SE-DETR: Explicit Semantic Exploration for Generalizability and Distinguishability in Video Temporal Grounding

Chengyang Hu[1†*], Guanshuo Wang[2†], Fufu Yu[2], Qiong Jia[2], Shouhong Ding[2‡], and Lizhuang Ma[1,3‡]

[1] Shanghai Jiao Tong University
[2] Tencent Youtu Lab
[3] MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
{huchengyang, lzma}@sjtu.edu.cn
{mediswang, fufuyu, boajia, ericshding}@tencent.com

Abstract. Video Temporal Grounding (VTG) aims to localize moments in untrimmed videos based on language queries. A fundamental challenge lies in learning cross-modal representations that capture both semantic correspondence and temporal specificity. Existing approaches predominantly rely on token-level visual-textual cross-attention, which may not fully capture two essential properties: generalizability across videos and distinguishability within videos. In practice, semantically related vocabularies may correspond to similar visual concepts, while temporally sparse yet crucial semantics can be easily diluted by frequently occurring visual patterns. To address these challenges, we propose Semantic-Explicit Detection Transformer (SE-DETR), a unified framework that explicitly models semantic structures for video-language grounding. We introduce a Semantics-Proxied Alignment (SPA) module that learns semantic proxies dynamically optimized during training, and enables concept-level alignment between multimodal representations. Furthermore, we design a Temporal Sparsity Modulation (TSM) module that estimates the temporal sparsity of semantic components and dynamically reweights word-guided visual features to highlight informative concepts for accurate moment localization. Extensive experiments on five benchmarks demonstrate that SE-DETR consistently achieves competitive performance across Video Moment Retrieval, Highlight Detection, and Video Summarization tasks. The source code is avaliable at: https://github.com/hu-cheng-yang/ECCV26-SE-DETR.

Keywords: Video Temporal Grounding · Cross-Modal Alignment · Semantic Representation

† Equal contribution. ‡ Corresponding authors.
* Work done during an internship at Tencent Youtu Lab.

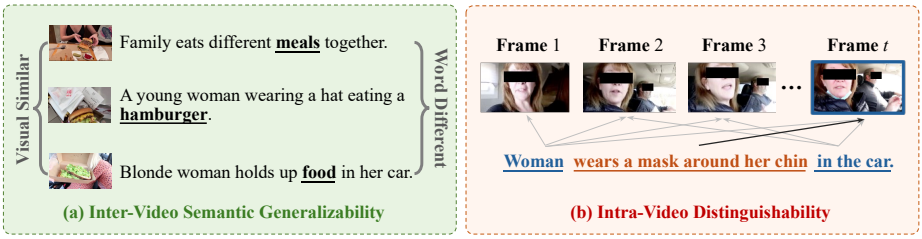

Fig. 1: Examples from QVHighlight that show the "generalizability" and "distinguishability". (a) For inter-video generalizability, different words like "meals", "hamburger", and "food" that contain overlapping concepts indicate similar visual features. (b) For in-video distinguishability, "woman" and "in the car" appear in most of the frames, and only the concept that "wears a mask around her chin" appears in the ground-truth interval, which should be distinguished from other semantic information.

1 Introduction

The rapid growth of video content on online platforms has prompted research [16, 19, 29, 50] into the specific moment searching within videos using natural language prompts, a field known as Video Temporal Grounding (VTG). VTG primarily encompasses three tasks: 1) **Video Moment Retrieval** (VMR), which aims to identify the time boundaries for given descriptions within a video; 2) **Highlight Detection** (HD), which predicts the saliency scores for individual video frames; and 3) **Video Summarization** (VS), which selects a series of shots to provide a quick overview for the provided video.

The core challenge of VTG lies in extracting visual representations that are semantically aligned with textual descriptions while preserving precise temporal boundaries. Starting with Moment-DETR [13], which first unified VTG tasks by proposing a transformer architecture, subsequent works, such as QD-DETR [28] and CG-DETR [27], primarily focus on designing cross-attention mechanisms between the textual and visual features to aggregate visual features related to query concepts. Subsequently, more fine-grained multi-modality alignment methods are introduced. R^2-Tuning [21] utilizes the patch-level visual features to aggregate refined information. Keyword-DETR [41] and CDTR [32] align the visual features to word-level features to identify subtle semantic differences visually. Recently, models like TimeChat [34], VTimeLLM [11], and LLaViLo [25] have leveraged vision-language large models to predict intervals based on textual outputs.

