DE^{2}TR: Dual Evidence Detection Transformer for Video Temporal Grounding

Yifan Zhang[1], Chengxu Liu[2], Yujie Dun[1(✉)], and Xueming Qian[1]

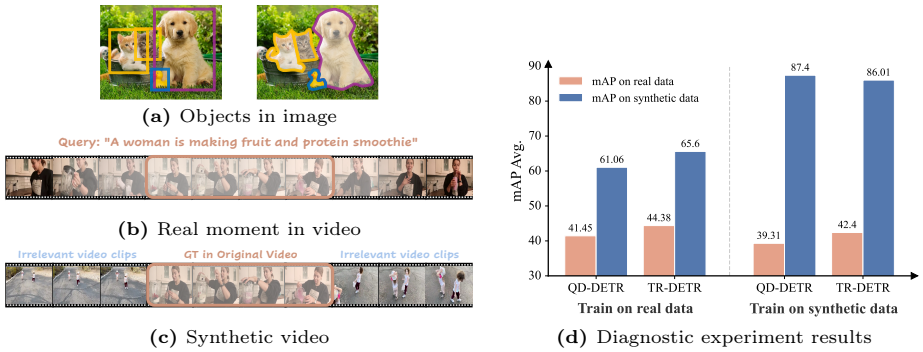
[1] School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an, China
[2] School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
ivanz@stu.xjtu.edu.cn, chengxu.liu@xjtu.edu.cn, {dunyj,qianxm}@mail.xjtu.edu.cn

Abstract. Video temporal grounding (VTG) aims to localize relevant moments and predict saliency scores in untrimmed videos based on natural language queries. Recently, DETR-inspired methods have achieved instance-level moment predictions through query-based semantic evidence aggregation. However, unlike spatial objects with clear physical edges, video moments often exhibit ambiguous temporal boundaries due to the semantic gradient. Consequently, relying solely on global semantics can lead to severe boundary misalignment, yielding sub-optimal grounding results. To address this issue, we propose a novel Dual Evidence Detection Transformer (DE^{2}TR) framework for better boundary localization. In particular, we introduce a dual-branch decoder that simultaneously captures global semantic evidence and fine-grained boundary cues. To prevent feature conflict between these distinct objectives, we channel-wise decouple the moment queries, facilitating independent evidence learning under the supervision of evidence alignment losses. Building upon this dual-branch design, we introduce a prior-guided refinement mechanism that employs explicit prior maps to steer queries toward potential regions of interest. Furthermore, a boundary augmentation strategy synthesizes ideal boundary patterns via simple temporal splicing, effectively enhancing the model's boundary awareness. Extensive experiments on multiple popular benchmarks demonstrate that DE^{2}TR significantly outperforms state-of-the-art approaches.

Keywords: Video temporal grounding · Moment retrieval · Detection transformer

1 Introduction

The explosive growth of online video data and users' rising preference for short-form content present significant challenges for efficient video retrieval and browsing [11,56]. Driven by this growing demand, video temporal grounding (VTG) [8, 24, 29, 52, 62] based on natural language queries has recently garnered extensive attention. This task encompasses moment retrieval (MR) [1,76] and highlight detection (HD) [21,61], aiming to localize text-relevant moments in an untrimmed video and predict clip-level saliency scores to pinpoint highlight frames.



(a) Objects in image

(b) Real moment in video

(c) Synthetic video

(d) Diagnostic experiment results

Fig. 1: Difference between objects in image and moments in video. (**a**) Objects in images have clear physical edges. (**b**) Moments in video exhibit semantic gradient, with ambiguous boundaries. (**c**) Synthetic videos consist of the original Ground Truth (GT) and irrelevant clips as background, which provide clear boundary signal as priors. (**d**) Diagnostic experiments with QD-DETR [40] and TR-DETR [51]. Models trained on both datasets perform excellently when provided with clear boundary signals as prior knowledge. Meanwhile, models trained on synthetic data still exhibit baseline-level performance on real data, highlighting the importance of boundary cues.

In recent years, numerous efforts have been devoted to addressing VTG [7, 40, 58, 62, 71]. In particular, inspired by the success of DETR [5] in object detection, Lei *et al*. proposed the Moment-DETR [24] framework. By utilizing a set of learnable decoder queries to directly predict target moments, it effectively circumvents complex components such as proposal generation [49] and non-maximum suppression (NMS) [46] found in traditional methods [1, 7, 30]. Recent works mainly focus on improving the paradigm to achieve more precise localization performance, centered on modal fusion and alignment [28,35,40,51], inter-task relationships [51, 62, 67], and query initialization [17, 35, 52].

However, existing methods follow DETR to predict moment spans at the instance level (*i.e*. using decoder queries to collect global semantic evidence for locating target moments), ignoring the difference between objects in image and moments in video. Specifically, objects in images typically possess clear physical edges (Fig. 1a). In contrast, temporal moments often exhibit ambiguous boundaries due to semantic gradient (Fig. 1b). Although the global semantic evidence helps identifying the clips most relevant to the text, the absence of boundary cues causes severe boundary misalignment, leading to sub-optimal localization. To validate our insights, we introduce hard boundary signals as prior knowledge for video moments via a simple splicing operation (Fig. 1c) and conduct diagnostic experiments on both real and synthetic data (Fig. 1d). The significant performance gap between synthetic and real-world videos exposes the vulnerability of existing methods to ambiguous boundaries, highlighting the importance of explicitly modeling boundary cues.

To bridge this gap, we propose a Dual Evidence Detection Transformer (DE^{2}TR) framework that simultaneously focuses on semantic alignment and

boundary cues for better boundary localization. In particular, we introduce a dual-branch transformer decoder, named DEFormer, to capture both types of evidence. Specifically, a semantic evidence branch captures the global semantic representation to predict the foreground score, while a boundary evidence branch focuses on fine-grained structural cues to refine moment boundaries. Since capturing distinct evidence with shared queries leads to feature conflicts, DEFormer decouples the moment queries channel-wise to facilitate independent evidence learning. To effectively supervise this process, an evidence alignment loss is further introduced to explicitly align these decoupled queries with global moments and local boundaries. Building upon the dual-branch structure, we introduce two efficient mechanisms to further enhance the decoding process. First, we design a prior-guided refinement mechanism to provide prior knowledge for the decoupled queries. By exploiting the dual nature of saliency, we learn branch-specific prior maps that direct the model to focus on potential target regions. Additionally, we propose a boundary augmentation strategy (BAS) that augments samples via a simple temporal splicing operation. This encourages the model to derive ideal boundary responses amidst ambiguous ones, effectively mitigating the boundary uncertainty caused by semantic gradients.

