

# Let Me Finish My Sentence: Video Temporal Grounding with Holistic Text Understanding

Jongbin Woo  
KAIST  
Daejeon, Republic of Korea  
jongbin.woo@kaist.ac.kr

Hyeonggon Ryu  
KAIST  
Daejeon, Republic of Korea  
gonhy.ryu@kaist.ac.kr

Youngjoon Jang  
KAIST  
Daejeon, Republic of Korea  
wgs01088@kaist.ac.kr

Jae Won Cho  
Sejong University  
Seoul, Republic of Korea  
chojw@sejong.ac.kr

Joon Son Chung  
KAIST  
Daejeon, Republic of Korea  
joonson@kaist.ac.kr

## ABSTRACT

Video Temporal Grounding (VTG) aims to identify visual frames in a video clip that match text queries. Recent studies in VTG employ cross-attention to correlate visual frames and text queries as individual token sequences. However, these approaches overlook a crucial aspect of the problem: a *holistic understanding* of the query sentence. A model may capture correlations between individual word tokens and arbitrary visual frames while possibly missing out on the global meaning. To address this, we introduce two primary contributions: (1) a visual frame-level gate mechanism that incorporates holistic textual information, (2) cross-modal alignment loss to learn the fine-grained correlation between query and relevant frames. As a result, we regularize the effect of individual word tokens and suppress irrelevant visual frames. We demonstrate that our method outperforms state-of-the-art approaches in VTG benchmarks, indicating that holistic text understanding guides the model to focus on the semantically important parts within the video.

## CCS CONCEPTS

• **Computing methodologies** → **Video summarization; Scene understanding; Activity recognition and understanding;** • **Information systems** → **Video search.**

## KEYWORDS

Video Temporal Grounding, Video Moment Retrieval, Video Highlight Detection

### ACM Reference Format:

Jongbin Woo, Hyeonggon Ryu, Youngjoon Jang, Jae Won Cho, and Joon Son Chung. 2024. Let Me Finish My Sentence: Video Temporal Grounding with Holistic Text Understanding. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28–November 1, 2024, Melbourne, VIC, Australia. *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24)*, October 28–November 1, 2024, Melbourne, Australia. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3664647.3681514>

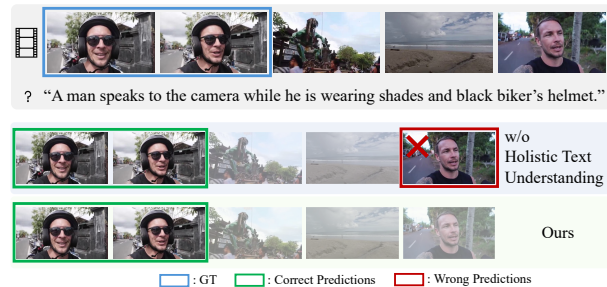
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/10.1145/3664647.3681514>



**Figure 1: This example shows the critical role of holistic text understanding in Video Temporal Grounding. Unlike previous works that do not take holistic text understanding into account, our method effectively filters out frames that do not correspond to the full context of the query. Here, our model does not predict the final frames due to the absence of the helmet and shades mentioned in the query.**

## 1 INTRODUCTION

Recently, the boom of video and streaming platforms such as Disney+, YouTube, TikTok, Netflix, etc., has led to an abundance of online video content. Naturally, the growing pool of videos from various platforms sparked an interest in efficiently searching videos using text inputs. Video Temporal Grounding (VTG) is a prominent research area within video-text search, that focuses on grounding visual frames that correspond to custom queries. Within VTG, various tasks such as Moment Retrieval (MR) [8, 19, 30, 33, 59, 60] and Highlight Detection (HD) [26, 45, 56] has been proposed.

The goal of MR is to identify time intervals that are highly relevant to text queries, while HD assesses the significance of each video frame to select the most significant segments. HD can be categorized into two perspectives: query-independent and query-dependent. This paper focuses exclusively on query-dependent HD, which utilizes text queries to analyze video content. Despite the distinct operational focuses of MR and HD, both tasks share the core aim of aligning video content with corresponding natural language queries. Recognizing their shared goal, [18] introduced the QVHighlights dataset. This allows for simultaneous training on both MR and HD, promoting a unified approach to VTG.

Prior approaches have focused on developing cross-modal interaction strategies [10, 18] or developed models specifically for

the demands of the MR and HD tasks [44, 51]. However, despite these advancements, existing models often treat the text query as a sequence of tokens rather than an entire sentence. These approaches may neglect the overall textual semantics, as individual text tokens either lack the capacity to convey the complete meaning and/or cause the model to attend to unrelated words or frames. This oversight can limit the model’s ability to fully capture the intent of the query. For instance, as shown in Fig. 1, the model without holistic text understanding may highlight the latter part of a video in response to the tokens inside the clause “A man speaks to the camera.” To tackle this issue, we focus on utilizing the global information contained within text queries, emphasizing its importance in accurately identifying the most relevant video frames.

To this end, we propose a novel framework that utilizes a holistic, or *global* text anchor—representing the full input sentence—to selectively suppress less relevant video frames while emphasizing the relevant ones. Leveraging this specialized token, our framework introduces a gated cross-attention mechanism that effectively filters out irrelevant video content. The gated cross-attention employs two gating mechanisms: *local* and *non-local* gates. The local gate assigns weights based on channel-wise similarity between the text anchor and individual frames (frame-level). In contrast, the non-local gate assigns weights by evaluating the overall relevance between the text anchor and the entire video (clip-level), prioritizing frames that exhibit greater contextual alignment with the global text query. The overall relevance is computed through an anchor-query cross-attention mechanism that employs the text anchor as a query to interact with video frames. An attention map derived from this interaction effectively assesses clip-level correlation, thereby emphasizing the focus on contextually pertinent frames. To further refine the precision in assessing similarity, we introduce two fine-grained alignment losses that optimize clip-level consistency and frame-level relevance. These losses use the text query as an anchor, enabling a more targeted and accurate alignment between the video content and the corresponding textual information. The clip-level consistency loss aims to minimize the discrepancy between the anchor and the clip-level video feature outputted from the anchor-query cross-attention layer. This reduction aims to enhance the accuracy in determining the relative significance of each video frame in relation to the anchor. On the other hand, the frame-level relevance loss aims to ensure that frames relevant to the text query are closely aligned with the global text query representation. It minimizes the distance between the global query and the corresponding frames, thus enhancing the model’s ability to more accurately align relevant video content with the text.