Despite previous advances, these approaches implicitly assume that each word corresponds to an independent visual concept, which is often violated in practice. We observe that semantic understanding in VTG requires addressing two fundamental challenges: First, inter-video semantic **generalizability** is necessary for modeling concept-level similarities across different lexical expressions. For example in Figure 1 (a), query words such as “meals”, “hamburgers”, and “food” correspond to closely related visual concepts but differ in lexical form. Word-

level alignment treats these tokens independently, which limits the ability of the model to generalize across semantically related vocabularies and may lead to concept fragmentation. Second, intra-video semantic **distinguishability** is critical for identifying temporally localized events. As demonstrated in Figure 1 (b), redundant query concepts (e.g., "woman" or "in the car") appear throughout most video frames, while the truly grounding-relevant concept (e.g., "wearing a mask around her chin") may occur only briefly along both spatial and temporal dimensions. Treating all query tokens equally may dilute the contribution of these sparse yet crucial semantics. However, most existing approaches focus on refining cross-attention mechanisms or introducing deeper architectures, largely overlooking explicit semantic modeling.

To address these limitations, we propose the Semantic-Explicit Detection Transformer (SE-DETR), a unified framework that explicitly models both semantic generalizability and distinguishability. For semantic generalization, we introduce a Semantics-Proxied Alignment (SPA) module that aligns word tokens and visual features through a set of learnable semantic proxies. Instead of directly aligning visual features with individual tokens, SPA maps them into a shared semantic space where related vocabularies can converge to similar concept representations. For semantic distinguishability, we propose a Temporal Sparsity Modulation (TSM) module that measures the temporal sparsity of query-related semantic components. By leveraging self-information derived from cross-frame similarity, TSM emphasizes sparse yet informative semantics while suppressing temporally redundant ones. Experiments are conducted for VTG tasks on five public benchmarks, including QVHighlight [13], Charades-STA [5], TACoS [33], TVSum [37], and Youtube-HL [40] with promising results.

Our main contributions are summarized as follows: **1)** We identify two key semantic properties in VTG: inter-video semantic generalizability and intra-video semantic distinguishability, and propose to explicitly model them in a unified framework. **2)** We propose Semantics-Proxied Alignment (SPA) to learn concept-level semantic anchors that improve cross-modal generalization across diverse vocabularies. **3)** We design Temporal Sparsity Modulation (TSM) to quantify the temporal sparsity of semantic components and emphasize informative semantics for precise grounding. **4)** Extensive experiments and visualizations demonstrate the effectiveness of our method against state-of-the-art competitors.

2 Related Work

The main issue of Video Temporal Grounding (VTG) tasks is extracting visual features from video that need to have semantic information matching the provided query. After the success of Moment-DETR [13], the following works mainly focus on aggregating effective query-related visual features via cross-attention operation. QD-DETR [28] first utilizes the cross-attention layer to inject query features into the visual feature. After that, CG-DETR [27] introduces dummy tokens in the query features to prevent the imposed attention between the visual and text features. On the other side, fine-grained multi-modality methods are trying

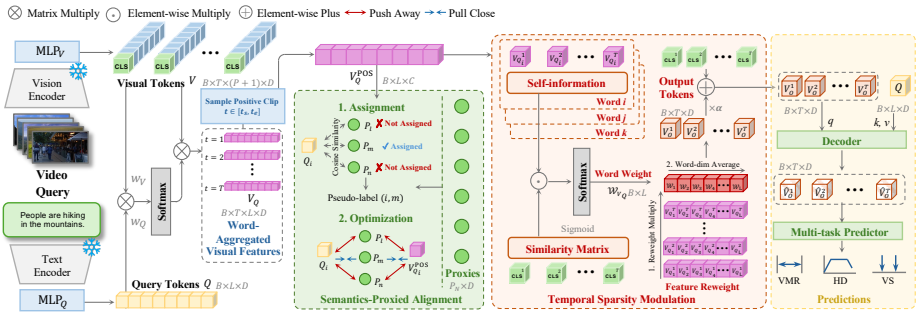

Fig. 2: The framework of SE-DETR. All subscripts in the variables (*e.g.* $Q_{(\cdot)}$, $P_{(\cdot)}$, $V_{(\cdot)}$) indicate the components along the word dimension, and all superscripts (*e.g.* ${V_Q}_i^{(\cdot)}$, $V_O^{(\cdot)}$) indicate components along the temporal dimension.

to align the patch features from video or word features from the query for refinement. R^2-Tuning [21] utilizes multi-layer patch features to obtain more faithful and comprehensive query-related visual features. To identify subtle semantic differences visually, Keyword-DETR [41] and CDTR [32] adopt the strategy of aligning visual features with word-level features. Also, works like UMT [22] and UniVTG [17] unified the VTG tasks and tend to propose simple but deep transformer layers with large parameters to obtain effective features. Recently, TimeChat [34], VTimeLLM [11], and LLaViLo [25] have utilized vision-language large models to convert the VTG tasks to visual question-answering tasks, requiring high computational cost and slow inference. However, all previous methods still lack the exploration of explicit semantic feature learning with both inter-video "generalizability" and in-video "distinguishability". For this, we design a Semantic-Explicit Detection Transformer (SE-DETR) framework to refine the explicit semantic information extraction in feature learning.