In summary, our contributions are as follows:

- We review the limitations of DETR-based methods in boundary misalignment, and propose a novel Dual Evidence Detection Transformer (DE^{2}TR) framework for video representation learning.
- We introduce a dual-branch transformer decoder, named DEFormer, to capture global semantic evidence and fine-grained boundary cues separately, which avoids feature conflicts.
- We design a branch-specific prior-guided refinement mechanism and a boundary augmentation strategy to provide explicit structural guidance and ideal boundary patterns for DEFormer.
- Extensive experiments on four challenging benchmarks show that DE^{2}TR surpasses previous state-of-the-art methods on VTG tasks.

2 Related Works

2.1 Video Temporal Grounding

VTG is a fine-grained video understanding task, including MR and HD, which were initially studied in isolation. MR aims to locate relevant moments in videos based on given text queries. Early work typically generates proposals using methods such as sliding windows [1, 12, 31, 69] and temporal anchors [7, 72, 76], which are then ranked based on their similarity to the text. Proposal-free approaches [54, 65, 70, 74, 75] further eliminate proposal generation by directly regressing boundaries, achieving improved efficiency. HD focuses on estimating the saliency of video clips. While traditionally treated as a uni-modal task [2, 61, 64], recent advancements in multi-modal learning and the increasing demand for personalized content have shifted attention toward query-dependent HD [21, 41].

Recognizing the correlation between MR and HD, recent research has gravitated towards unified frameworks for joint optimization. Moment-DETR [24] pioneered this direction by introducing the query-based paradigm and the QVHighlights dataset, aiming to simultaneously output moment spans and saliency scores within a unified framework. Building upon this, UMT [35] extended the framework to accommodate multi-modal inputs. To enhance query discriminability, QD-DETR [40] introduced query-dependent video representations via cross-attention and negative pair learning. Furthermore, TR-DETR [51] explored task reciprocity by injecting highlight scores into retrieval pipeline and utilizing retrieved moments to refine highlight predictions. TD-DETR [79] proposed a dynamic learning approach that combined video synthesis with dynamic text interaction to alleviate over-association with background frames. Despite remarkable progress, the over-reliance on global semantic alignment in existing query-based methods often leads to sub-optimal localization. In this work, we extend this paradigm by integrating dual evidence for more precise boundary localization.

2.2 Query-based Detection Transformer

DETR [5] revolutionized object detection by formulating it as an end-to-end set prediction problem via learnable queries. Since its inception, numerous variants have emerged to enhance this architecture. Some improvements focus on accelerating training convergence and computational efficiency [47, 57, 78, 80]. Others concentrate on extending the DETR paradigm to broader vision-language tasks through various multi-modal fusion strategies, such as visual grounding [19, 33], video object segmentation [3] and spatio-temporal video grounding [66, 77]. Furthermore, a crucial line of research focuses on optimizing the decoder queries. For instance, several methods incorporate explicit spatial priors into the query for better performance [32, 38, 60], while others introduce query denoising tasks to stabilize the bipartite matching process [25, 73].

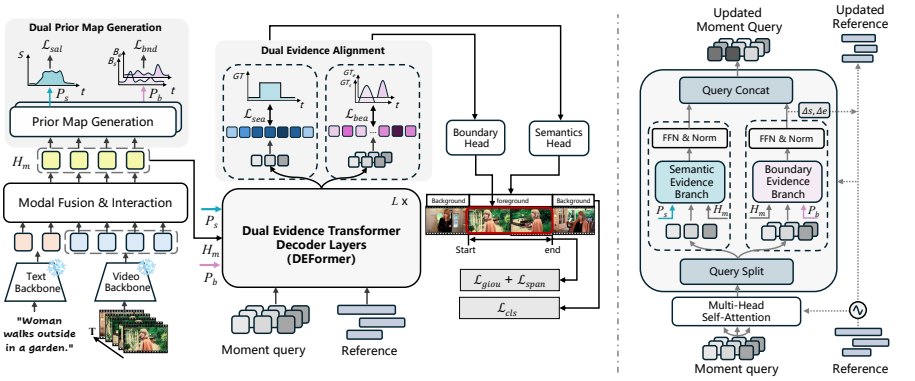
In this work, we also focus on the decoder query. However, different from the previous methods, we concentrate on the evidence decoupling learning of moment queries to mitigate the inherent boundary ambiguity in videos.

3 Method

3.1 Overview

Given a text query with L_t words and an untrimmed video with L_v clips, the goal of VTG is to predict the start and end timestamps of the relevant moments $\mathcal{M} = \{(t_k^s, t_k^e)|_{k=1}^K\}$ in the video, while providing clip-wise saliency scores $\mathcal{S} = \{s_1, s_2, \cdots, s_{L_v}\}$ of the text to select the most highlights frames. Here, t_k^s and t_k^e denote the start and end timestamps of the k-th moment, respectively. K represents the number of moments in the video, typically $K \geq 1$.

As illustrated in Fig. 2, we present the DE^{2}TR, a novel framework designed to disentangle semantic and boundary perception for precise moment localization. DE^{2}TR follows the encoder-decoder pipeline. Taking text features and



(a) Overall framework of DE^2 TR **(b)** Illustration of DEFormer layer

Fig. 2: (a) Overview of our proposed DE^{2}TR framework. DE^{2}TR first performs modal fusion and interaction to generate saliency and boundary scores as structural priors. Guided by these priors, the DEFormer explicitly decouples the probing of semantic and boundary evidence, supervised by dual evidence alignment losses. Finally, the target moments are predicted via the prediction heads. **(b)** Illustration of the DEFormer layer. It features a semantic evidence branch to aggregate salient visual context, and a parallel boundary evidence branch dedicated to capturing fine-grained cues, iteratively refining the temporal reference anchors across consecutive decoding stages.

video features as input, the model first employs modal fusion and interaction to produce text-aware video representations. These representations are then used to generate saliency and boundary scores as structural prior maps(Sec. 3.2). Subsequently, DEFormer aggregates dual evidence assisted by the prior-guided refinement mechanism and evidence alignment losses(Sec. 3.3). Finally, aggregated evidence is fed to prediction heads to predict the target moments.

3.2 Dual Prior Map Generation

Following standard protocols [24, 40], we utilize frozen backbones (e.g. Slowfast [10], CLIP [42]) to extract visual and textual features, projecting them into a shared d-dimensional space. To align the modalities, we employ a modal fusion and interaction pipeline [40, 51, 52] to generate the text-aware video representation $H_m \in \mathbb{R}^{L_v \times d}$, which serves as the decoder memory. More details about the pipeline can be found in Appendix. Typically, standard DETR-like methods utilize such fused representations to predict 1D saliency scores for locating highlight frames [17, 24, 55, 79]. Building upon this paradigm, we uncover a hidden dual nature within the saliency scores, *i.e.* high values indicate strong semantic alignment, whereas sharp drops suggest potential boundaries. Motivated by this, we extract dual structural prior maps from H_m to provide prior knowledge for the subsequent dual-branch decoding.