To demonstrate the effectiveness of our proposed framework, we conduct extensive experiments on the QVHighlights [18] dataset, as well as on other notable VTG benchmarks, including CharadesSTA [8] and TACoS [33]. Our experimental results show that our proposed method, to the best of our knowledge, achieves state-of-the-art performance compared to previous methods. Additionally, we carry out a detailed ablation study to further assess and validate the advantages of our proposed method. The contributions of our approach are summarized as follows:

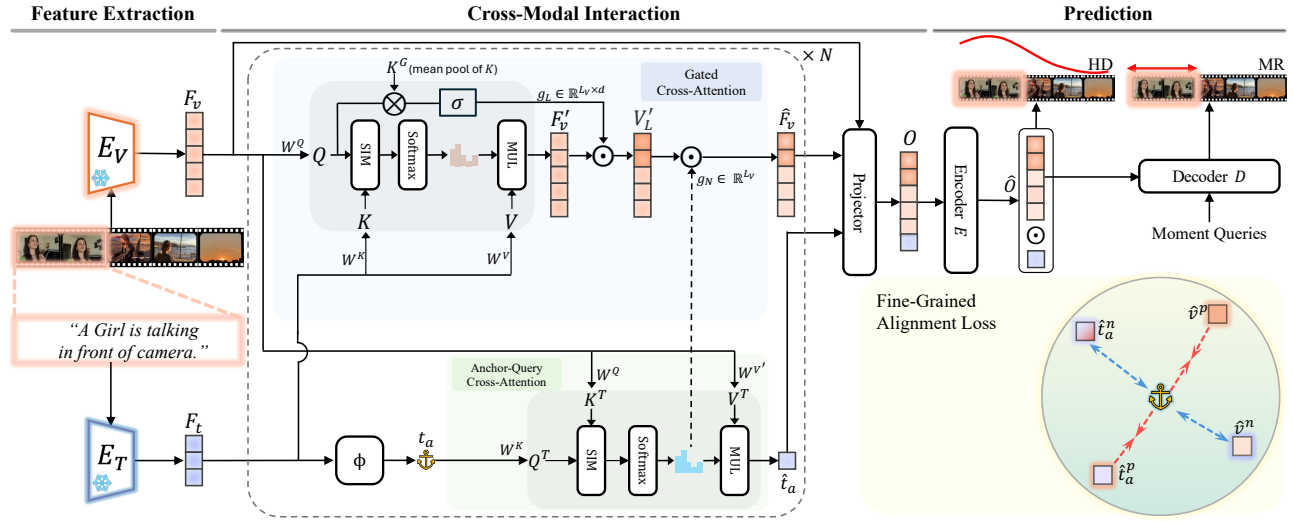
- We introduce a novel framework that utilizes a *global text-anchor* based approach to enhance video grounding, leveraging holistic textual information to accurately filter and prioritize relevant frames.
- To best exploit our global text-anchor, we introduce two cross-attention mechanisms where we integrate both gated and anchor-query cross-attention for deeper video-text interaction.
- We propose fine-grained alignment loss functions (clip-level and frame-level) anchored by the text query, designed to refine the accuracy in measuring similarities and aligning video content with the given text queries.

## 2 RELATED WORKS

**Moment Retrieval (MR)** is a task that identifies temporal moments that are highly relevant to a given natural language query. Within MR, previous works have been categorized in two sections: proposal-based and proposal-free. As in the name, proposal-based methods have generally followed a pipeline where a model generates candidate windows from the entirety of the video, then grants a rank based on the matched scores. This approach often relies on predefined temporal structures like sliding windows [1, 8, 24, 39, 61] or temporal anchors [3, 41, 53, 55, 58, 60, 62] to formulate candidate moments. On the other hand, proposal-free methods solve the task as a regression problem where they directly regress the start and end time frames through various multimodal methods. Within this line of works, previous methods have utilized methods such as multimodal coattention [9, 22, 29, 57, 59], dynamic filters [35], and additional features [4, 35] to varying degrees of success. Both proposal-based and proposal-free methods are effective but rely on hand-crafted processes such as proposal generation and non-maximum suppression.

**Highlight Detection (HD)** is a task that focuses on the evaluation of individual video clips’ significance by assigning them clip-wise saliency scores and highlighting the segments with the highest scores. However, HD datasets [36, 42, 46] are usually domain-specific and operate independently of textual queries. The common datasets [26, 56] available for query-based highlight detection offer a limited number of annotated frames for training and evaluation. The scarcity of query-dependent HD datasets underscores the perception of HD as primarily a vision-only problem. Unlike earlier HD datasets that were query-agnostic, we explore and test on an HD task that offers a saliency score for query-relevant clips, facilitating models to perform query-dependent highlight detection.

**Video Temporal Grounding** seeks to combine the two aforementioned tasks. Despite their similar objectives, the absence of a unified dataset supporting both tasks has constrained simultaneous exploration and combination of these two fields. To this end, [18] released QVHighlights and proposed Moment-DETR as a simple baseline. Subsequently, UMT [27] explored the addition of audio cues to enrich query generation, and QD-DETR [28] enhanced query-dependent video representations through a specialized cross-attention module. MH-DETR [54] explores further cross-modal integration by merging visual and textual features through a pooling mechanism, while UniVTG [23] proposes a unified grounding



**Figure 2: The pipeline of our framework consists of four components: feature extraction, cross-modal interaction, fine-grained alignment loss, and prediction. First, we extract visual and text features with frozen pre-trained encoders. Since the task requires cross-modal understanding and suppression of irrelevant information, we incorporate the gated cross-attention mechanism for the cross-modal interaction. The encoded features of cross-modal interaction are leveraged through the fine-grained alignment loss, which guides the model to enhance cross-modal alignment. Finally, the visual and textual representations from this aligned embedding space are fed into the prediction section to produce task-specific outputs.**

model including video summarization. More recent efforts by TR-DETR [44] and UVCOM [51] incorporate distinct task characteristics into their frameworks. Even with these advancements, existing models often interpret text queries as a sequence of individual tokens, neglecting the holistic semantics of the entire query. To address this issue, we introduce a novel framework that emphasizes holistic textual understanding to accurately identify and emphasize relevant video segments.

**Cross-modal alignment** is an important concept in tasks that involve multiple modalities [5, 6, 17]. This type of alignment ensures that representations from various modalities, such as visual and textual data, are positioned in a shared embedding space. This is often achieved through contrastive learning methods [15, 16, 20, 32, 37, 38, 50, 52]. Particularly in the video-text domain, approaches such as [2, 10, 13, 14, 49] aim to align these modalities in a more fine-grained manner.