3 Method

In this section, we propose Semantic-Explicit Detection Transformer (SE-DETR), as shown in Figure 2. Given the video features V and text query features Q, the goal is to localize the target moment $[t_s, t_e]$ (VMR task), to predict the saliency scores for clips $S = \{s_1, s_2, \cdots, s_T\}, s_t \in [0, 1]$ (HD task), or to obtain the selection of frames $F = \{f_1, f_2, \cdots, f_T\}, f_t \in \{0, 1\}$ (VS task). Different from the conventional configuration [13, 27, 28], we introduce fine-grained features from CLIP [31] as $V \in \mathbb{R}^{B\times T\times (P+1)\times D_V}, Q \in \mathbb{R}^{B\times L\times D_Q}$, where B, T, P, L, D_V and D_Q are the batch size, number of video frames, number of patches, number of query tokens, dimensions of the video and text query features. To align the multimodal features into common feature spaces, MLP layers are first utilized to map the features into the same dimension D as: $V = \text{MLP}_V(V) \in \mathbb{R}^{B\times T\times (P+1)\times D}, Q = \text{MLP}_Q(Q) \in \mathbb{R}^{B\times L\times D}$, where D is the common feature dimension. Our framework builds upon the DETR-style grounding architecture while introducing explicit semantic modeling modules before the decoder.

3.1 Semantics-Proxied Alignment

To align visual features with precise semantics within a query in a **generalizable** way, we propose the Semantics-Proxied Alignment (SPA) module to align visual features to fine-grained query tokens via semantic proxies. Previous work directly aligns visual features to word-level tokens [32]; however, CLIP text embeddings are known to form clusters that correspond to semantic concepts rather than individual lexical tokens [24]. We introduce learnable semantic proxies to obtain shared concepts across different vocabularies in queries. We initialize learnable proxies as $C \in \mathbb{R}^{C_N \times D}$, where C_N is the number of semantic proxies.

To prevent trivial solutions caused by arbitrary proxy initialization, we initialize the semantic proxies using K-means [7] centroids computed from the global [CLS] embeddings of training queries. These centroids provide a coarse semantic coverage of the query feature space, serving as anchor points for optimization rather than fixed semantic clusters. Instead of just static prototype clustering, proxies are further updated during training and are not constrained to represent specific lexical groups.

Then, we aggregate the visual features corresponding to the word tokens in the query. For visual patch features $V_P \in \mathbb{R}^{B \times T \times P \times D}$ without the [CLS] feature and Q, we broadcast the Q into $\mathbb{R}^{B \times T \times L \times D}$ and calculate the aggregated visual features to word features as:

$$V_Q = \text{Softmax}(\frac{w_Q Q \cdot w_V V_P^\top}{\sqrt{D}}) \cdot V_P \in \mathbb{R}^{B \times T \times L \times D}, \tag{1}$$

where w_Q, w_V are the learnable mapping matrices, V_Q is the aggregated visual features corresponding to different words.

To enable semantically related vocabularies to share higher-level concept anchors for generalization improvement, we first map the word query features to semantic proxies to explore potential shared semantic concepts across different vocabularies (Assignment Step). Subsequently, both word and visual tokens are optimized such that the distribution of visual tokens aligns with that of the semantic feature space (Optimization Step).

Assignment step. For feature alignment, we just utilize positive clips from the video. Specifically, one clip is randomly sampled from the ground-truth interval, denoted as $V_Q^{\text{POS}} \in \mathbb{R}^{B \times L \times D}$. To explore shared semantic concepts among known vocabularies in a self-supervised manner, we assign each word token in the query to its nearest semantic proxy as a word-concept pair:

$$\begin{aligned} S_{Q-C} &= \text{Softmax}(\cos(Q, C)) \in [0,1]^{B \times L \times C_N}, \\ A &= \arg\max_{C_N}(S_{Q-C}), \end{aligned} \tag{2}$$

where S_{Q-C} is the similarity matrix between the word tokens and all semantic proxies. A is the assignment result obtained by selecting the proxy with the highest similarity.

Optimization step. The assignment result A guides the refinement of multi-modal feature distributions through contrastive learning. Specifically, for the

word tokens, we minimize their distance to the assigned semantic proxy while maximizing separation from all other proxies. Simultaneously, the visual features are optimized to maintain a distribution consistent with their corresponding word tokens. This joint alignment is achieved by minimizing the cross-entropy losses:

$$\begin{aligned} S_{V-C} &= \text{Softmax}(\cos\left(V_Q^{\text{POS}}, C\right)) \in [0,1]^{B\times L\times C_N}, \\ \mathcal{L}_{Q-C} &= -\frac{1}{BL}\sum_{BL}\sum_{c=1}^{C_N} \mathbb{1}(c=A)\log(S_{Q-C}[c]), \\ \mathcal{L}_{V-C} &= -\frac{1}{BL}\sum_{BL}\sum_{c=1}^{C_N} \mathbb{1}(c=A)\log(S_{V-C}[c]), \end{aligned} \tag{3}$$

where S_{V-C} denotes the similarity matrix between the visual tokens V_Q^{POS} and the semantic proxies C. $\mathbb{1}(\cdot) \in \{0,1\}$ is the condition function. $S[c]$ selects the similarity score corresponding to the c-th proxy.

