-specific referential search areas. In moment reasoning, we introduce the gated fusion transformer layer to enhance the sentence-relevant moment queries and filter out the irrelevant queries, producing more reliable referential search areas. The dynamic moment queries interact with the video-sentence representations through the transformer decoder layers, enabling more accurate video grounding over state-of-the-art methods. Extensive experiments demonstrated the effectiveness and efficiency of the event-aware dynamic moment queries.

While we have explored the dynamic moment query based on visual information, consideration of sentence information is still underexplored. We hope our study will promote the potential of research and provide a foundation for variants of the moment query.

Acknowledgement. This research was supported by the Yonsei Signature Research Cluster Program of 2022 (2022-22-0002) and the KIST Institutional Program (Project No.2E31051-21-203).

References

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1, 2, 6, 7
- [2] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. In *EMNLP*, 2021. 2, 3, 5
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2, 4, 5
- [4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 4, 6
- [5] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018. 1, 2
- [6] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, 2021. 1
- [7] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, 2020. 1, 2
- [8] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, 2019. 2
- [9] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. *NeurIPS*, 2021. 2, 6
- [10] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *ICCV*, 2021. 2
- [11] Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. Support-set based cross-supervision for video grounding. In *ICCV*, 2021. 6, 7
- [12] Debidatta Dwivedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In *CVPR*, 2020. 4
- [13] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*, 2019. 6, 7
- [14] Xiang Fang, Daizong Liu, Pan Zhou, Zichuan Xu, and Ruixuan Li. Hierarchical local-global transformer for temporal sentence grounding. *arXiv preprint arXiv:2208.14882*, 2022. 2
- [15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 6
- [16] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 1, 2, 5, 7, 12, 13
- [17] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. In *ICCV*, 2021. 2, 3
- [18] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *WACV*, 2019. 2
- [19] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *NAACL*, 2019. 2
- [20] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *ECCV*, 2022. 2
- [21] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *CVPR*, 2019. 2

- [22] Hao Jiang and Yadong Mu. Joint video summarization and moment localization by cross-task sample transfer. In *CVPR*, 2022. 1
- [23] Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim. Uboco: Unsupervised boundary contrastive learning for generic event boundary detection. In *CVPR*, 2022. 2, 3, 4, 12
- [24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 2, 5, 7, 12, 13
- [25] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2, 1955. 4
- [26] Jie Lei, Tamara L Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. *NeurIPS*, 2021. 2, 3, 4, 5, 6, 7, 8, 12, 13
- [27] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 6, 7
- [28] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022. 2, 3
- [29] Daizong Liu and Wei Hu. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *ACM MM*, 2022. 6, 7
- [30] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *AAAI*, 2022. 2, 6, 7
- [31] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. In *EMNLP*, 2021. 1, 2, 6, 7
- [32] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, 2021. 2, 6, 7
- [33] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *ACM MM*, 2022. 6, 7
- [34] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *ACM MM*, 2020. 2, 6, 7
- [35] Daizong Liu, Xiaoye Qu, and Pan Zhou. Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. In *EMNLP*, 2021. 5, 6, 7
- [36] Daizong Liu, Xiaoye Qu, Pan Zhou, and Yang Liu. Exploring motion and appearance information for temporal sentence grounding. In *AAAI*, 2022. 2
- [37] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *ACM SIGIR*, 2018. 1, 2
- [38] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *ACM MM*, 2018. 2
- [39] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *ICLR*, 2022. 2, 3, 5
- [40] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*, 2015. 6, 7
- [41] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, 2022. 1, 2, 6, 7
- [42] 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. 2, 3, 4, 12
- [43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017. 7
- [44] Chujiu Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *EMNLP-IJCNLP*, 2019. 1, 2
- [45] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *ICCV*, 2021. 2, 3, 5
- [46] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*, 2023. 1, 2, 3, 5, 6, 7
- [47] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 2, 6, 7
- [48] Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. Zero-shot natural language video localization. In *ICCV*, 2021. 4
- [49] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *CVPR*, 2021. 2
- [50] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *BMVC*, 2020. 2
- [51] Costas Panagiotakis, Giorgos Karvounas, and Antonis Argiros. Unsupervised detection of periodic segments in videos. In *ICIP*, 2018. 4
- [52] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Sumgraph: Video summarization via recursive graph modeling. In *ECCV*, 2020. 1
- [53] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *CVPR*, 2022. 1
- [54] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metzger, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 1
- [55] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Fine-grained iterative attention network for temporal language localization in videos. In *ACM MM*, 2020. 6, 7
- [56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 6, 7
- [57] Hamid Rezaatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized in-

- tersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5
- [58] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*, 2020. 2
- [59] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *CVPR*, 2021. 2
- [60] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. Find and focus: Retrieve and localize video events with natural language queries. In *ECCV*, 2018. 2
- [61] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *ICCV*, 2021. 2, 3
- [62] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *CIKM*, 2016. 6, 7
- [63] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *ICCV*, 2021. 2
- [64] Jiaqi Tang, Zhaoyang Liu, Chen Qian, Wayne Wu, and Limin Wang. Progressive attention on multi-level dense difference maps for generic event boundary detection. In *CVPR*, 2022. 3
- [65] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 6
- [66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2, 3
- [67] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *CVPR*, 2021. 6, 7
- [68] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *AAAI*, 2022. 2, 3
- [69] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, 2019. 5
- [70] Sangmin Woo, Jinyoung Park, Inyong Koo, Sumin Lee, Minki Jeong, and Changick Kim. Explore and match: End-to-end video grounding with transformer. *arXiv preprint arXiv:2201.10168*, 2022. 2
- [71] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *CVPR*, 2021. 1
- [72] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *AAAI*, 2021. 2, 6, 7
- [73] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019. 2
- [74] Yifang Xu, Yunzhuo Sun, Yang Li, Yilei Shi, Xiaoxiang Zhu, and Sidan Du. Mh-detr: Video moment and highlight detection with cross-modal transformer. In *ACM MM*, 2023. 2, 3, 6, 7
- [75] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *ACM SIGIR*, 2021. 2
- [76] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021. 2
- [77] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *NeurIPS*, 2019. 1, 2
- [78] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019. 2
- [79] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020. 1, 2, 6, 7
- [80] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. Multi-modal relational graph for cross-modal video moment retrieval. In *CVPR*, 2021. 2
- [81] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019. 2
- [82] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, and Shijian Lu. Accelerating detr convergence via semantic-aligned matching. In *CVPR*, 2022. 2
- [83] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, 2020. 2
- [84] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Towards debiasing temporal sentence grounding in video. *arXiv preprint arXiv:2111.04321*, 2021. 2
- [85] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020. 2
- [86] Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *ACM MM*, 2019. 2
- [87] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *ACM SIGIR*, 2019. 1, 2
- [88] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. Cascaded prediction network via segment tree for temporal video grounding. In *CVPR*, 2021. 6, 7
- [89] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and debias for robust temporal grounding. In *CVPR*, 2021. 6, 7
- [90] Jiahao Zhu, Daizong Liu, Pan Zhou, Xing Di, Yu Cheng, Song Yang, Wenzheng Xu, Zichuan Xu, Yao Wan, Lichao Sun, et al. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. In *EMNLP*, 2022. 2, 6, 7
- [91] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 2, 3, 5

Appendix

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. A), and additional experimental results (Sec. B), including ablation studies and qualitative results.

A. 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 [23], 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 [23], we define the contrastive kernel $\mathbf{Z} \in \mathbb{R}^{z \times z}$ with the kernel size $z = 5$ as follows:

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & 0 & -1 & -1 \\ 1 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \end{bmatrix} \quad (18)$$

Since the kernel is designed to imitate the boundary pattern in the TSM [23], 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.

B. Additional Experiments

In this section, we present additional component analysis on QVHighlights [26] (Sec. B.1), ablation studies on Charades-STA [16] and ActivityNet Captions [24] (Sec. B.2), and qualitative results for video grounding (Sec. B.3).