### 3 METHOD

In this section, we provide an in-depth explanation of our approach that centers on the global text-anchor. As shown in Fig. 2, our framework integrates four components: feature extraction, cross-modal interaction, fine-grained alignment loss, and prediction. We first outline the task while detailing the feature extraction process in Sec. 3.1. We then introduce the core method of our framework in Sec. 3.2, where the concept of the global text anchor is introduced to encapsulate the holistic textual context. In Sec. 3.3, we detail two critical cross-attention mechanisms—gated cross-attention and anchor-query cross-attention. Both mechanisms utilize the global text anchor to enhance the interaction between the video content and the text query. We then present newly devised fine-grained

alignment losses in Sec. 3.4, aimed at improving the alignment between video content and text queries. Finally, we detail the inference procedure, explaining how our model concurrently predicts for both Moment Retrieval (MR) and Highlight Detection (HD).

#### 3.1 Preliminaries

Given a video  $V \in \mathbb{R}^{L_v \times H \times W \times 3}$ , consisting of  $L_v$  frames sampled from the original video at specific intervals and a natural language text query  $T$  comprising  $L_t$  tokens, the objective is to identify all relevant moments  $\{m_n = (m_{c_n}, m_{\sigma_n})\}_{n=1}^N$ , where  $m_{c_n}$  denotes the center coordinate of a moment and  $m_{\sigma_n}$  represents its width. Additionally, the model predicts a frame-level saliency score  $\{s_i\}_{i=1}^{L_v}$  concurrently. For feature extraction, following previous works [18, 23, 28, 44, 51], we employ pre-trained encoders  $E_V$  and  $E_T$  to extract visual features  $F_v = [v_1, v_2, \dots, v_{L_v}] \in \mathbb{R}^{L_v \times d_v}$  and text features  $F_t = [t_1, t_2, \dots, t_{L_t}] \in \mathbb{R}^{L_t \times d_t}$ , respectively. Separate Multi-layer Perceptron (MLP) are used to project the video and text features into a shared embedding space of the same dimension  $d$ . In the following sections, we describe in detail our contributions of cross-modal interaction, fine-grained alignment loss, and how we train our model with our proposed methods.

#### 3.2 Global Text Anchor

The primary goal of the VTG task is to identify video sequences that correspond to the *entire* query. In order to do this effectively, an essential aspect of VTG is that it needs to capture the *overall meaning of a given text query*. Therefore, we propose to use a global text anchor that embodies a holistic understanding of the sentence, as emphasized in our title: *Let Me Finish My Sentence*. Adopting the global text query as an anchor offers two main advantages: (1) the text modality is generally less noisy compared to the visual

modality due to its discreteness, and (2) we can effectively leverage prior knowledge from language models trained on a large corpora, which are well-generalized across various domains. We employ a global mean pooling operation, denoted as  $\phi$ , to derive a global text anchor from token-level text representations. This anchor is then integrated into our cross-modal interaction module and is further detailed in the following section.

### 3.3 Cross-Modal Interaction

As mentioned, we hypothesize the need for holistic understanding of the text query so that the model may take into account the entire text query instead of being biased to certain words in the query. Hence, we propose and introduce a global text anchor to address the aforementioned issues in these ways: 1) suppressing non-essential video frames through a gating mechanism with the global text anchor, 2) enhancing cross-modal alignment between video content and global textual information. The versatile use of the global text-anchor enhance the contextual relevance of video-text interactions and cross-model understanding.

**Gated Cross-Attention.** As shown in Fig. 2, we extend the conventional cross-attention mechanism [48] by incorporating *local* and *non-local* gates. The standard cross-attention is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V, \quad (1)$$

where  $Q$ ,  $K$ , and  $V$  represent the query, key, and value matrices respectively, and  $d_k$  is the dimensionality of the key. In our model, this formula becomes:

$$\text{Attention}(Q, K, V) = \text{Attention}(W^Q F_v, W^K F_t, W^V F_t) = F'_v, \quad (2)$$

where  $W^Q$ ,  $W^K$ , and  $W^V$  are learned weight that project the video features ( $F_v$ ) and text features ( $F_t$ ) into the query, key, and value.

The local gate weight, denoted by  $g_L$ , employs element-wise multiplication and sigmoid activation to assess the relevance between the video clip and the text query, utilizing the global key  $K^G \in \mathbb{R}^d$ . This key  $K^G$  is derived by mean pooling  $K$  across time dimensions:

$$g_L = \sigma \left( W_q^g Q \odot W_k^g K^G \right) \in \mathbb{R}^{L_v \times d}, \quad (3)$$

where  $W_q^g$  and  $W_k^g$  are trainable weights applied to the query  $Q$  and the global key  $K^G$ , respectively, to capture channel-wise relevance between single frame and global text. The output,  $g_L$ , modulates the feature vector  $V'$  through element-wise multiplication, resulting in a relevance-enhanced feature representation  $V'_L$ :

$$V'_L = g_L \odot F'_v. \quad (4)$$

The non-local gate weight  $g_N$ , designed for broad relevance assessment across video clips, modifies the interaction to enhance contextually pertinent video frames:

$$\hat{F}_v = g_N \odot V'_L = \{\hat{v}_1, \hat{v}_2, \dots, \hat{v}_{L_v}\}. \quad (5)$$

This mechanism,  $g_N \in \mathbb{R}^d$ , computes min-max normalized attention scores between the global text anchor and video frames, effectively weighting the relative significance of each video frame in relation to the global text. It serves as an effective filter that emphasizes frames with substantial contextual relevance while suppressing the less relevant ones.

Through the integration of both local and non-local gates, our model emphasizes clips closely aligned with the text query while minimizing the influence of non-relevant clips in subsequent steps.

**Anchor-Query Cross-Attention.** Here, the global text anchor  $t_a$ , obtained through the mean pooling operator  $\phi(F_t)$  is encoded into the visually enriched text-query. The video clip representations  $F_v$  are adapted as both keys and values within the cross-attention mechanism. While the gated cross-attention approach learns the alignment between individual video frames and text tokens, the anchor-query cross-attention addresses the relative relevance between the global text representation and each of the video frames. The operational equation is defined as follows:

$$\begin{aligned} \text{Attention}(Q_a, K_a, V_a) &= \text{Attention}(W^K t^a, W^Q F_v, W^{V'} F_v) \\ &= \text{softmax} \left( \frac{Q_a K_a^T}{\sqrt{d_k}} \right) V_a = \hat{t}_a. \end{aligned} \quad (6)$$

In this configuration,  $W^Q$  and  $W^K$  serve as projection layers for video and text modalities within the gated cross-attention mechanism, respectively. They aid in establishing the understanding between video content and textual information. Since  $\text{softmax} \left( \frac{Q_t K_t^T}{\sqrt{d_k}} \right)$  represents the similarity score between the text query and video frames, we represent this calculation as  $g_N$ .

The two parts in the cross-modal interaction module refine the intermediate features based on cross-modal understanding. The module guides the model to emphasize the frames which are contextually relevant to the text query.