Unlike explicit word-level clustering, proxy assignment does not enforce discrete semantic groups. Instead, semantic proxies serve as intermediate alignment anchors that facilitate cross-modal distribution matching between visual tokens and textual semantics. Although the nearest proxy is used for supervision, the optimization operates over the full similarity distribution, which preserves semantic diversity and avoids collapsing tokens into fixed clusters. Also, proxies are only used during training to regularize the semantic space.

In addition to fine-grained multi-modality alignment, we incorporate a global clip-sentence alignment for [CLS] tokens through InfoNCE loss [8]:

$$\mathcal{L}_{Q-V} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\cos(\bar{V}_{\text{CLS}}^{i}, Q_{\text{CLS}}^{i}))}{\sum_j \exp(\cos(\bar{V}_{\text{CLS}}^{i}, Q_{\text{CLS}}^{j}))}, \tag{4}$$

where i, j denote the indices in the batch. $\bar{V}_{\text{CLS}}$ is the average of V_{CLS} over the relative interval.

3.2 Temporal Sparsity Modulation

To accurately localize video moments, semantically essential concepts should be emphasized over those that appear uniformly across the video. However, when visual content is abstracted into high-level feature representations, the matched signals produced by sparse concepts sometimes become easily diluted by dominant and consistently present visual patterns, since their contribution to the overall feature variation is relatively small. We introduce the Temporal Sparsity Modulation (TSM) module to explicitly quantify the temporal sparsity of each word-guided semantic component and modulate its contribution for improved **distinguishability**.

Based on our premise, cross-frame similarity as a proxy can be interpreted as the occurrence probability of semantic components, since concepts that consistently appear across frames tend to produce stable feature similarity patterns.

We first transform the V_Q into $\mathbb{R}^{B\times L\times T\times D}$ and calculate a normalized temporal similarity matrix:

$$\mathcal{A}_{V_Q} = \frac{\cos(V_Q, V_Q^\top) + 1}{2} \in [0,1]^{B\times L\times T\times T}. \tag{5}$$

Then, we utilize self-information as a principled probabilistic formulation to measure temporal sparsity, which is a criterion to assess concept essentiality. Self-information [35] is a classical information-theoretic measure: rare events naturally receive higher importance. The self-information $\mathcal{S}_{V_Q}$ is then calculated as:

$$\mathcal{S}_{V_Q} = -\log((\frac{T(T-1)}{2})^{-1}\sum \mathbf{U}(\mathcal{A}_{V_Q})) \in {\mathbb{R}_+}^{B\times L}, \tag{6}$$

$\mathbf{U}(\cdot)$ is the upper triangular matrix excluding the diagonal of $T\times T$ matrix, and $\frac{T(T-1)}{2}$ is the number of such elements. Here, a low temporal similarity yields a low probability value, leading to a high self-information score, which naturally highlights temporally sparse but informative semantic components that are critical for VTG tasks.

Next, we calculate the overall temporal consistency of the video to measure whether the scarcer semantic concepts will be diluted. The attention matrix for the global [CLS] token of visual features is formulated as:

$$\mathcal{A}_{V_{\text{CLS}}} = \frac{V_{\text{CLS}}\cdot V_{\text{CLS}}^\top}{\sqrt{D}} \in \mathbb{R}^{B\times T\times T}, \tag{7}$$

The global similarity score $\mathcal{S}_{\text{CLS}}$ to modulate the final word weight is calculated from the average attention:

$$\mathcal{S}_{\text{CLS}} = \text{Sigmoid}((\frac{T(T-1)}{2})^{-1}\sum \mathbf{U}(\mathcal{A}_{V_{\text{CLS}}})) \in [0,1]^B, \tag{8}$$

The modulation score $\mathcal{W}_{V_Q}$ for V_Q is finally obtained by:

$$\mathcal{W}_{V_Q} = \text{Softmax}(\mathcal{S}_{V_Q}\cdot\mathcal{S}_{\text{CLS}}) \in [0,1]^{B\times L}. \tag{9}$$

When $\mathcal{S}_{\text{CLS}}\to 1$, the high overall consistency indicates that scarcer semantics are difficult to detect. $\mathcal{W}_{V_Q} = \text{Softmax}(\mathcal{S}_{V_Q})$ modulates V_Q to enhance the weight of features with higher self-information (scarcely appear temporally). If a concept appears only in a small portion of frames, the similarity of its aggregated features across frames will be low. On the other side, when $\mathcal{S}_{\text{CLS}}\to 0$, the overall video features have already been semantically varied, $\mathcal{W}_{V_Q} = 1/L$ treats all word-aggregated visual features uniformly. The fused output visual features V_O are

$$V_O = \alpha\cdot\sum_{i=1}^{L}(\mathcal{W}_{V_{Q_i}}\cdot V_{Q_i}) + V_{\text{CLS}} \in \mathbb{R}^{B\times T\times D}, \tag{10}$$

where $\alpha\in[0,1]$ is a learnable score for different layers.

