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

Jinhyun Jang\textsuperscript{1} Jungin Park\textsuperscript{1} Jin Kim\textsuperscript{1} Hyeongjun Kwon\textsuperscript{1} Kwanghoon Sohn\textsuperscript{1,2*}
\textsuperscript{1}Yonsei University \textsuperscript{2}Korea 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
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 \((i.e., \text{distinctive TSM})\) 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 events referring to a given sentence \(S\) in an untrimmed video \(V\). Recent DETR-based methods [26, 2, 74, 46] formulate learnable query embeddings \(Q\) \((i.e., \text{moment query})\) 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 \((C, P)\) of the moment reasoning network.

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 \( V \) of length \( L_v \) and a sentence \( S \) of length \( L_s \), we first encode the video and sentence representations using corresponding pretrained backbone networks (e.g. 3D [4], CLIP [56]) as follows:

\[
h_v = f_v(V) + PE \in \mathbb{R}^{L_v \times d}, \quad h_s = f_s(S) \in \mathbb{R}^{L_s \times d},
\]

where \( h_v \) and \( h_s \) 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 \( h_{enc} \):

\[
h_{enc} = f_{enc}(h_v \| h_s),
\]

where \( f_{enc}(\cdot) \) is the transformer encoder and \( \| \) is the concatenate operation. Following [26], we feed the partial representations corresponding to the video in \( h_{enc} \) 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]:

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

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 \( \alpha \) 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 \( h_v \) using a set of \( N \) learnable event slots \( E \in \mathbb{R}^{N \times d} \) and the slot attention mechanism [42]. Formally, the event slots iteratively interact with the video representations \( h_v \) for \( K \) iterations to group the visually similar frames and obtain the final event units \( E^K \). In \( k \)-th iteration, we first embed \( h_v \) and \( E^{k-1} \) using layer normalization followed by linear projections, such that:

\[
h'_v = (\text{LN}(h_v))W_1, \quad E^{k-1} = (\text{LN}(E^{k-1}))W_2,
\]

where \( h'_v \) and \( E^{k-1} \) are the embedded video representations and event slots, \( \text{LN}(\cdot) \) is layer normalization, \( W_1 \), and \( W_2 \) are the linear projection matrices. The \( k \)-th interaction matrix between \( h'_v \) and \( E^{k-1} \) can be computed as:

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

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

\[
U = (\hat{A}^k)\top(h_v)W_3 + E^{k-1}, \quad \hat{A}^k_{i,n} = \frac{A^k_{i,n}}{\sum_k A^k_{i,n}},
\]

\[
E^k = (\text{LN}(U))W_4 + U,
\]

with additional linear projection matrices \( W_3 \) and \( W_4 \). Different from the conventional slot attention [42], we replace the GRU layer with residual summation to avoid inefficient computations. Since each event slot \( e^K_n \in E^K \) contains visual information corresponding to the individual event of the video, we use \( E^K \) as the initial content queries \( C \). In addition, we project \( E^K \) to the 2-dimensional embedding space to derive the initial positional queries \( P \) that represent the center and duration of the referential search area for each moment query as:

\[
P = E^KW_p = \{(c_n, w_n)\}_{n=1}^N,
\]

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. \( W_1 \)) 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 matrix [51, 12, 48, 23]. Specifically, we employ the recent contrastive kernel [23] that computes the event boundary matrix [51, 12, 48, 23]. Specifically, we employ the recent contrastive kernel [23] that computes the event boundary matrix [51, 12, 48, 23]. Specifically, we employ the recent contrastive kernel [23] that computes the event boundary matrix [51, 12, 48, 23]. Specifically, we employ the recent contrastive kernel [23] that computes the event boundary matrix [51, 12, 48, 23].
where $L_{iou}$ is generalized temporal IoU [57]. $\lambda_{l_t}$ and $\lambda_{sal}$ are the balancing parameters. Given the optimal assignment, the event localization loss is defined as:

$$L_{event} = \sum_{i}^{N} C(\hat{P}_i, P_{\sigma(i)}). \quad (9)$$

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 $p_n = (c_n, w_n)$ is first extended to $d$-dimensional space through the sinusoidal positional encoding (PE) [45, 39], concatenation, and MLP layer as:

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

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 $C'$ can be obtained by:

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

where $\oplus$ is an element-wise summation.

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

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

The similarity between the $n$-th moment query $c_n'$ and aggregated sentence representation $\hat{c}_n$ 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:

$$g_n = \text{Sigmoid}(c_n', \hat{c}_n^T). \quad (13)$$

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

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

where $\text{Linear}$ is a single linear layer and $\odot$ is the element-wise multiplication. The enhanced moment queries then interact with the video-sentence representations $h_{enc}$ 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 query [91, 39]. The output of the transformer decoder $h_{dec}$ is fed to FFN for predicting the moment span $M$. 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 $M_t \in [0, 1]^2$ 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:

$$\hat{M}_t = \text{arg min}_{\sigma' \in \sigma_N} \sum_{i}^{N} \left[ -\lambda_{c} p_c, \sigma'(i) + C(M_t, M_{\sigma'(i)}) \right], \quad (15)$$

$$C(M_t, M_{\sigma'(i)}) = \lambda_{l_t} ||M_t - M_{\sigma'(i)}||_1 + \lambda_{sal}L_{iou}(\hat{M}_t, M_{\sigma'(i)}),$$

where $\lambda_{l_t}$, $\lambda_{sal}$ and $\lambda_{c}$ are the balancing parameters. With the optimal assignment $\hat{\sigma}'$, the moment localization loss [2, 26] is defined as:

$$L_{moment} = \sum_{i}^{N} \left[ -\lambda_{l_t} \log p_{c, \hat{\sigma}'(i)} + C(\hat{M}_t, M_{\hat{\sigma}'(i)}) \right]. \quad (16)$$