### 3.4 Fine-Grained alignment loss

Understanding of the fine-grained correlation between text queries and video clips is essential for MR and HD tasks. Moreover, the validity of the local and non-local gates depends on the reliability of cross-modal understanding. To address this, we propose two loss functions to better learn cross-modal alignment: 1) preserving the consistency of the global text representation before and after the anchor-query cross-attention step, and 2) inducing the model to capture the frame-level relevance between the visual frame and global text representation.

**Clip-Level Consistency Loss.** Suppose we are given a mini-batch of feature pairs  $\{(F_v^i, F_t^i)\}_{i=1}^B$ , where  $B$  refers to the size of the mini-batch. The loss function incorporates the global representation  $t_a^i = \phi(F_t^i)$  of text query  $F_t^i$ , and the output of the anchor-query cross-attention layer  $\psi$ ,  $\hat{t}_a^{ij} = \psi(t_a^i, F_v^j)$ . We employ  $\hat{t}_a^p$  and  $\hat{t}_a^n$ , positive and negative visual-enriched global text representation from paired ( $i = j$ ), and unpaired ( $i \neq j$ ) video-text respectively. The proposed loss function is designed to minimize the distance between the global text anchor  $t_a$  and  $\hat{t}_a^p$ , while maximizing the distance between  $t_a$  and  $\hat{t}_a^n$ , expressed as:

$$\begin{aligned} \mathcal{L}_{\text{clip}} &= -\frac{1}{B} \sum_{i=1}^B \log \left( \frac{\exp(\hat{t}_a^{ii} \cdot t_a^i)}{\sum_{j=1}^B \exp(\hat{t}_a^{ij} \cdot t_a^i)} \right) \\ &\quad - \frac{1}{B} \sum_{j=1}^B \log \left( \frac{\exp(\hat{t}_a^{jj} \cdot t_a^j)}{\sum_{i=1}^B \exp(\hat{t}_a^{ij} \cdot t_a^j)} \right) \end{aligned} \quad (7)$$

Optimizing the model with this objective enhances the cross-modal alignment between the text anchor  $t_a$  and semantically relevant video clips. By minimizing the distance between  $t_a$  and the visually enriched text query  $\hat{t}_a^p$ , which incorporates both relevant and irrelevant video frames, the approach guides the text anchor’s attention strongly towards semantically relevant video frames.

The training objective aims to refine the video clip representation  $\hat{F}_v$ , derived from the gated cross-attention, to align closely with the text anchor  $t_a$  for relevant text queries, and to diverge when the queries are irrelevant. This method conditions the model to enhance the semantic correlation between video clips and the corresponding text query, ensuring their representations in the embedding space accurately reflect their contextual relevance.

**Frame-Level Relevance Losses.** This loss function refines the representation of video clips  $\hat{F}_v$ , which has been processed through gated cross-attention, by optimizing the alignment between video frames and the text anchor. Specifically, it enhances the similarity between relevant video frames  $\hat{v}^p$  and the text anchor  $t_a$ , while reducing the similarity with irrelevant frames  $\hat{v}^n$ . This loss guides the model to learn the fine-grained correlation between visual frames and the corresponding text query, ensuring their representations in the embedding space accurately reflect their contextual relevance.

The similarity score between the  $i$ -th frame  $\hat{v}_i$  and the text anchor  $t_a$  is given by  $D^i = \sigma(\hat{v}_i \cdot a)$ , where  $\sigma$  denotes a sigmoid that incorporates the similarity score. The loss function then is:

$$\mathcal{L}_{\text{frame}} = \sum_{i=1}^{L_v} C^i \log(D^i) + (1 - C^i) \log(1 - D^i) \quad (8)$$

Here,  $C^i$  is a binary indicator reflecting whether the  $i$ -th clip is relevant (1) or irrelevant (0) to the text query.

### 3.5 Prediction and Losses

In our cross-modal interaction module, which consists of  $N$  transformer layers, we aim to generate a composite representation. This is achieved by channel-wise concatenating the intermediate outputs from each layer  $O_l$  with the input video features  $F_v$ , and then projecting this concatenation into a  $d$ -dimensional space using a linear projection layer  $f$ :

$$O' = f(\text{Concat}_{\text{channel}}(F_v, O_1, O_2, \dots, O_N)) \in \mathbb{R}^{L_v \times d}. \quad (9)$$

Subsequently, the output  $\hat{t}_a$  is concatenated with  $O$  in a temporal manner to form the basis for a query-dependent adaptive classifier [28, 43]:

$$O = \text{Concat}_{\text{temporal}}(O', \hat{t}_a) \in \mathbb{R}^{(L_v+1) \times d}. \quad (10)$$

This concatenated output, denoted as  $O$ , is considered the final feature set of the cross-modal interaction module.

**Highlight Prediction.** The final feature set  $O$  is processed through a transformer encoder  $E$ . Separate Multi-layer Perceptron (MLP) is used to project the video and text features into a shared embedding space of the same dimension  $d$ , producing  $\hat{O} = \{\hat{o}_1, \dots, \hat{o}_{L_v}, \hat{t}'_a\}$ . Following QD-DETR [11, 28], we calculate a vector of saliency scores  $S \in \mathbb{R}^{L_v}$  for every frame in the video as follows:

$$S_i = \frac{w_s \cdot \hat{t}'_a \cdot (w_v \cdot \hat{o}_i)}{d}, \quad \text{for } i = 1, \dots, L_v, \quad (11)$$

where  $w_s$  and  $w_v$  are learnable parameters applied to the query representation and each video frame representation respectively.

**Moment Retrieval Prediction.** Following decoder strategies from prior research [25, 28, 44], we utilize dynamic anchor boxes to represent moment queries (which is clearly separate from text queries). Together with the output  $\hat{O} = \{\hat{o}_1, \dots, \hat{o}_{L_v}\}$ , this input is provided to the decoder  $D$ , resulting in moment features  $Q$ .  $Q$  undergoes processing via a Multi-Layer Perceptron (MLP) and a sigmoid activation to generate predictions for moment dimensions ( $\hat{m}$ ), yielding  $M$  moment predictions. Concurrently, another linear layer equipped with a softmax activation categorizes each predicted moment as foreground or background ( $\hat{p}$ ).