Subsequently, triplet loss [9] is used to constrain the modulation scheme through video content annotations. Positive and negative video clips are randomly sampled within and outside the ground-truth moment boundary $[t_s, t_e]$ respectively, for both V_O and V_{CLS}:

$$\begin{aligned} V_{\text{CLS}}^{\text{p}} = V_{\text{CLS}}^{t}, t \in [t_s, t_e]; V_{\text{CLS}}^{\text{n}} = V_{\text{CLS}}^{t}, t \notin [t_s, t_e], \\ V_{O}^{\text{p}} = V_{O}^{t}, t \in [t_s, t_e]; V_{O}^{\text{n}} = V_{O}^{t}, t \notin [t_s, t_e]. \end{aligned} \tag{11}$$

The temporal-view triplet loss is formulated by:

$$\begin{aligned} \mathcal{L}_{\text{T}} = \frac{1}{B} \sum_{i}^{B} \text{Triplet}(Q_i, V_i^{\text{p}}, V_i^{\text{n}}), V \in \{V_{\text{CLS}}, V_O\}, \\ \text{Triplet}(\mathbf{a}, \mathbf{p}, \mathbf{n}) = \max(m - \cos(\mathbf{a}, \mathbf{p}) + \cos(\mathbf{a}, \mathbf{n}), 0), \end{aligned} \tag{12}$$

where $\mathbf{a}, \mathbf{p}, \mathbf{n}$ are the anchor, positive, negative features for the triplet loss, and the margin m is set to 0.5. Triplet loss ensures the intra-video distinguishability between related and unrelated moment intervals and promotes the discovery of subtle temporal semantic differences. The global alignment loss $\mathcal{L}_{Q-V}$ is also extended to incorporate V_O:

$$\mathcal{L}_{Q-V} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\cos(\bar{V}^i, Q^i))}{\sum_j \exp(\cos(\bar{V}^i, Q^j))}, V \in \{V_{\text{CLS}}, V_O\}. \tag{13}$$

3.3 Multi-Task Predictions and Losses

In this section, following previous works [17, 21], we employ different prediction heads for different video-text grounding (VTG) tasks. A Transformer decoder is applied to aggregate temporal features:

$$V_F = \text{Decoder}(V_O, Q) \in \mathbb{R}^{B \times T \times D}, \tag{14}$$

where the final feature V_F is obtained as output. The fused feature V_F is then downsampled with various strides using 1D CNN layers, and the final predictions are obtained from visual features at different temporal resolutions.

Video Moment Retrieval. The temporal boundaries are predicted using a 1D convolutional layer for $\hat{t}_s$ and $\hat{t}_e$, denoted as the predicted start and end times. The optimization is performed with an L1 loss defined as:

$$\mathcal{L}_{\text{VMR}} = |t_s - \hat{t}_s| + |t_e - \hat{t}_e|, \tag{15}$$

Highlight Detection. The saliency score is computed as the cosine similarity between its visual feature V_F and the corresponding query feature Q_F:

$$s = \cos(V_F, Q_F), \tag{16}$$

where $Q_F \in \mathbb{R}^{B \times D}$ denotes the adaptively pooled features from the query representations Q. This HD task is optimized with a contrastive loss that distinguishes positive video-query pairs from negatives:

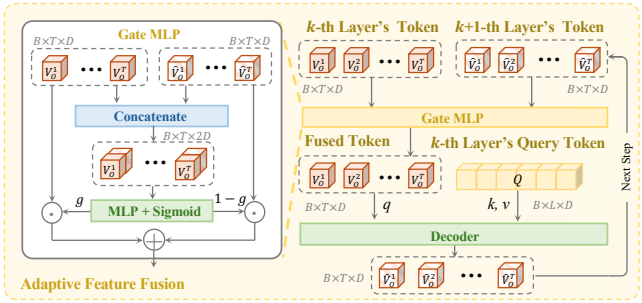

Fig. 3: The illustration of the Adaptive Feature Fusion (AFF) module.

$$\mathcal{L}_{\text{HD}} = -\log \frac{\exp(s_p/\tau)}{\sum_i \exp(s_i/\tau)}, \tag{17}$$

where s_p is the saliency score from the similarity of the positive pairs between video clips and queries, s_i denotes the similarity of all combinations of V_F and Q_F within a batch. The temperature parameter τ is set to 0.07.