B.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.

Table A1. 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 A2. 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

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 semantics in a given video. As shown in Tab. A1, 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. A2. 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 [42]. 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. A3. 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.

Table A3. 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 A4. 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

B.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 [16] and ActivityNet Captions [24] in Tab. A4. 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 [16] and ActivityNet Captions [24] in Fig. A1a and Fig. A2a. 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 [26] 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.

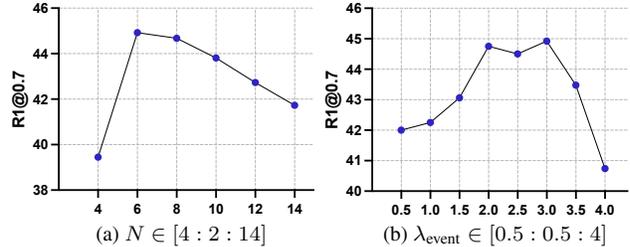


Figure A1. Hyper-parameter analysis on Charades-STA test split.

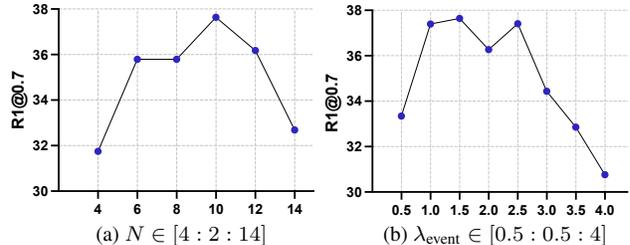


Figure A2. Hyper-parameter analysis on ActivityNet Captions val_2 split.

Effect of λ_{event} . The sensitivity of $\mathcal{L}_{\text{event}}$ on Charades-STA [16] and ActivityNet Captions [24] are in Fig. A1b and Fig. A2b. 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.

B.3. Qualitative results

We provide the qualitative results on QVHighlights [26] and Charades-STA [16] in Fig. A3 and Fig. A4 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. A3, 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. A4. 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



Figure A3. Qualitative results of our EaTR on QVHighlights val split.

Fig. A5, our EaTR fails to localize the given sentence on the fine-grained videos composed of visually similar frames.