**Highlight Loss.** The highlight loss  $L_{hl}$  is comprised of a margin contrastive loss  $L_{\text{margin}}$  and a rank-aware contrastive loss  $L_{\text{rank}}$ . Margin loss contrasts high and low scoring frames within ( $t_{\text{high}}, t_{\text{low}}$ ) and outside ( $t_{\text{in}}, t_{\text{out}}$ ) ground-truth moments, given by:

$$\mathcal{L}_{\text{margin}} = \max(0, \Delta + S(t_{\text{low}}) - S(t_{\text{high}})) + \max(0, \Delta + S(t_{\text{out}}) - S(t_{\text{in}})), \quad (12)$$

with  $\Delta$  denoting the margin. Following QD-DETR [28], the rank-aware loss is:

$$\mathcal{L}_{\text{rank}} = - \sum_{r=1}^R \log \frac{\sum_{x \in X_r^{\text{pos}}} \exp(S_x / \tau)}{\sum_{x \in (X_r^{\text{pos}} \cup X_r^{\text{neg}})} \exp(S_x / \tau)}, \quad (13)$$

where  $X_r^{\text{pos}}$  and  $X_r^{\text{neg}}$  are the indexes of positive and negative frames within the  $r$ -th rank group, respectively, and  $\tau$  is a temperature scaling parameter. The total highlight loss is as follows:

$$\mathcal{L}_{\text{hd}} = \mathcal{L}_{\text{margin}} + \mathcal{L}_{\text{rank}}. \quad (14)$$

**Moment Retrieval Loss.** To address the set prediction challenge in moment retrieval without a direct one-to-one correspondence between ground truth and predictions, we apply the Hungarian matching algorithm. This algorithm pairs ground truth moments with predictions, where  $\hat{\sigma}(i)$  indexes the predicted moment matched to the  $i$ -th ground truth moment. The span loss for matched pairs is as follows:

$$\mathcal{L}_{\text{span}}(m_i, \hat{m}_{\hat{\sigma}(i)}) = \lambda_{L1} \|m_i - \hat{m}_{\hat{\sigma}(i)}\| + \lambda_{\text{iou}} \mathcal{L}_{\text{iou}}(m_i, \hat{m}_{\hat{\sigma}(i)}), \quad (15)$$

incorporating the generalized IOU loss [34]. The moment retrieval loss combines classification and span losses:

$$\mathcal{L}_{\text{mr}} = \sum_{i=1}^N [-\lambda_{\text{cls}} \log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{I}(c_i \neq \emptyset) \mathcal{L}_{\text{span}}(m_i, \hat{m}_{\hat{\sigma}(i)})], \quad (16)$$

where  $\mathbb{I}(\cdot)$  applies span loss only to non-empty ground truth moments.

**Overall Loss.** The overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{hd}} + \mathcal{L}_{\text{mr}} + \lambda_{\text{clip}} \mathcal{L}_{\text{clip}} + \lambda_{\text{frame}} \mathcal{L}_{\text{frame}}, \quad (17)$$

where the coefficients  $\lambda_{\text{clip}}$  and  $\lambda_{\text{frame}}$  are parameters that balance the contribution of clip-level and frame-level losses to the overall loss, respectively.

## 4 EXPERIMENTS

In this section, we outline the experimental setup and list details on the dataset, evaluation metrics, and implementation specifics as discussed in Sec. 4.1. Then, we compare the performance of

**Table 1: Experimental results on the QVHighlights test split, comparing performance in Moment Retrieval (MR) and Highlight Detection (HD). All models listed utilized uniform video (SlowFast and CLIP) and text features (CLIP).**

Method	R1		MR			HD	
	@0.5	@0.7	@0.5	mAP	Avg.	$\geq$ Very Good	
				@0.75		mAP	HIT@1
XML+ [19] <small>ECCV2020</small>	46.69	33.46	47.89	34.67	34.90	35.38	55.06
Moment-DETR [18] <small>NeurIPS2021</small>	52.89	33.02	54.82	29.40	30.73	35.69	55.60
UMT [27] <small>CVPR2022</small>	56.23	41.18	53.83	37.01	36.12	38.18	59.99
MomentDiff [21] <small>NeurIPS2023</small>	57.42	39.66	54.02	35.73	35.95	-	-
MH-DETR [54] <small>ACM MM2023</small>	60.05	42.48	60.75	38.13	38.38	38.22	60.51
QD-DETR [28] <small>CVPR2023</small>	62.40	44.98	62.52	39.88	39.86	38.94	62.40
UniVTG [23] <small>ICCV2023</small>	58.86	40.86	57.60	35.59	35.47	38.20	60.96
TR-DETR [44] <small>AAAI2024</small>	64.66	48.96	63.98	43.73	42.62	39.91	63.42
UVCOM [51] <small>CVPR2024</small>	63.55	47.47	63.37	42.67	43.18	39.74	64.20
<b>Ours</b>	<b>65.95</b>	<b>49.74</b>	<b>65.82</b>	<b>44.14</b>	<b>43.57</b>	<b>40.27</b>	<b>65.60</b>

our framework with established baselines in Sec. 4.2 and show a detailed ablation study in Sec. 4.3. Lastly, we show qualitative results showcasing the effectiveness of our approach in Sec. 4.4.

#### 4.1 Experimental Setup

**Datasets.** As our task is to detect highlights while retrieving moments, we use commonly used MR and HD dataset QVHighlights [18] as our main benchmark. The QVHighlights dataset currently stands as the sole dataset available to test both MR and HD tasks concurrently and comprises of over 10,000 YouTube videos accompanied by human-written, free-form text queries. For moment labels, each video-text pair is annotated with one or more relevant moments, and highlight labels are provided with 5-scale saliency scores (ranging from 1 being very bad to 5 being very good). Additionally, in order to ensure a fair benchmark for evaluation, the performance on the test set is assessed exclusively through submissions to the QVHighlights server.<sup>1</sup> In addition to this, to further test the efficacy of our method, we evaluate it on other VTG datasets, namely Charades-STA [8] and TACoS [33]. Charades-STA features 9,848 videos with 16,128 query-moment pairs focusing on indoor activities. TACoS comprises 127 videos annotated specifically for cooking scenarios.

**Evaluation Metrics.** We follow the conventions established in previous research [12, 18, 23, 27, 28, 44, 51, 54]. In MR, we apply Recall@1 (R@1) at IoU thresholds of 0.5 and 0.7. We also calculate mean Average Precision (mAP) for IoU thresholds [0.5 : 0.05 : 0.95]. For HD, we measure mAP and HIT@1, where HIT@1 is determined by the hit ratio of the clip with the highest score.

**Baseline Architecture.** Our framework is built upon the QD-DETR [28] architecture, which is a widely used baseline in Video Temporal Grounding due to its effective use of cross-attention layers for injecting text query information into video frames. We build upon this baseline while retaining its standard cross-attention mechanism, decoder structure, and rank-aware loss. Our method introduces novel approaches aimed to effectively leverage the holistic context of text queries for improved video frame selection and alignment.