Semantic Prior Map (P_s). Since the saliency scores directly measure semantic relations, we follow standard protocols to produce saliency scores as our

semantic prior map $P_s \in \mathbb{R}^{L_v}$. Specifically, we use a saliency prediction head to estimate the textual relevance of each clip from the memory H_m. The standard saliency loss $\mathcal{L}_{sal}$ [40,53,62] is applied to assist the model better understand the semantic relationship between the video and text.

Boundary Prior Map (P_b). Although the temporal gradients of saliency scores indicate potential boundaries, directly deriving P_b from P_s via element-wise difference makes it vulnerable to prediction noise. Instead, we employ a separate boundary projection, composed of two linear layers with ReLU [14] activation, to project the H_m into start and end probabilities $P_b^s, P_b^e \in \mathbb{R}^{L_v}$.

To avoid overfitting to the annotation noise caused by hard boundary labels [15, 63], we propose an adaptive Gaussian labeling strategy to provide ambiguity-aware soft supervision. Consider a video with a set of GT moments $\mathcal{M} = \{(t_k^s, t_k^e)|_{k=1}^K\}$, we explicitly model each boundary timestamp in $\mathcal{M}$ as a probabilistic Gaussian distribution rather than a deterministic point. Taking start boundary as an example, for the k-th GT moment, we first generate an instance-level soft label sequence $y_k^s \in [0,1]^{L_v}$. Specifically, the target value at the i-th clip is defined as:

$$y_{k,i}^s = \exp\left(-\frac{(i - t_k^s)^2}{2\sigma_k^2}\right), \quad i \in [1, L_v]. \tag{1}$$

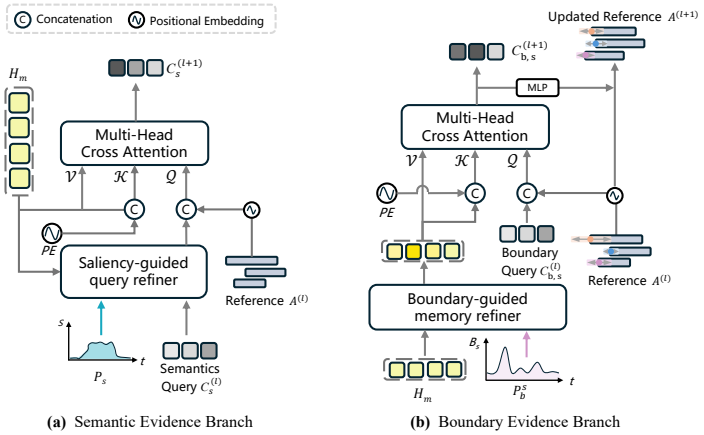
Here, the bandwidth is defined as $\sigma_k = \delta_k \cdot \sigma_0$, where the base scale σ_0 is determined by the IoU-based radius rule from keypoint-based object detection [9,22]. The coefficient $\delta_k = \exp(\cos(\mathbf{v}_{in}, \mathbf{v}_{out}))$ quantifies the semantic ambiguity to dynamically modulate the bandwidth, where $\mathbf{v}_{in}$ and $\mathbf{v}_{out}$ denote the averaged video features inside and outside the boundary, respectively. More details about the Gaussian labeling strategy are provided in the Appendix. This design naturally imposes strict supervision ($\sigma_k \downarrow$) on sharp boundaries while relaxing constraints ($\sigma_k \uparrow$) for ambiguous ones. To supervise the boundary prior map, we generate video-level soft target $Y^s \in [0,1]^{L_v}$ by aggregating Gaussian distributions from all K moments, *i.e.* $Y_i^s = \max_{k=1}^K y_{k,i}^s, i \in [1, L_v]$. The boundary regular loss is then formulated as:

$$\mathcal{L}_{bnd}^s = -\frac{1}{L_v}\sum_{i=1}^{L_v}\left[Y_i^s \log P_{b,i}^s + (1 - Y_i^s)\log(1 - P_{b,i}^s)\right]. \tag{2}$$

where $P_{b,i}^s$ denotes the predicted start probability at the i-th clip. The total boundary loss is the summation of both boundary: $\mathcal{L}_{bnd} = \mathcal{L}_{bnd}^s + \mathcal{L}_{bnd}^e$.

3.3 Dual Evidence Transformer

As illustrated in Fig. 2b, DEFormer comprises a semantic evidence branch and a boundary evidence branch for independent evidence aggregation. Specifically, we employ a set of learnable content embedding $C^{(l)} \in \mathbb{R}^{N \times d}$ as moment queries, and maintain a set of reference anchors $A^{(l)} = \{m_s, m_e\} \in \mathbb{R}^{N \times 2}$ across the decoder layers. In each layer, to eliminate redundant predictions, $C^{(l)}$ is first



(a) Semantic Evidence Branch **(b)** Boundary Evidence Branch

Fig. 3: Illustration of the (a) Semantic and (b) Boundary Evidence Branches.

fed into a multi-head self-attention module to enable inter-query communication [24, 32]. Subsequently, to facilitate dual-evidence learning, the queries are split channel-wise into semantic queries $C_s^{(l)} \in \mathbb{R}^{N\times d/2}$ and boundary queries $C_{b,s}^{(l)}, C_{b,e}^{(l)} \in \mathbb{R}^{N\times d/4}$. They aggregate both types of evidence from H_m under the guidance of prior maps. Through continuous evidence aggregation, the reference anchors are layer-wisely revised to yield the final predicted moments.

Semantic Evidence Branch. As illustrated in Fig. 3a, the semantic evidence branch aims to aggregate global representations from the decoder memory to estimate the foreground confidence of the predicted moment. To guide queries toward semantically salient segments within video sequences, we introduce a Saliency-guided Query Refiner to enhance queries with prior knowledge. Specifically, guided by the semantic prior map P_s, we first perform attention pooling on the decoder memory H_m to extract the global semantic representation H_s that condenses high-confidence clips, *i.e.* $H_s = \text{Linear}\left(\sum \text{Softmax}(P_s) \odot H_m\right)$. Subsequently, a gated residual fusion operation is employed to adaptively inject this global semantic prior into the semantic queries $C_s^{(l)}$:

$$\begin{aligned} \widehat{g} &= \text{sigmoid}\left(C_s^{(l)} \odot H_s\right), \\ \widehat{C}_s^{(l)} &= \text{Linear}\left(\widehat{g} \odot H_s\right) + C_s^{(l)} \end{aligned} \tag{3}$$

where $\widehat{g} \in \mathbb{R}^{N\times d/2}$ represents the gate value, $\odot$ denotes the element-wise multiplication. Then, a standard cross-attention layer is utilized to aggregate the semantic evidence. To endow the refined queries $\widehat{C}_s^{(l)}$ with positional awareness, we apply sinusoidal positional encoding to the reference anchor $A^{(l)}$ to obtain $P_{C_s}^{(l)} = \text{PE}(c, w)$, where $c = (m_e + m_s)/2$ and $w = m_e - m_s$ denote the center and width of the moment, respectively. The position embedding is explicitly decoupled from the content by concatenating them along the channel dimen-

sion [32, 38]. The queries, keys, and values are obtained via linear projections:

$$\mathcal{Q}^{(l)} = [\widehat{C}_s^{(l)} W_q^{(l)} \parallel P_{C_s}^{(l)} W_{P_c}^{(l)}], \quad \mathcal{K}^{(l)} = [H_m W_k^{(l)} \parallel P_k W_{P_k}^{(l)}], \quad \mathcal{V}^{(l)} = H_m W_v^{(l)} \tag{4}$$

where $\parallel$ denotes channel-wise concatenation, and $W_{(\cdot)}$ are learnable projection matrices. Subsequently, the context-aggregated queries $\widetilde{C}_s^{(l)}$ are derived via the standard attention mechanism:

$$\widetilde{C}_s^{(l)} = \text{Softmax}\left(\frac{\mathcal{Q}^{(l)} {\mathcal{K}^{(l)}}^\top}{\sqrt{2d}}\right)\mathcal{V}^{(l)} + \widehat{C}_s^{(l)}. \tag{5}$$