Video Summarization. This task is formulated as a binary classification problem to predict the foreground probability of each video segment. The model is optimized using the focal loss [18]:

$$\mathcal{L}_{\text{VS}} = -\beta(1-\hat{f}_i)^{\gamma}\log(\hat{f}_i), \tag{18}$$

where $\hat{f}_i$ is the predicted probability, the hyper-parameters β, γ are set to 0.9 and 0.2.

The final loss function is:

$$\begin{aligned}\mathcal{L} =& \lambda_{\text{VMR}}\mathcal{L}_{\text{VMR}} + \lambda_{\text{HD}}\mathcal{L}_{\text{HD}} + \lambda_{\text{VS}}\mathcal{L}_{\text{VS}} \\ &+ \lambda_{\text{Align}}(\mathcal{L}_{\text{Q-C}} + \mathcal{L}_{\text{V-C}} + \mathcal{L}_{\text{Q-V}} + \mathcal{L}_{\text{T}}).\end{aligned} \tag{19}$$

3.4 Extension for Multi-layer Features

While SE-DETR focuses on semantic modeling, we further extend it to multi-layer feature inputs as SE-DETR+, to utilize previous findings [21, 30] for complementary representations. To extend SE-DETR to the multi-layer setting, we introduce K interval layer features as input. Specifically, the visual and query features are given by $V \in \mathbb{R}^{K\times B\times T\times(P+1)\times D_V}$, $Q \in \mathbb{R}^{K\times B\times L\times D_Q}$. The features of different layers are first computed independently via previous processes in parallel, producing the output feature $V_O \in \mathbb{R}^{K\times B\times T\times D}$.

To integrate features from different layers K, we propose an Adaptive Feature Fusion (AFF) module to adaptively fuse multi-layer features from deep to shallow ($K \to 1$) through a learnable gate, as shown in Figure 3. For the k-th layer feature V_{O_k} ($k < K$, which is not the last layer feature) and the $(k+1)$-th layer

Table 1: Performances on QVHighlight [13] *test* and *val* splits. For features, C+SF is the CLIP+Slowfast features, and C is the features from CLIP. The first group of methods is traditional methods, and the second group uses the transformer architecture. The **bolded** and the underlined scores are the best and the second-best.

Split		*test*							*val*						
Method	Feature	VMR					HD		VMR					HD	
		mAP			R1		≥VeryGood		mAP			R1		≥VeryGood	
		0.5	0.75	Avg.	0.5	0.7	mAP	HIT1	0.5	0.75	Avg.	0.5	0.7	mAP	HIT1
B.Thumb [36]	C+SF	-	-	-	-	-	14.36	20.88	-	-	-	-	-	-	-
DVSE [20]	C+SF	-	-	-	-	-	18.75	21.79	-	-	-	-	-	-	-
MCN [1]	C+SF	24.94	8.22	10.67	11.41	2.72	-	-	-	-	-	-	-	-	-
CAL [3]	C+SF	23.40	7.65	9.89	25.49	11.54	-	-	-	-	-	-	-	-	-
XML [14]	C+SF	44.63	31.73	32.14	41.83	30.35	34.49	55.25	-	-	-	-	-	-	-
XML+ [14]	C+SF	47.89	34.67	34.90	46.69	33.46	35.38	55.06	-	-	-	-	-	-	-
M-DETR [13]	C+SF	54.82	29.40	30.73	52.89	33.02	35.69	55.60	53.94	34.84	32.20	60.26	44.26	35.65	55.55
UMT [22]	C+SF	53.83	37.01	36.12	56.23	41.18	38.18	59.99	60.26	44.26	38.59	60.26	44.26	39.85	64.19
QD-DETR [28]	C+SF	62.52	39.88	39.86	62.40	44.98	38.94	62.40	62.68	46.66	41.22	62.68	46.66	39.13	63.03
UniVTG [17]	C+SF	57.60	35.59	35.47	58.86	40.86	38.20	60.96	59.74	-	36.13	59.74	-	38.83	61.81
EaTR [12]	C+SF	-	-	-	-	-	-	-	61.36	45.79	41.74	61.36	45.79	37.15	58.65
TR-DETR [38]	C+SF	63.98	43.73	42.62	64.66	48.96	39.91	63.42	-	-	-	-	-	-	-
CG-DETR [27]	C+SF	64.50	42.80	42.90	65.40	48.40	40.30	66.20	65.60	45.70	44.90	67.40	52.10	40.80	66.70
Kw-DETR [41]	C+SF	67.73	46.24	45.69	66.86	**51.23**	**40.94**	64.79	-	-	47.69	68.97	53.35	**41.67**	**67.03**
CDTR [32]	C+SF	66.44	45.96	44.37	65.79	49.60	-	-	66.62	46.97	45.85	68.03	52.68	-	-
MS-DETR [23]	C+SF	66.41	44.91	44.89	64.72	48.77	40.45	65.95	67.19	46.05	46.00	66.90	51.68	40.57	66.58
R^2-Tuning [21]	C($K=4$)	69.04	47.56	46.17	68.03	49.35	40.75	64.20	68.55	49.94	47.86	67.74	51.87	39.45	64.13
SE-DETR	C($K=1$)	68.81	48.92	45.73	68.03	48.88	40.61	65.84	69.19	50.67	47.57	69.35	52.45	40.04	66.58
SE-DETR+	C($K=4$)	**69.74**	**49.16**	**46.42**	**68.22**	49.96	40.82	**66.40**	**70.16**	**51.02**	**48.29**	**69.51**	**53.61**	40.27	66.97