<sup>1</sup><https://codalab.lisn.unpsaclay.fr/competitions/6937>

**Implementation Details.** We configure the number of layers in the transformer encoder  $E$  and decoder  $D$  as 3. We set the cross-modal interaction layer count to 2. The loss balancing parameters are set with  $\lambda_{L1} = 10$ ,  $\lambda_{iou} = 1$ , and  $\lambda_{cls} = 4$  adopted from previous works, while  $\lambda_{frame} = 1$  and  $\lambda_{clip} = 1$  are for our proposed loss function, and these parameters are consistent across all datasets. We set the batch size to 32 and the learning rate (LR) to 0.0001 for QVHighlights, maintain a batch size of 32 with an LR of 0.0002 for Charades. We adjust the batch size of 16 with an LR of 0.0002 for TACoS following previous works. Across all datasets, training proceeds for 200 epochs with a learning rate reduction at epoch 100, using the Adam optimizer. Additionally, for all datasets, we set the hidden dimension  $d$  to 256 and the number of moment queries  $M$  to 10. Whereas otherwise stated, we employ a pre-trained SlowFast and CLIP [32] model for video feature extraction. Specifically for Charades-STA, additionally features were extracted using VGG [40], C3D [47], and GloVe [31]. All models were trained on a single NVIDIA RTX 4090 with an average training time of 3 hours for all 200 epochs on our machines.

#### 4.2 Main Result

We compare how our method performs in relation to recent state-of-the-art methods for MR and HD and summarize our findings in the following sections.

**Results on QVHighlights.** In Table 1, we list the experimental results of our method as well as other established methods on the QVHighlights dataset. Our method diverges from XML [19], which adopts a proposal-free strategy, and aligns with the transformer-based and end-to-end trainable nature of current baselines [12, 18, 21, 23, 27, 28, 44, 51, 54]. MH-DETR [54] and QD-DETR [28] focus more on cross-modal interaction before transformer encoder and TR-DETR [44] explores the inherent reciprocity between MR and HD, and UVCOM [51] is tailored to address the unique demands of both MR and HD tasks effectively. Although EaTR [12] is a fairly recent work, we do not list them in our main table as they do not evaluate on the QVHighlights test split. Our model capitalizes on global text semantics and novel loss functions within a refined cross-modal interaction framework, surpasses all compared methods. Notably, our method outperforms the baseline model,

**Table 2: Experimental results on the Charades-STA test split. All models employed uniform features for fairness in evaluation, with ‘SF+C, C’ denoting SlowFast and CLIP for video and CLIP for text, respectively, ‘VGG, GloVe’ indicating VGG features for video and GloVe embeddings for text, and ‘C3D, GloVe’ representing C3D features for video and GloVe embeddings for text.**

Method	feat	R@0.5	R@0.7
Moment-DETR [18] <small>NeurIPS2021</small>	SF+C, C	53.63	31.37
MomentDiff [21] <small>NeurIPS2023</small>	SF+C, C	55.57	32.42
QD-DETR [28] <small>CVPR2023</small>	SF+C, C	57.31	32.55
UniVTG [23] <small>ICCV2023</small>	SF+C, C	58.01	35.65
TR-DETR [44] <small>AAAI2024</small>	SF+C, C	57.61	33.52
UVCOM [51] <small>CVPR 2024</small>	SF+C, C	59.25	36.64
<b>Ours</b>	SF+C, C	<b>60.73</b>	<b>39.49</b>
MAN [58] <small>CVPR2019</small>	VGG, GloVe	41.24	20.54
2D-TAN [60] <small>AAAI2020</small>	VGG, GloVe	40.94	22.85
MomentDiff [21] <small>NeurIPS2022</small>	VGG, GloVe	51.94	28.25
QD-DETR [28] <small>CVPR2023</small>	VGG, GloVe	52.77	31.13
TR-DETR [44] <small>AAAI2024</small>	VGG, GloVe	53.47	30.81
UVCOM [51] <small>CVPR2024</small>	VGG, GloVe	54.57	34.13
<b>Ours</b>	VGG, GloVe	<b>56.56</b>	<b>37.28</b>
IVG-DCL [30] <small>CVPR2021</small>	C3D, GloVe	50.24	32.88
MomentDiff [21] <small>NeurIPS2023</small>	C3D, GloVe	53.79	30.18
QD-DETR [28] <small>CVPR2023</small>	C3D, GloVe	50.67	31.02
<b>Ours</b>	C3D, GloVe	<b>54.78</b>	<b>35.13</b>

**Table 3: Experimental results on the TACoS test split. All models utilized uniform video (SlowFast and CLIP) and text features (CLIP).**

Method	R@0.3	R@0.5	R@0.7	mIoU
2D-TAN [60] <small>AAAI2020</small>	40.01	27.99	12.92	27.22
VSLNet [59] <small>ACL2022</small>	35.54	23.54	13.15	24.99
Moment-DETR [18] <small>NeurIPS2021</small>	37.97	24.67	11.97	25.49
UniVTG [23] <small>ICCV2023</small>	51.44	34.97	17.35	33.60
UVCOM [51] <small>CVPR2024</small>	-	36.39	23.32	-
<b>Ours</b>	<b>52.59</b>	<b>37.89</b>	<b>23.97</b>	<b>36.31</b>

QD-DETR [28], in MR by achieving a 3.5% improvement in R@1 at IoU 0.5, 4.76% improvement at IoU 0.7, and 3.7% increase in average mAP, and in HD by 1.63 mAP and 3.20 in HIT@1. Our method outperforms the most recent method, UVCOM [51], in MR by 2.40% R@0.5, 2.24% R@0.5, and 0.39% average mAP, and in HD by 0.50 mAP and 1.40 HIT@1 respectively. We find that our method outperforms all previous baselines across all metrics in both MR and HD, to the best of our knowledge, setting the new state-of-the-art.

**Results on Charades-STA and TACoS.** To test the generalizability and efficacy of our method, we extend our experimentation to other VTG benchmarks such as Charades-STA [8] and TACoS [33]. On top of our comparisons with transformer-based models, we further test our approach against prior proposal-based approaches [58, 60]. As evidenced in Tables 2 and 3, our method surpasses previous techniques, establishing new state-of-the-arts on these benchmarks. Particularly on Charades-STA, our approach

**Table 4: Ablation study results on QVHighlights val split regarding gated cross-attention mechanisms. The ‘Local’ and ‘Non-local’ columns refer to the use of local and non-local gates, respectively.**

Local	Non-local	MR		HD	
		R1 @0.5	mAP @0.7	≥Very Good Avg.	mAP
✓	✓	63.29	48.45	42.22	39.69
		62.45	48.90	42.5	39.82
		65.87	49.29	43.61	40.79
✓	✓	<b>67.61</b>	<b>50.65</b>	<b>44.8</b>	<b>40.98</b>

demonstrates its robustness by consistently outperforming existing methods across a variety of backbones, including learned multi-modal features from CLIP [32], as well as 2D and 3D features from VGG [40] and C3D [47], respectively. Notably, with features combining SlowFast [7] and CLIP (SF+C, C), our method surpasses the recent state-of-the-art model, UVCOM [51], by achieving improvements of 1.48% in R@1 at an IoU of 0.5 and 2.85% at an IoU of 0.7. On TACoS, our method once again outperforms all previous baselines, affirming its effectiveness across diverse domains and datasets.