Finally, a feed-forward network (FFN) is applied to yield the completely updated semantic queries $\widetilde{C}_s^{(l+1)}$. To ensure that semantic queries can fully capture the global context of the target moment, we draw inspiration from the differentiable IoU loss [16, 43] in image segmentation and introduce a semantic evidence alignment loss $\mathcal{L}_{sea}$ to supervise this process. Specifically, we compute the cosine similarities between the semantic queries $\widetilde{C}_s^{(l+1)}$ and the memory H_m, and project them into a normalized score matrix $\hat{m} \in [0,1]^{N \times L_v}$. Our objective is to maximize their semantic similarities with all foreground frames within the GT span, while strictly penalizing any spurious responses to out-of-span background frames. The $\mathcal{L}_{sea}$ is defined as:

$$\mathcal{L}_{sea} = -\frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{j=1}^{L_v} \hat{m}_{i,j} \cdot m_{i,j}}{\sum_{j=1}^{L_v} \hat{m}_{i,j}(1 - m_{i,j}) + \sum_{j=1}^{L_v} m_{i,j}}, \tag{6}$$

where N denotes the number of queries, and $\hat{m}_{i,j}$ represents the cosine similarity between the i-th query and the j-th video clip. $m_{i,j}$ serves as a binary indicator denoting whether the j-th video clip is a text-relevant foreground for the i-th query, *i.e.* $m_{i,j} = 1$ if relevant, and 0 otherwise. By minimizing $\mathcal{L}_{sea}$, we force the semantic queries to strictly align with the internal semantics of the moment while effectively suppressing external noise.

Boundary Evidence Branch. Unlike global semantics, boundary evidence is typically transient and highly localized [27, 48]. Therefore, this branch focuses on fine-grained cues to refine the boundaries. To achieve this, as shown in Fig. 3b, we first introduce a Boundary-guided memory refiner designed to enhance the discriminative feature response of potential boundaries within the video memory. Taking the start boundary as an example:

$$\widehat{H}_m = H_m \odot (1 + \gamma^s) + \beta^s \tag{7}$$

where $\gamma^s, \beta^s \in \mathbb{R}^d$ are adaptive scaling and shifting factors derived from the start boundary prior map P_b^s via a two layer MLP. Similar to the semantic branch, the boundary evidence branch then utilizes a cross-attention layer to aggregate fine-grained boundary cues from the modulated video memory:

$$C_{b,s}^{(l+1)} = \text{FFN}(\text{CrossAttn}(C_{b,s}^{(l)}, P_{C_{b,s}}^{(l)}, \widehat{H}_m)). \tag{8}$$

Here, the position embedding is defined as $P^{(l)}_{C_{b,s}} = \mathrm{PE}(m_s, w)$ to explicitly constrain the query's focus to the local temporal neighborhood of the boundary. Finally, the updated queries are fed into an MLP to predict the boundary offsets Δs [80], which are used to iteratively refine the reference anchors layer by layer. To ensure that the boundary queries precisely capture the fine-grained boundary cues, we introduce the boundary evidence alignment loss, denoted as $\mathcal{L}_{bea}$. Similar to the semantic evidence branch, we compute the cosine similarity between $C^{(l+1)}_{b,s}$ and H_m to derive the predicted similarity score matrix $\hat{b}^s$. Given the corresponding GT start boundary labels $b^s \in [0,1]^{N \times L_v}$, which are dynamically assigned from the instance-level soft labels y^s (generated in Sec. 3.2) via bipartite matching, the alignment loss $\mathcal{L}^s_{bea}$ is formulated as:

$$\mathcal{L}^s_{bea} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L_v}\left[b^s_{i,j}\log\hat{b}^s_{i,j} + (1-b^s_{i,j})\log(1-\hat{b}^s_{i,j})\right], \tag{9}$$

where $b^s_{i,j}$ and $\hat{b}^s_{i,j}$ represent the target soft label and the predicted boundary probability for the i-th query at the j-th clip, respectively. The total boundary evidence alignment loss is given by: $\mathcal{L}_{bea} = \mathcal{L}^s_{bea} + \mathcal{L}^e_{bea}$.

Finally, the decoupled sub-queries are concatenated channel-wise at the end of each layer to reconstruct a unified instance query. After multi-layer evidence aggregation and alignment, the specific sub-queries from the last layer are routed to their respective heads to decode the temporal coordinates $(\hat{t}^s, \hat{t}^e)$ and foreground scores $\hat{f}$, yielding the final prediction set $\hat{\Psi} = \{\hat{\psi}_n = (\hat{t}^s_n, \hat{t}^e_n, \hat{f}_n)\}_{n=1}^{N}$.

3.4 Training

Boundary Augmentation Strategy. To overcome the inherent boundary ambiguity, we incorporate BAS to augment the videos. Following the process in Fig. 1c, for a given sample, we retain its GT moments and replace the contextual background with clips from randomly sampled videos. During training, the model learns these ideal boundary responses and effectively transfers this boundary awareness to real-world videos. More details are provided in Appendix.

Loss Functions. Following the Hungarian matching [24] between predictions $\hat{\Psi}$ and ground truth $\mathcal{M}$, the loss function incorporates $\mathcal{L}_{MR}$ and $\mathcal{L}_{HD}$ as in [24, 62, 79]. Besides, the boundary prior map is optimized by $\mathcal{L}_{bnd}$, while the two branches of DEFormer are constrained by $\mathcal{L}_{sea}$ and $\mathcal{L}_{bea}$, respectively. The overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{MR} + \mathcal{L}_{HD} + \lambda_{bnd}\mathcal{L}_{bnd} + \lambda_{sea}\mathcal{L}_{sea} + \lambda_{bea}\mathcal{L}_{bea} \tag{10}$$

where $\lambda_{(\cdot)}$ are the balancing factors.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate our method on four widely used benchmarks: **QVHighlights** [24], **Charades-STA** [12], **TACoS** [45], and **TVSum** [50]. QVHighlights

is a benchmark for joint MR and HD, including 10,148 YouTube videos covering diverse topics. It contains 10,310 text queries and 18,367 different moment annotations, with most queries corresponding to multiple moments. Charades-STA focuses on daily indoor activities with 16,128 query-moment pairs. Following prior work [40, 51], we use standard split of 12,408 samples for training and 3,720 for testing. TACoS contains 127 cooking videos with an average length of 4.5 minutes. TVSum is an HD benchmark consisting of 50 videos with dense frame-level saliency scores, which we use to evaluate the HD generalization.