feature $\hat{V}_{O_{k+1}}$, feature fusion is performed via a GateMLP layer as:

$$\begin{aligned} V'_{O_k} &= \text{Concat}(V_{O_k}, \hat{V}_{O_{k+1}}) \in \mathbb{R}^{B\times T\times 2D}, \\ g &= \text{Sigmoid}(\text{MLP}(V'_{O_k})) \in [0,1]^{B\times T}, \\ \hat{V}_{O_k} &= g \cdot V_{O_k} + (1-g)\cdot \hat{V}_{O_{k+1}}, \end{aligned} \tag{20}$$

where g is the gate score, Concat$(\cdot)$ is the concatenation operation along the D dimension. MLP$(\cdot)$ consists of stacked linear layers. Subsequently, a Transformer decoder layer is applied to aggregate temporal features like previous steps:

$$\hat{V}_{O_k} = \text{Decoder}(\hat{V}_{O_k}, Q_k) \in \mathbb{R}^{B\times T\times D}, \tag{21}$$

For the case where $k=K$, $\hat{V}_{O_K} = \text{Decoder}(V_{O_K}, Q_K)$ and the GateMLP is not applied. To enhance feature learning across different layers, we introduce a layer-wise contrastive learning objective defined as:

$$\mathcal{L}_K = -\frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K}\log\frac{\exp(\cos(\bar{V}_i^k, Q_i^k))}{\sum_l \exp(\cos(\bar{V}_i^k, Q_i^l))}, V \in \{V_{\text{CLS}}, V_O\}. \tag{22}$$

where i is the index of the batch, k, l are the layer indices. This loss will be added into the final loss $\mathcal{L}$ with the weight λ_{Align}.

4 Experiment

4.1 Benchmarks and Evaluation Metrics

We conduct experiments on five datasets: QVHighlight [13] for both VMR and HD tasks, TACoS [33] and Charades-STA [5] for the VMR task, and Youtube-HL [40]

Table 2: Moment retrieval results. For feature, C+SF is the CLIP+Slowfast features, C is CLIP features. **Bolded** and underlined scores are the best and the second-best.

Method	Feature	Charades-STA [5]				TACoS [33]			
		mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
2D-TAN [49]	C+SF	41.3	58.8	46.0	27.5	27.2	40.0	28.0	12.9
VSLNet [47]	C+SF	41.6	60.3	42.7	24.1	25.0	35.5	23.5	13.2
M-DETR [13]	C+SF	45.5	65.8	52.1	30.6	25.5	38.0	24.7	12.0
UniVTG [17]	C+SF	50.1	70.8	58.0	35.7	33.6	**51.4**	35.0	17.4
R^2-Tuning [21]	C($K=4$)	50.9	70.9	59.8	37.0	35.9	49.7	38.7	**25.1**
SE-DETR	C($K=1$)	48.6	69.2	55.4	34.3	35.1	50.1	38.8	21.4
SE-DETR+	C($K=4$)	**51.2**	**71.1**	**60.2**	**39.7**	**36.2**	50.6	**39.2**	22.6

Table 3: Highlight detection results on Youtube-HL [40]. For feature, V means only using visual feature, V+A means using visual and audio features. The **bolded** and the underlined scores are the best and the second-best results.

Method	Feature	Avg.	Dog	Ska.	Ski.	Sur.	Gym.	Par.
RRAE [45]	V	38.3	49.0	25.0	22.0	49.0	35.0	50.0
GIFs [6]	V	46.4	30.8	55.4	32.8	54.1	33.5	54.0
LSVM [40]	V	53.6	60.0	62.0	36.0	61.0	41.0	61.0
LIM-S [43]	V	56.4	57.9	57.8	48.6	65.1	41.7	67.0
SL-Module [44]	V	69.3	70.8	72.5	66.1	76.2	53.2	77.2
QD-DETR [28]	V	74.4	72.2	72.7	72.8	80.6	**77.4**	71.0
UniVTG [17]	V	75.2	71.8	73.3	73.2	82.2	76.5	73.9
R^2-Tuning [21]	V($K=4$)	76.1	75.6	74.6	74.8	**84.8**	73.5	73.0
MINI-Net [10]	V+A	64.4	58.2	72.2	58.7	65.1	61.7	70.2
TC [46]	V+A	63.0	55.4	69.1	60.1	59.8	62.7	70.9
Joint-VA [2]	V+A	71.8	64.5	62.0	73.2	78.3	71.9	80.8
UMT [22]	V+A	74.9	65.9	71.8	72.3	82.7	75.2	**81.6**
SE-DETR	V	76.4	76.2	73.8	**76.9**	82.7	71.6	77.2
SE-DETR+	V($K=4$)	**77.0**	**76.7**	**75.5**	76.4	83.4	72.5	77.6