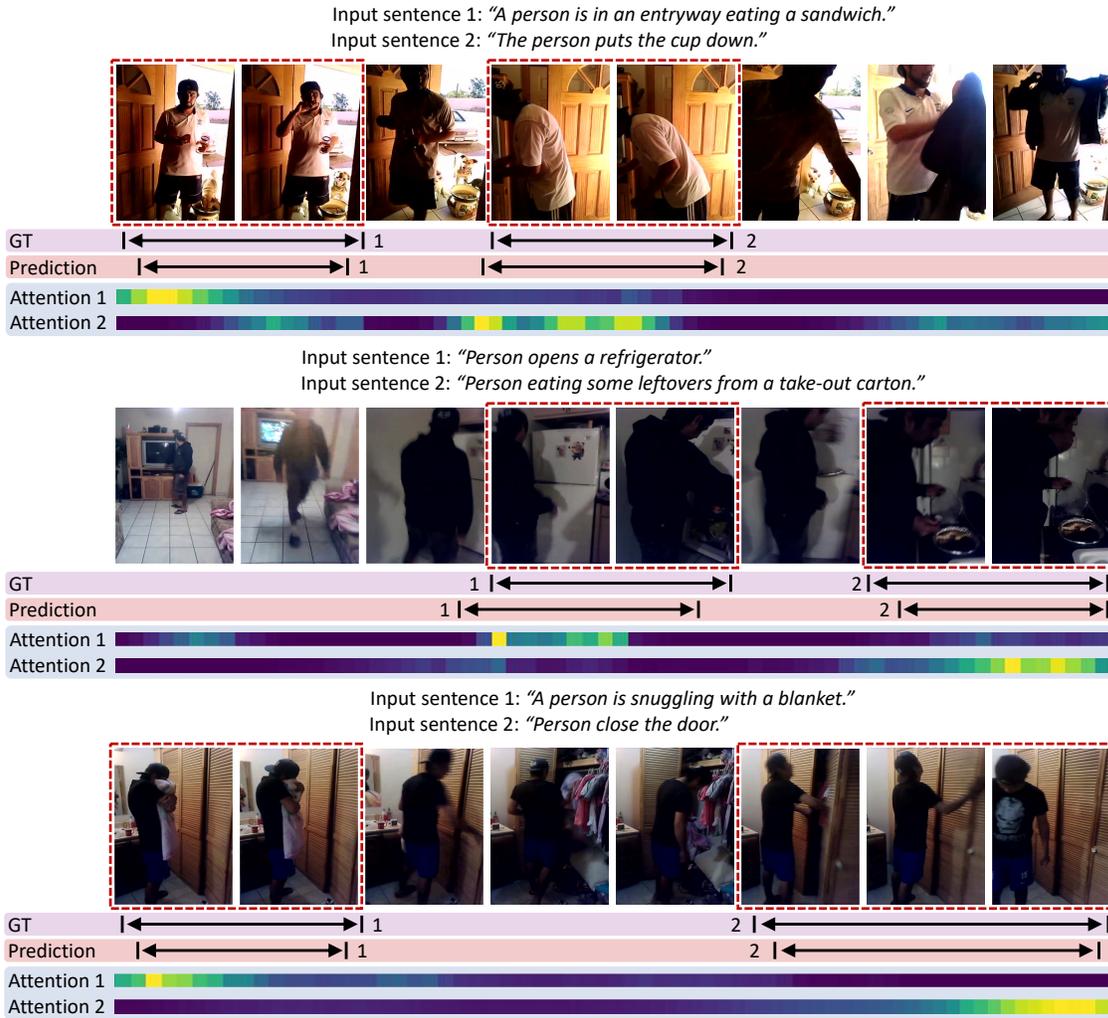


Figure A4. Qualitative results of our EaTR on Charades-STA test split.

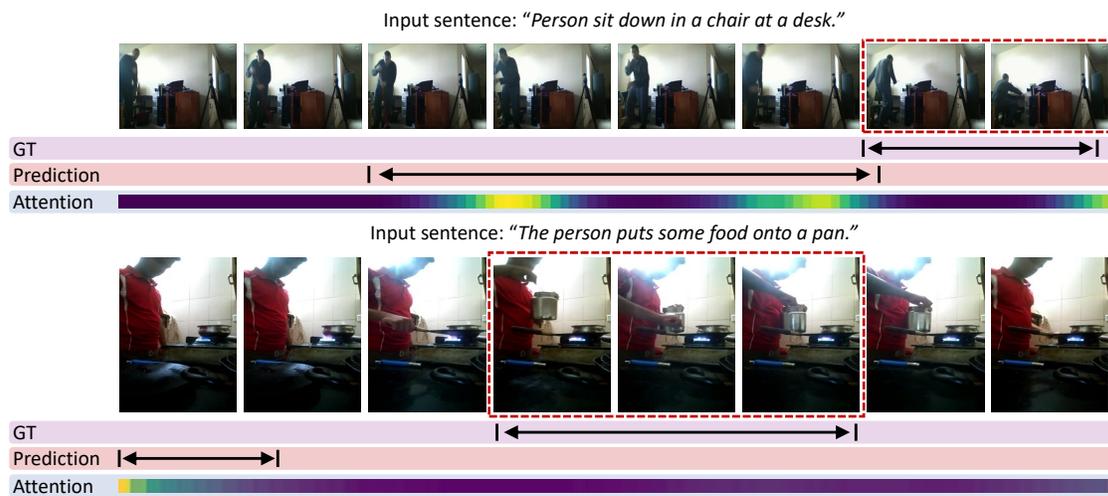


Figure A5. Failure cases of our EaTR on Charades-STA test split.