Evaluation Metrics. Following standard protocols [24, 40, 53], we adopt task-specific metrics for fair comparison. For the MR task, we report Recall@1 (R@1) and mean Average Precision (mAP) at various IoU thresholds on QVHighlights, while utilizing R@1 at IoU $\in \{0.3, 0.5, 0.7\}$ alongside the mean IoU (mIoU) of top-1 predictions for Charades-STA and TACoS. For the HD task, we evaluate using mAP and HIT@1 on QVHighlights, and report the standard Top-5 mAP scores on TVSum.

4.2 Implementation details

Feature Representations. Following previous works [40,51,79], video features are extracted using pre-trained SlowFast [10] and CLIP [42] at 2-second intervals for QVHighlights and TACoS, and 1-second intervals for Charades-STA. For TVSum, we employ the I3D [6] pre-trained on Kinetics-400 for visual feature extraction. Across all datasets, textual queries are consistently encoded using the pre-trained CLIP text encoder.

Training Settings. We implement our method in PyTorch and conduct all experiments on a single NVIDIA RTX 3090 GPU. The weights are initialized using Xavier initialization [13] and optimized via AdamW [36] with a weight decay of $1e-4$. For QVHighlights, TACoS, and Charades-STA, we use a batch size of 32 and an initial learning rate of $1e-4$. The training spans 200 epochs for QVHighlights and TACoS, and 100 epochs for Charades-STA. For TVSum, the model is trained for 2000 epochs with a batch size of 4 and a learning rate of $1e-3$. Finally, the loss balancing factors are empirically set to $\lambda_{bnd} = 2$ and $\lambda_{sea} = \lambda_{bea} = 6$. Further implementation details are provided in the Appendix.

4.3 Comparison with State-of-the-arts

Results on QVHighlights. Tab. 1 presents a comprehensive comparison of DE^{2}TR against recent SOTA methods on the QVHighlights *test* and *val* splits. Our method consistently outperforms previous approaches across various settings and metrics. Specifically, compared to the previous best method, TD-DETR [79], DE^{2}TR achieves a +2.05% absolute improvement in Avg. mAP on the test set. Notably, even though TD-DETR relies on its inherent data augmentation mechanism, our model exhibits superior robustness at strict IoU thresholds, yielding gains of +2.51% in mAP@0.75 and +1.02% in R1@0.7. Furthermore, equipping our model with the proposed boundary augmentation strategy widens this performance gap significantly, delivering absolute gains of +3.18%,

Table 1: Experimental results (%) on the QVHighlights *test* and *val* splits. The results are reported on Moment Retrieval (MR) and Highlight Detection (HD). SlowFast [10] and CLIP [42] are utilized as the backbones. **Bold** letters indicate the best results, while underlined results are suboptimal.

Method	test							val						
	MR					HD		MR					HD	
	R1		mAP			≥ Very Good		R1		mAP			≥ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
M-DETR [24]	52.89	33.02	54.82	29.40	30.73	35.69	55.60	53.94	34.84	-	-	32.20	35.65	55.55
UMT† [35]	56.23	41.18	53.83	37.01	36.12	38.18	59.99	60.26	44.26	56.70	39.90	38.59	39.85	64.19
QD-DETR [40]	62.40	44.98	62.52	39.88	39.86	38.94	62.40	62.68	46.66	62.23	41.82	41.22	39.13	63.03
MomentDiff [26]	57.42	39.66	54.02	35.73	35.95	-	-	-	-	-	-	-	-	-
UniVTG [29]	58.86	40.86	57.60	35.59	35.47	38.20	60.96	-	-	-	-	36.13	38.83	-
CG-DETR [39]	65.40	48.40	64.50	42.80	42.90	40.30	66.20	67.40	52.10	65.60	45.70	44.90	40.80	66.70
TR-DETR [51]	64.66	48.96	63.98	43.73	42.62	39.91	63.42	67.10	51.48	66.27	46.42	45.09	40.55	64.77
UVCOM [62]	63.55	47.47	63.37	42.67	43.18	39.74	64.20	65.10	51.81	-	-	45.79	40.03	63.29
BAM-DETR [23]	62.71	48.64	64.57	46.33	45.36	-	-	65.10	51.61	65.41	48.56	47.61	-	-
TD-DETR‡ [79]	64.53	50.37	66.21	47.32	46.69	-	-	65.88	53.67	66.43	49.86	49.05	-	-
KDA [44]	66.70	50.88	67.57	46.31	45.67	-	-	69.11	53.46	68.17	48.04	47.41	-	-
MS-DETR [37]	64.72	48.77	66.41	44.91	44.89	40.45	65.95	66.90	51.68	67.19	46.05	46.00	40.57	66.58
DE^{2}TR	67.24	51.39	67.92	49.83	48.74	**41.08**	66.34	68.13	53.81	68.89	51.61	50.67	40.93	66.65
DE^{2}TR‡	**67.98**	**52.46**	**68.36**	**51.24**	**49.87**	40.51	**66.75**	**69.81**	**55.29**	**70.52**	**53.43**	**52.12**	**41.20**	**66.97**

†represents additional use of audio modality. ‡indicates the use of data augmentation method.

+2.09%, and +3.92% in Avg. mAP, R1@0.7, and mAP@0.75, respectively. This firmly validates the efficacy of our method for highly precise moment localization. Finally, owing to the explicit supervision from the dual prior maps, our model also attains highly competitive performance on the HD task.

Results on Charades-STA and TACoS. To further validate our approach, we evaluate DE^{2}TR on the Charades-STA and TACoS datasets for the MR task. As presented in Tab. 2, our method consistently surpasses all recent state-of-the-art competitors. In particular, under the strict IoU threshold, DE^{2}TR achieves significant absolute gains of +1.35% and +2.40% in R1@0.7 on Charades-STA and TACoS, respectively. These substantial improvements underscore the superior localization capability of our dual-evidence paradigm and its robust generalization across diverse video scenarios.

Results on TVSum. Following evaluation protocols of previous works [40, 53, 62], Tab. 3 summarizes the test results on the TVSum. Our DE^{2}TR demonstrates superior performance compared to recent advanced methods. Specifically, our approach outperforms previous SOTA models across 7 out of the 10 video categories. Furthermore, it achieves a 1.20% absolute improvement in average performance, firmly verifying the robust capability of our method for HD tasks.