### 4.3 Ablation Studies

To understand the individual components of our framework and its effects, we present a series of ablation studies on QVHighlights validation split.

**Gated Cross-Attention.** We analyse the significance of employing both local and non-local gates within our gated cross-attention framework and present it in Table 4. Implementing the non-local gate alone enhances performance in MR and HD tasks, underscoring its utility. The synergy of local and non-local gates, however, yields the best outcome, underlining their collective importance in refining video-text alignment and enhancing prediction accuracy.

To further assess the validity of our non-local gate weight ( $g_N$ ) within the gated cross-attention framework, we test to see if it can be used directly as a saliency score for Highlight Detection on the QVHighlights validation split. Table 5 demonstrates that using  $g_N$  directly to predict saliency not only exceeds Moment-DETR’s HD performance but also shows the utility of our gate mechanism in emphasizing relevant video sections in cross-modal interaction. Additionally, ‘ $g_N$  w/o Non-local,’ represents a scenario where  $g_N$  is computed without applying the non-local gate, which corresponds to the the second row of Table 4, and this results in a significant drop in performance. This underlines the non-local gate’s critical role in emphasizing relevant video frame and enhancing the accuracy of the similarity measure between global text and video frames.

**Fine-Grained Alignment Losses.** We further analyse the impact of the frame-level and clip-level alignment losses in Table 6. Implementing each loss individually offers noticeable improvements; however, integrating both simultaneously provides a substantial performance boost. Specifically, without these alignment losses, the model achieves lower scores across all metrics. With both losses applied, we observe a 5.29% increase in R@1 at IoU 0.5, a 3.3% rise

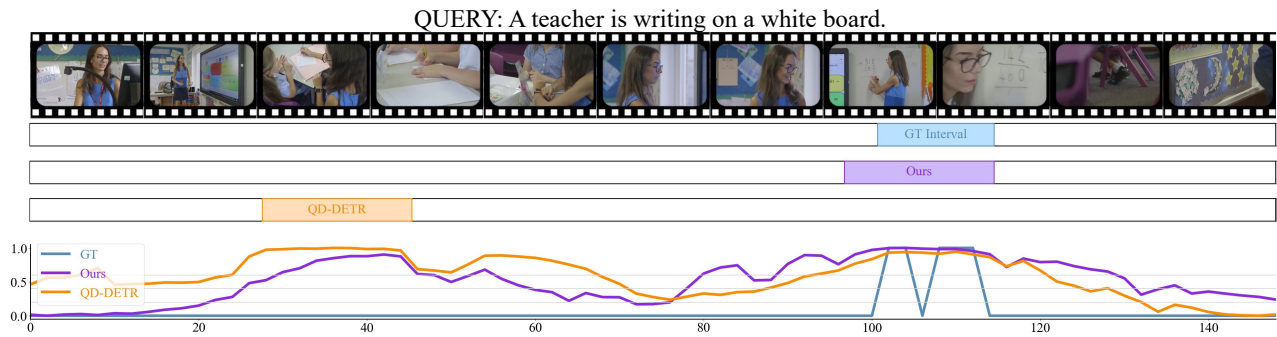


Figure 3: Qualitative results of predictions on QVHighlights validation split. We show the effectiveness of our method compared to the baseline, QD-DETR. From top to bottom are the text queries, along with the predicted moments and highlights corresponding to each method.

Table 5: Highlight Detection performance on QVHighlights val split, showcasing the validity of non-local gate weights in gated cross-attention. † denotes intermediate outputs assessed directly for HD.

Method	HD	
	$\geq$ Very Good mAP	HIT@1
$g_N$ w/o Non-local†	25.31	41.16
$g_N^\dagger$	35.83	56.71
Moment-DETR [18]	35.69	55.60
<b>Ours</b>	<b>40.98</b>	<b>65.35</b>

Table 6: Ablation study results on QVHighlights val split on fine-grained alignment loss. The ‘Frame’ and ‘Clip’ columns denote the frame-level consistency loss and clip-level relevance loss, respectively.

Frame	Clip	MR		HD	
		R1 @0.5	@0.7	mAP Avg.	$\geq$ Very Good mAP
		62.32	47.35	41.92	39.02
	✓	65.48	49.23	43.20	40.46
✓		63.48	48.13	43.43	39.56
✓	✓	<b>67.61</b>	<b>50.65</b>	<b>44.80</b>	<b>40.98</b>

at IoU 0.7, and a 2.88% improvement in mAP, highlighting their effectiveness in refining the model’s capability to align video content with textual queries accurately.

**Global Text Anchor.** We also explore various methods for generating the global text anchor, which is pivotal to our framework as it focuses attention on the relevant parts of the video corresponding to the text query. To determine the most effective approach, as detailed in Table 7, we compared mean pooling, max pooling, weighted pooling, and the use of a transformer layer. We find that mean pooling outperforms other methods. This underscores the effectiveness of a simple yet powerful mean pooling strategy in capturing the holistic semantics of the text query for VTG.

Table 7: Ablation study on various global text anchor generation methods on QVHighlights val split.

Method	MR		HD	
	R1 @0.5	@0.7	mAP Avg.	$\geq$ Very Good mAP
Max pooling	64.45	49.55	44.07	40.48
Weighted pooling	65.87	49.29	43.61	40.79
Transformer	65.68	50.06	44.24	40.30
Mean pooling	<b>67.61</b>	<b>50.65</b>	<b>44.80</b>	<b>40.98</b>

#### 4.4 Qualitative Results

We show qualitative results in Fig. 3 compared with the baseline model, QD-DETR. By adopting a holistic approach to understanding text queries, our method consistently identifies moments that fully align with the intent of the textual queries. This approach allows our model to capture the essence of the entire query, preventing the oversight of integral query components such as ‘a white board’, which QD-DETR sometimes overlooks. This underscores the advantage of our method’s integration of global text semantics for more accurate video grounding.

## 5 CONCLUSION

In this work, we present a novel approach to Video Temporal Grounding (VTG) that emphasizes holistic understanding of text queries and suppresses irrelevant visual frames. By integrating the entire query text into a global representation and employing visual frame-level gate mechanisms within a cross-modal interaction framework, our approach significantly enhances the alignment of text queries with accurate video segments. We demonstrate state-of-the-art performance on several VTG benchmarks, highlighting the importance of considering the entire text query and selectively focusing on relevant video.

## ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00212845)



## REFERENCES

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proc. ICCV*.
- [2] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. 2022. Locvtv: Video-text pre-training for temporal localization. In *Proc. ECCV*.
- [3] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proc. EMNLP*.
- [4] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. 2021. End-to-end multi-modal video temporal grounding. In *NeurIPS*.
- [5] Jae Won Cho, Dawit Mureja Argaw, Youngtaek Oh, Dong-Jin Kim, and In So Kweon. 2023. Empirical study on using adapters for debiased Visual Question Answering. *Computer Vision and Image Understanding* (2023).
- [6] Jae Won Cho, Dong-Jin Kim, Hyeonggon Ryu, and In So Kweon. 2023. Generative Bias for Robust Visual Question Answering. In *Proc. CVPR*.
- [7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slow-fast networks for video recognition. In *Proc. ICCV*.
- [8] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proc. ICCV*.
- [9] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Proc. ACL*.
- [10] Ning Han, Jingjing Chen, Guangyi Xiao, Hao Zhang, Yawen Zeng, and Hao Chen. 2021. Fine-grained cross-modal alignment network for text-video retrieval. In *Proc. ACM MM*.
- [11] David T Hoffmann, Nadine Behrmann, Juergen Gall, Thomas Brox, and Mehdi Noroozi. 2022. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In *Proc. AAAI*.
- [12] Jinyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. 2023. Knowing where to focus: Event-aware transformer for video grounding. In *Proc. ICCV*.
- [13] Youngjoon Jang, Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, Joon Son Chung, and In So Kweon. 2022. Signing Outside the Studio: Benchmarking Background Robustness for Continuous Sign Language Recognition. In *Proc. BMVC*.
- [14] Youngjoon Jang, Youngtaek Oh, Jae Won Cho, Myungchul Kim, Dong-Jin Kim, In So Kweon, and Joon Son Chung. 2023. Self-Sufficient Framework for Continuous Sign Language Recognition. In *Proc. ICASSP*.
- [15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. ICLR*.
- [16] Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. 2023. Shatter and Gather: Learning Referring Image Segmentation with Text Supervision. In *Proc. ICCV*.
- [17] Dong-Jin Kim, Jae Won Cho, Jinsoo Choi, Yunjae Jung, and In So Kweon. 2021. Single-Modal Entropy based Active Learning for Visual Question Answering. In *Proc. BMVC*.
- [18] Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*.
- [19] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Proc. ECCV*.
- [20] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.
- [21] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. 2023. Momentdiff: Generative video moment retrieval from random to real. In *NeurIPS*.
- [22] Sizhe Li, Chang Li, Minghang Zheng, and Yang Liu. 2022. Phrase-level prediction for video temporal localization. In *Proc. ICMR*.
- [23] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. Univtq: Towards unified video-language temporal grounding. In *Proc. ICCV*.
- [24] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *Proc. ACM MM*.
- [25] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. 2022. Dab-detr: Dynamic anchor boxes are better queries for detr. In *Proc. ICLR*.
- [26] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proc. CVPR*.
- [27] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proc. CVPR*.
- [28] Wonjun Moon, Sangeek Hyun, Sanguk Park, Dongchan Park, and Jae-Pil Heo. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *Proc. CVPR*.
- [29] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proc. CVPR*.
- [30] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional video grounding with dual contrastive learning. In *Proc. CVPR*.
- [31] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. EMNLP*.
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICLR*.
- [33] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. In *Proc. ACL*.
- [34] Hamid Rezaatofghi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proc. CVPR*.
- [35] Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, and Stephen Gould. 2021. DORI: Discovering object relationships for moment localization of a natural language query in a video. In *Proc. WACV*.
- [36] Yong Rui, Anoop Gupta, and Alex Acero. 2000. Automatically extracting highlights for TV baseball programs. In *Proc. ACM MM*.
- [37] Hyeonggon Ryu, Arda Senocak, In So Kweon, and Joon Son Chung. 2023. Hindi as a second language: Improving visually grounded speech with semantically similar samples. In *Proc. ICASSP*.
- [38] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. 2023. Sound source localization is all about cross-modal alignment. In *Proc. ICCV*.
- [39] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. 2018. Find and focus: Retrieve and localize video events with natural language queries. In *Proc. ECCV*.
- [40] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*.
- [41] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. 2021. Vlg-net: Video-language graph matching network for video grounding. In *Proc. ICCV*.
- [42] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proc. CVPR*.
- [43] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. 2021. Segmenter: Transformer for semantic segmentation. In *Proc. CVPR*.
- [44] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. 2024. TR-DETR: Task-Reciprocal Transformer for Joint Moment Retrieval and Highlight Detection. In *Proc. AAAI*.
- [45] Min Sun, Ali Farhadi, and Steve Seitz. 2014. Ranking domain-specific highlights by analyzing edited videos. In *Proc. ECCV*.
- [46] Min Sun, Ali Farhadi, and Steve Seitz. 2014. Ranking domain-specific highlights by analyzing edited videos. In *Proc. ECCV*.
- [47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proc. ICCV*.
- [48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. [n. d.]. Attention is all you need. In *NeurIPS*.
- [49] Xiaohan Wang, Linchao Zhu, and Yi Yang. 2021. T2vлад: global-local sequence alignment for text-video retrieval. In *Proc. CVPR*.
- [50] Jongbin Woo, Hyeonggon Ryu, Arda Senocak, and Joon Son Chung. 2024. Speech Guided Masked Image Modeling for Visually Grounded Speech. In *Proc. ICASSP*.
- [51] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. 2024. Bridging the Gap: A Unified Video Comprehension Framework for Moment Retrieval and Highlight Detection. In *Proc. CVPR*.
- [52] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metzke, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proc. ACL*.
- [53] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proc. AAAI*.
- [54] Yifang Xu, Yunzhuo Sun, Yang Li, Yilei Shi, Xiaoxiang Zhu, and Sidan Du. 2023. Mh-detr: Video moment and highlight detection with cross-modal transformer. In *Proc. ACM MM*.
- [55] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NeurIPS*, Vol. 32.
- [56] Yitian Yuan, Lin Ma, and Wenwu Zhu. 2019. Sentence specified dynamic video thumbnail generation. In *Proc. ACM MM*.
- [57] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proc. AAAI*.
- [58] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proc. CVPR*.

- [59] Hao Zhang, Aixun Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based Localizing Network for Natural Language Video Localization. In *Proc. ACL*.
- [60] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proc. AAAI*.
- [61] Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019. Exploiting temporal relationships in video moment localization with natural language. In *Proc. ACM MM*.
- [62] Minghang Zheng, Sizhe Li, Qingchao Chen, Yuxin Peng, and Yang Liu. 2023. Phrase-level temporal relationship mining for temporal sentence localization. In *Proc. AAAI*.