**Overall objectives.** The overall objective is defined as:

$$L_{overall} = L_{moment} + \lambda_{sal}L_{sal} + \lambda_{event}L_{event}, \quad (17)$$

where $\lambda_{sal}$ and $\lambda_{event}$ 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Video Grounding</th>
<th>HD</th>
<th>GFLOPs</th>
<th>Params</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1 @0.5 @0.7</td>
<td>mAP @0.5 @0.75</td>
<td>Avg.</td>
<td>mAP</td>
</tr>
<tr>
<td>BeautyThumb [62]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DVSE [40]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MCN [1]</td>
<td>11.41</td>
<td>2.72</td>
<td>24.94</td>
<td>8.22</td>
</tr>
<tr>
<td>CAL [13]</td>
<td>25.49</td>
<td>11.54</td>
<td>23.40</td>
<td>7.65</td>
</tr>
<tr>
<td>CLIP [56]</td>
<td>16.88</td>
<td>5.19</td>
<td>18.11</td>
<td>7.00</td>
</tr>
<tr>
<td>XML [27]</td>
<td>41.83</td>
<td>30.35</td>
<td>44.63</td>
<td>31.73</td>
</tr>
<tr>
<td>XML+ [27]</td>
<td>46.69</td>
<td>33.46</td>
<td>47.89</td>
<td>34.67</td>
</tr>
<tr>
<td>Moment-DETR [26]</td>
<td>52.89 ±2.3</td>
<td>33.02 ±1.7</td>
<td>54.82 ±1.7</td>
<td>29.40 ±1.7</td>
</tr>
<tr>
<td>UMT† [41]</td>
<td>56.23</td>
<td>41.18</td>
<td>53.83</td>
<td>37.01</td>
</tr>
<tr>
<td>MH-DETR [74]</td>
<td>60.05</td>
<td>42.48</td>
<td>60.75</td>
<td>38.13</td>
</tr>
<tr>
<td>QD-DETR [46]</td>
<td>62.40 ±1.1</td>
<td>44.98 ±0.8</td>
<td>62.52 ±0.6</td>
<td>39.88 ±0.7</td>
</tr>
<tr>
<td>Ours</td>
<td>61.36 ±1.2</td>
<td>45.79 ±0.7</td>
<td>61.86 ±0.6</td>
<td>41.91 ±0.6</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Charades-STA R1@0.5 R1@0.7</th>
<th>ActivityNet Captions R1@0.5 R1@0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPNet [72]</td>
<td>50.75 31.64</td>
<td>42.07 24.69</td>
</tr>
<tr>
<td>DRN [79]</td>
<td>53.09 31.75</td>
<td>45.45 24.36</td>
</tr>
<tr>
<td>FIAN [55]</td>
<td>58.55 37.72</td>
<td>47.90 29.81</td>
</tr>
<tr>
<td>LGI [47]</td>
<td>59.46 35.48</td>
<td>41.51 23.07</td>
</tr>
<tr>
<td>DeNet [89]</td>
<td>59.70 38.52</td>
<td>43.79 -</td>
</tr>
<tr>
<td>CPN [88]</td>
<td>59.77 36.67</td>
<td>45.10 28.10</td>
</tr>
<tr>
<td>CSMGAN [34]</td>
<td>60.04 37.34</td>
<td>49.11 29.15</td>
</tr>
<tr>
<td>SSCS [11]</td>
<td>60.75 36.19</td>
<td>46.67 27.56</td>
</tr>
<tr>
<td>CBLN [32]</td>
<td>61.13 38.22</td>
<td>48.12 27.60</td>
</tr>
<tr>
<td>IA-Net [35]</td>
<td>61.29 37.91</td>
<td>48.57 27.95</td>
</tr>
<tr>
<td>APGN [31]</td>
<td>62.58 38.86</td>
<td>48.92 28.64</td>
</tr>
<tr>
<td>MGSL-Net [30]</td>
<td>63.98 41.03</td>
<td>51.87 31.42</td>
</tr>
<tr>
<td>SMIN [67]</td>
<td>64.04 40.75</td>
<td>48.46 30.34</td>
</tr>
<tr>
<td>SLIP [29]</td>
<td>64.35 40.43</td>
<td>52.89 32.04</td>
</tr>
<tr>
<td>D-TSG [33]</td>
<td>65.05 42.77</td>
<td>54.29 33.64</td>
</tr>
<tr>
<td>SSRN [90]</td>
<td>65.59 42.65</td>
<td>54.49 33.15</td>
</tr>
<tr>
<td>Ours</td>
<td>68.47 44.92</td>
<td>58.18 37.64</td>
</tr>
</tbody>
</table>

• **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
balancing parameters for the total loss function are set to $\lambda_1 = 10$, $\lambda_{iou} = 1$, $\lambda_c = 4$, and $\alpha = 0.2$. $\lambda_{sal}$ 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.

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
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 $L_{\text{event}}$. 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 $\lambda_{\text{event}}$. The balancing parameter $\lambda_{\text{event}}$ controls the impact of a newly introduced event loss $L_{\text{event}}$ in the total training objective. To study the sensitivity of $L_{\text{event}}$, we report the performance according to $\lambda_{\text{event}}$ in Fig. 3b. The result shows the performance monotonically increases with $1 \leq \lambda_{\text{event}} \leq 3$. The values of $\lambda_{\text{event}}$ 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