4.4 Ablation Studies and Discussions

In this section, we organize ablation studies and discussions of DE^{2}TR, all the experiments are conducted on the *val* split of QVHighlights.

Table 2: Experimental results (%) on Charades-STA and TACoS. All the methods utilize SlowFast and CLIP as backbones.

Method	Charades-STA				TACoS			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
M-DETR [24]	65.83	52.07	30.59	45.54	37.97	24.67	11.97	25.49
QD-DETR [40]	-	57.31	32.55	-	-	-	-	-
MomentDiff [26]	-	55.57	32.42	-	44.78	33.68	-	-
UniVTG [29]	70.81	58.01	35.65	50.10	51.44	34.97	17.35	33.60
UVCOM [62]	-	59.25	36.64	-	-	36.39	23.32	-
LLMEPET [18]	70.91	-	36.49	50.25	52.73	-	22.78	36.55
DualGround [20]	-	**61.11**	38.52	-	-	-	-	-
MS-DETR [37]	71.34	59.62	36.48	50.59	53.16	39.65	23.42	37.01
FlashVTG [4]	-	60.11	38.01	-	53.71	41.76	24.74	37.61
KDA [44]	-	60.23	37.63	-	-	40.13	24.34	-
DE^{2}TR	**72.31**	60.39	**39.87**	**51.58**	**56.29**	**42.52**	**27.14**	**39.76**

Table 3: Comparison of Top-5 mAP results (%) with recent sota methods on TVSum.

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Avg.
TCG [68]	85.0	71.4	81.9	78.6	80.2	75.5	71.6	77.3	78.6	68.1	76.8
UMT [35]	87.5	81.5	88.2	78.8	81.4	87.0	76.0	86.9	84.4	79.6	83.1
QD-DETR [40]	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7	85.0
UniVTG [29]	83.9	85.1	89.0	80.1	84.6	81.4	70.9	91.7	73.5	69.3	81.0
TR-DETR [51]	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	81.6	88.1
UVCOM [62]	87.6	91.6	91.4	86.7	86.9	86.9	76.9	92.3	87.4	75.6	86.3
R2-Tuning [34]	85.0	85.9	91.0	81.7	88.8	87.4	78.1	89.2	90.3	74.7	85.2
MS-DETR [37]	89.8	93.3	94.8	88.7	88.5	89.0	80.5	94.0	88.7	76.9	88.4
MQVTG [53]	87.7	91.6	92.3	85.2	85.7	**91.3**	78.5	**96.5**	90.6	**82.9**	88.2
DE^{2}TR	**91.7**	**94.1**	**95.4**	**88.9**	**89.5**	90.4	**81.2**	92.3	**91.6**	80.9	**89.6**

Effect of different components. In Tab. 4, we investigate the impact of each component in DE^{2}TR. Specifically, applying only our DEFormer architecture improves the Avg. mAP by 0.67%, demonstrating the necessity of the dual-branch design. Adding the semantic and boundary evidence alignment losses further brings a +3.61% absolute improvement in Avg. mAP over the baseline, verifying their effectiveness. In addition, incorporating the prior-guided refinement mechanism and the boundary augmentation strategy yields Avg. mAP gains of +1.43% and +1.45%, respectively. Finally, utilizing all components leads to a +6.49% overall increase in Avg. mAP, which fully demonstrates the effectiveness of the proposed components in DE^{2}TR.

Features from different backbones. As shown in Tab. 5a, we present the effect of different backbones on performance. The model exhibits stronger performance when using InternVideo2 [59] (IV2) compared to SlowFast [10] (SF) + CLIP [42] (C), which benefits from its naturally aligned multimodal features. Our model achieves SOTA results across different backbones. This confirms the effectiveness of our method independent of feature extractors.

Table 4: Effects of different components on QVHighlights *val* split. *DEF* denotes the *DEFormer* structure, and *BAS* indicates *Boundary Augmentation Strategy*.

DEF	$\mathcal{L}_{bea}$	$\mathcal{L}_{sea}$	P_s	P_b	*BAS*	R1		mAP
						@0.5	@0.7	Avg.
						65.68	49.87	45.63
✓						65.23	50.52	46.30
✓	✓					66.71	50.97	48.08
✓		✓				67.03	49.35	47.69
✓	✓	✓				66.84	52.13	49.24
✓	✓	✓	✓			67.42	52.39	49.74
✓	✓	✓		✓		66.77	52.71	49.69
✓	✓	✓	✓	✓		68.13	53.81	50.67
✓	✓	✓	✓	✓	✓	**69.81**	**55.29**	**52.12**

Table 5: Detailed discussions of the proposed DE^{2}TR on QVHighlights *val* split.

(a) Robustness on different backbones

Method	**Backbone**	R1		mAP
		@0.5	@0.7	Avg.
TD-DETR [79]	SF+C	65.88	53.67	49.05
DE^{2}TR	SF+C	**68.13**	**53.81**	**50.67**
TD-DETR [79]	IV2	71.19	55.81	51.72
DE^{2}TR	IV2	**71.84**	**57.68**	**53.91**

(b) Effect of query strategies

Method	R1		mAP
	@0.5	@0.7	Avg.
Merged	66.71	52.90	49.53
Separate	**68.13**	**53.81**	**50.67**

(c) Choice of prior map injection

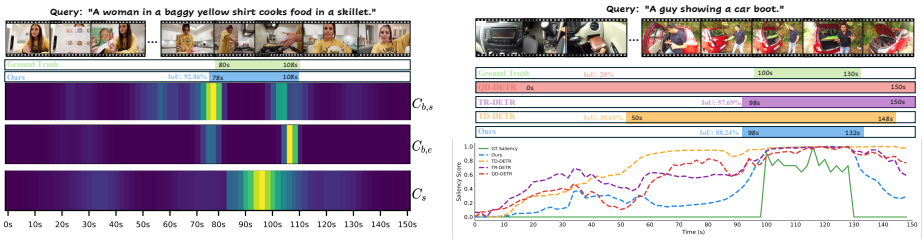
Sem. Branch	Bnd. Branch	R1		mAP
		@0.5	@0.7	Avg.
memory	memory	67.74	52.97	50.14
query	query	**68.97**	53.34	50.32
memory	query	67.51	52.05	49.86
query	memory	68.13	**53.81**	**50.67**

(d) Influence of query numbers N

N	R1		mAP
	@0.5	@0.7	Avg.
5	67.68	53.48	48.56
10	68.13	53.81	50.67
15	68.52	53.35	**51.22**
20	**69.35**	**54.00**	51.19

Query strategies for prediction heads. As shown in Tab. 5b, we compare a "merged" strategy (using a shared instance-level query for all predictions) against our "separate" strategy (routing decoupled queries to their respective heads). Results indicate that the "separate" design yields superior performance by ensuring strict task-feature alignment, where semantic and boundary queries exclusively focus on global relevance and fine-grained refinement, respectively.

Prior-guided refinement mechanism. In Tab. 5c, we investigate strategies for prior map injection. Empirical results reveal that injecting semantic priors into queries yields superior performance, as queries are inherently suited for capturing global intent. Conversely, embedding boundary priors directly into the memory sequence proves more effective for preserving fine-grained local cues essential for precise localization. Consequently, our prior-guided refinement mechanism employ a branch-specific injection design to achieve optimal synergy.



(a) Attention maps for queries. **(b)** Qualitative results.

Fig. 4: Analytical experiments on the QVHighlights *val* split. **(a)** Visualization of the attention maps for boundary queries $C_{b,s}, C_{b,e}$ and semantic query C_s. Diverse regions are learned for different queries. **(b)** Qualitative results for our methods and recent advanced methods. For more qualitative results, please refer to the Appendix.

Numbers of the moment query. We study the impact of different query numbers on performance in Tab. 5d. The performance improves steadily and eventually saturates as the number of queries increases. For a fair comparison, we set $N = 10$ in alignment with previous works [24, 40, 79].

Branch-specific attention. In Fig. 4a, we visualize the attention maps learned by the semantic and boundary queries in DEFormer. Specifically, the semantic query captures global context with broad activations, whereas boundary queries sharply pinpoint the moment edges. Our dual-branch design effectively guides different queries to focus on distinct temporal regions, thereby capturing the complementary evidence required for precise moment localization.

Qualitative results. As shown in Fig. 4b, previous methods (*e.g.*, TD-DETR [79]) suffer from severe boundary misalignment. Relying solely on global semantic alignment, they struggle with ambiguous boundaries caused by semantic gradient. In contrast, thanks to the dual-branch design, our model effectively captures fine-grained boundary cues while maintaining global semantic alignment, thereby predicting well-aligned boundaries. The sharp transitions of the saliency scores around ground-truth boundaries further demonstrate our model's ability to learn discriminative boundary patterns.

5 Conclusion

In this paper, we proposed DE^{2}TR to tackle the boundary misalignment issue in DETR-based VTG methods. To simultaneously align global semantics and capture boundary cues, we designed a dual-branch decoder (DEFormer) for independent evidence aggregation. To further guide this process, we introduced a prior-guided refinement mechanism that steers DEFormer toward potential regions of interest. Furthermore, a boundary augmentation strategy was proposed to fortify the model's boundary awareness in real-world videos. Extensive experiments on multiple challenging benchmarks demonstrate that DE^{2}TR surpasses previous state-of-the-art methods. We hope that our dual-evidence learning paradigm provides a new perspective for fine-grained video understanding.

Acknowledgements

This work was supported in part by the Shaanxi Provincial Natural Science Foundation (Grant No. 2025JC-YBMS-733), the Open Fund for Scientific and Technological Innovation of the Shaanxi Provincial Engineering Technology Innovation Center for Farmland Protection, Monitoring, and Supervision (Grant No. 2025GDBHJCJG07), the National Natural Science Foundation of China (Grant No. 62272380), and the Fundamental Research Funds for the Central Universities, China (Grant No. xzy012026026).

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Int. Conf. Comput. Vis. pp. 5803–5812 (2017)
2. Bhattacharya, U., Wu, G., Petrangeli, S., Swaminathan, V., Manocha, D.: Highlightme: Detecting highlights from human-centric videos. In: Int. Conf. Comput. Vis. pp. 8157–8167 (2021)
3. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4985–4995 (2022)
4. Cao, Z., Zhang, B., Du, H., Yu, X., Li, X., Wang, S.: Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 9226–9236. IEEE (2025)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229 (2020)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6299–6308 (2017)
7. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: Conf. Empir. Methods Nat. Lang. Process. pp. 162–171 (2018)
8. Chen, Y.W., Tsai, Y.H., Yang, M.H.: End-to-end multi-modal video temporal grounding. Adv. Neural Inform. Process. Syst. **34**, 28442–28453 (2021)
9. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Int. Conf. Comput. Vis. pp. 6569–6578 (2019)
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Int. Conf. Comput. Vis. pp. 6202–6211 (2019)
11. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: Eur. Conf. Comput. Vis. pp. 214–229. Springer (2020)
12. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Int. Conf. Comput. Vis. pp. 5267–5275 (2017)
13. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Int. Conf. Artif. Intell. Stat. pp. 249–256 (2010)
14. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Int. Conf. Artif. Intell. Stat. pp. 315–323 (2011)
15. Huang, J., Jin, H., Gong, S., Liu, Y.: Video activity localisation with uncertainties in temporal boundary. In: Eur. Conf. Comput. Vis. pp. 724–740. Springer (2022)

16. Iglovikov, V., Shvets, A.: Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 (2018)
17. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: Int. Conf. Comput. Vis. pp. 13846–13856 (2023)
18. Jiang, Y., Zhang, W., Zhang, X., Wei, X.Y., Chen, C.W., Li, Q.: Prior knowledge integration via llm encoding and pseudo event regulation for video moment retrieval. In: ACM Int. Conf. Multimedia. pp. 7249–7258 (2024)
19. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Int. Conf. Comput. Vis. pp. 1780–1790 (2021)
20. Kang, M., Lee, M., Kim, M., Kim, D., Lee, S.: Empower words: Dualground for structured phrase and sentence-level temporal grounding. Adv. Neural Inform. Process. Syst. **38**, 92472–92499 (2025)
21. Kudi, S., Namboodiri, A.M.: Words speak for actions: Using text to find video highlights. In: Asian Conf. Pattern Recognit. pp. 322–327. IEEE (2017)
22. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Eur. Conf. Comput. Vis. pp. 734–750 (2018)
23. Lee, P., Byun, H.: Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In: Eur. Conf. Comput. Vis. pp. 220–238. Springer (2024)
24. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Adv. Neural Inform. Process. Syst. **34**, 11846–11858 (2021)
25. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13619–13627 (2022)
26. Li, P., Xie, C.W., Xie, H., Zhao, L., Zhang, L., Zheng, Y., Zhao, D., Zhang, Y.: Momentdiff: Generative video moment retrieval from random to real. Adv. Neural Inform. Process. Syst. **36**, 65948–65966 (2023)
27. Lin, C., Xu, C., Luo, D., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Learning salient boundary feature for anchor-free temporal action localization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3320–3329 (2021)
28. Lin, J., Liu, D., Chen, X., Qu, X., Yang, X., Zhu, J., Zhang, S., Dong, J.: Audio does matter: Importance-aware multi-granularity fusion for video moment retrieval. In: ACM Int. Conf. Multimedia. pp. 6027–6036 (2025)
29. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Int. Conf. Comput. Vis. pp. 2794–2804 (2023)
30. Liu, M., Wang, X., Nie, L., He, X., Chen, B., Chua, T.S.: Attentive moment retrieval in videos. In: Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. pp. 15–24 (2018)
31. Liu, M., Wang, X., Nie, L., Tian, Q., Chen, B., Chua, T.S.: Cross-modal moment localization in videos. In: ACM Int. Conf. Multimedia. pp. 843–851 (2018)
32. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. In: Int. Conf. Learn. Represent. (2022)
33. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: Eur. Conf. Comput. Vis. pp. 38–55. Springer (2024)
34. Liu, Y., He, J., Li, W., Kim, J., Wei, D., Pfister, H., Chen, C.W.: r 2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In: Eur. Conf. Comput. Vis. pp. 421–438. Springer (2024)

35. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3042–3051 (2022)
36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Int. Conf. Learn. Represent. (2019)
37. Ma, H., Wang, G., Yu, F., Jia, Q., Ding, S.: Ms-detr: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning. In: ACM Int. Conf. Multimedia. pp. 4514–4523 (2025)
38. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: Int. Conf. Comput. Vis. pp. 3651–3660 (2021)
39. Moon, W., Hyun, S., Lee, S., Heo, J.P.: Correlation-guided query-dependency calibration for video temporal grounding. arXiv preprint arXiv:2311.08835 (2023)
40. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23023–23033 (2023)
41. Narasimhan, M., Rohrbach, A., Darrell, T.: Clip-it! language-guided video summarization. Adv. Neural Inform. Process. Syst. **34**, 13988–14000 (2021)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PmLR (2021)
43. Rahman, M.A., Wang, Y.: Optimizing intersection-over-union in deep neural networks for image segmentation. In: Int. Symp. Visual Comput. pp. 234–244. Springer (2016)
44. Ran, R., Wei, J., He, S., Ma, Z., Zhang, C., Xie, N., Yang, Y.: Kda: Knowledge diffusion alignment with enhanced context for video temporal grounding. In: Int. Conf. Comput. Vis. pp. 23311–23320 (October 2025)
45. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Annu. Meet. Assoc. Comput. Linguist. **1**, 25–36 (2013)
46. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inform. Process. Syst. **28** (2015)
47. Roh, B., Shin, J., Shin, W., Kim, S.: Sparse detr: Efficient end-to-end object detection with learnable sparsity. In: Int. Conf. Learn. Represent. (2022)
48. Shao, J., Wang, X., Quan, R., Zheng, J., Yang, J., Yang, Y.: Action sensitivity learning for temporal action localization. In: Int. Conf. Comput. Vis. pp. 13457–13469 (2023)
49. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1049–1058 (2016)
50. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5179–5187 (2015)
51. Sun, H., Zhou, M., Chen, W., Xie, W.: Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In: AAAI Conf. Artif. Intell. pp. 4998–5007 (2024)
52. Sun, X., Shi, L., Wang, L., Zhou, S., Xia, K., Wang, Y., Hua, G.: Diversifying query: Region-guided transformer for temporal sentence grounding. In: AAAI Conf. Artif. Intell. pp. 7131–7139 (2025)

53. Sun, X., Wang, L., Zhou, S., Shi, L., Xia, K., Liu, M., Wang, Y., Hua, G.: Moment quantization for video temporal grounding. In: Int. Conf. Comput. Vis. pp. 20137–20146 (October 2025)
54. Tang, H., Zhu, J., Liu, M., Gao, Z., Cheng, Z.: Frame-wise cross-modal matching for video moment retrieval. IEEE Trans. Multimedia **24**, 1338–1349 (2021)
55. Um, S.J., Kim, D., Lee, S., Kim, J.U.: Watch video, catch keyword: Context-aware keyword attention for moment retrieval and highlight detection. In: AAAI Conf. Artif. Intell. pp. 7473–7481 (2025)
56. Wang, J., Sun, G., Wang, P., Liu, D., Dianat, S., Rabbani, M., Rao, R., Tao, Z.: Text is mass: Modeling as stochastic embedding for text-video retrieval. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16551–16560 (2024)
57. Wang, T., Yuan, L., Chen, Y., Feng, J., Yan, S.: Pnp-detr: Towards efficient visual analysis with transformers. In: Int. Conf. Comput. Vis. pp. 4661–4670 (2021)
58. Wang, W., Huang, Y., Wang, L.: Language-driven temporal activity localization: A semantic matching reinforcement learning model. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 334–343 (2019)
59. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: Eur. Conf. Comput. Vis. pp. 396–416. Springer (2024)
60. Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor detr: Query design for transformer-based object detection. arXiv preprint arXiv:2109.07107 (2021)
61. Wei, F., Wang, B., Ge, T., Jiang, Y., Li, W., Duan, L.: Learning pixel-level distinctions for video highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3073–3082 (2022)
62. Xiao, Y., Luo, Z., Liu, Y., Ma, Y., Bian, H., Ji, Y., Yang, Y., Li, X.: Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18709–18719 (2024)
63. Xie, T.T., Tzelepis, C., Patras, I.: Boundary uncertainty in a single-stage temporal action localization network. arXiv preprint arXiv:2008.11170 (2020)
64. Xiong, B., Kalantidis, Y., Ghadiyaram, D., Grauman, K.: Less is more: Learning highlight detection from video duration. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1258–1267 (2019)
65. Xu, Z., Chen, D., Wei, K., Deng, C., Xue, H.: Hisa: Hierarchically semantic associating for video temporal grounding. IEEE Trans. Image Process. **31**, 5178–5188 (2022)
66. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16442–16453 (2022)
67. Yang, J., Wei, P., Li, H., Ren, Z.: Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18308–18318 (2024)
68. Ye, Q., Shen, X., Gao, Y., Wang, Z., Bi, Q., Li, P., Yang, G.: Temporal cue guided video highlight detection with low-rank audio-visual fusion. In: Int. Conf. Comput. Vis. pp. 7950–7959 (2021)
69. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. Adv. Neural Inform. Process. Syst. **32** (2019)
70. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10287–10296 (2020)

71. Zeng, Y., Cao, D., Wei, X., Liu, M., Zhao, Z., Qin, Z.: Multi-modal relational graph for cross-modal video moment retrieval. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2215–2224 (2021)
72. Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1247–1257 (2019)
73. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: Int. Conf. Learn. Represent. (2023)
74. Zhang, H., Sun, A., Jing, W., Zhen, L., Zhou, J.T., Goh, R.S.M.: Natural language video localization: A revisit in span-based question answering framework. IEEE Trans. Pattern Anal. Mach. Intell. **44**(8), 4252–4266 (2021)
75. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Annu. Meet. Assoc. Comput. Linguist. pp. 6543–6554 (2020)
76. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: AAAI Conf. Artif. Intell. vol. 34, pp. 12870–12877 (2020)
77. Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10668–10677 (2020)
78. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16965–16974 (2024)
79. Zhou, X., Wei, F., Duan, L., Yao, A., Li, W.: The devil is in the spurious correlations: Boosting moment retrieval with dynamic learning. In: Int. Conf. Comput. Vis. pp. 20981–20990 (October 2025)
80. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: Int. Conf. Learn. Represent. (2021)