and TVSum [37] for the VS task. Following prior work [27], for VMR task on QVHighlight, we measure Recall@1 and mean average precision (mAP) at IoU thresholds $\theta = 0.5$ and 0.7, along with mAP over the range $[0.5 : 0.05 : 0.95]$; highlight detection performance is assessed using mAP and HIT@1 with "Very Good" saliency segments as positives; Charades-STA, and TACoS are evaluated via Recall@1 ($\theta = \{0.3, 0.5, 0.7\}$) and mean IoU; while Youtube-HL and TVSum utilize mAP and Top-5 mAP as respective metrics.

4.2 Implementation Details

As previous works indicate that intermediate layers of the CLIP visual encoder retain richer local and structural information [4, 15, 39], we utilize the second-to-last layer features from CLIP-B/32 [31] and $P = 49$; For multi-layer features, we utilize the last four layers' feature ($K = 4$) following previous work [21]. The feature dimension is set as $D = 256$. The number of proxies C_N is set to 1000 for QVHighlight, 100 for Charades, TACoS and Youtube-HL, 10 for TVSum, depending on vocabulary size. QVHighlight, Charades, and TACoS downsample

Table 4: Video summarization results on TVSum [37]. For feature, V means only using visual feature, V+A means using visual and audio features. The **bolded** and the underlined scores are the best and the second-best results.

Method	Feature	Avg.	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS
sLSTM [48]	V	45.1	41.1	46.2	46.3	47.7	44.8	46.1	45.2	40.6	47.1	45.5
SG [26]	V	46.2	42.3	47.2	47.5	48.9	45.6	47.3	46.4	41.7	48.3	46.6
LIM-S [43]	V	56.3	55.9	42.9	61.2	54.0	60.3	47.5	43.2	66.3	69.1	62.6
Trailer [42]	V	62.8	61.3	54.6	65.7	60.8	59.1	70.1	58.2	64.7	65.6	68.1
SL-Module [44]	V	73.3	86.5	68.7	74.9	86.2	79.0	63.2	58.9	72.6	78.9	64.0
QD-DETR [28]	V	85.0	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7
UniVTG [17]	V	81.0	83.9	85.1	89.0	80.1	84.6	81.4	70.9	91.7	73.5	69.3
R^2-Tuning [21]	V($K=4$)	85.2	85.0	85.9	91.0	81.7	88.8	87.4	78.1	89.2	90.3	74.7
TR-DETR [38]	V	88.1	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	81.6
Keyword-DETR [41]	V	88.7	89.9	**93.8**	**94.4**	85.9	89.2	89.4	**81.5**	**92.6**	90.1	80.6
MINI-Net [10]	V+A	73.2	80.6	68.3	78.2	81.8	78.1	65.8	57.8	75.0	80.2	65.5
TCG [46]	V+A	76.8	85.0	71.4	81.9	78.6	80.2	75.5	71.6	77.3	78.6	68.1
Joint-VA [2]	V+A	76.3	83.7	57.3	78.5	86.1	80.1	69.2	70.0	73.0	**97.4**	67.5
UMT [22]	V+A	83.1	87.5	81.5	88.2	78.8	81.4	87.0	76.0	86.9	84.4	79.6
SE-DETR	V	89.1	89.7	91.9	93.8	87.4	89.5	90.8	79.8	91.0	93.9	**83.2**
SE-DETR+	V($K=4$)	**89.6**	**90.2**	92.5	94.3	**88.1**	**90.2**	**91.4**	80.6	91.6	94.5	82.6

the V_F with strides of $1, 2, 4, 8$. TVSum and Youtube-HL are not downsampled. The weights for losses are $\lambda_{\text{VMR}} = 1$, $\lambda_{\text{VS}} = 0.2$, $\lambda_{\text{HD}} = 0.1$ and $\lambda_{\text{Align}} = 0.1$. Batch sizes are 128 for QVHighlight, 32 for Charades, 16 for TACoS, 4 for TVSum, and Youtube-HL.

4.3 Comparison Results

Joint Video Moment Retrieval and Highlight Detection Tasks We conduct the experiments on QVHighlight [13] benchmark *test* and *val* split with the VMR task and HD tasks. Our method shows promising results on both *test* and *val* splits in Table 1. We have the following observations: 1) Compared to the traditional methods in the first group, the unified architectures show a better performance, which indicates that such a unified design prompts the effective feature learning via jointly multi-task learning. 2) Compared to previous methods, SE-DETR shows promising results. Our method learns inter-video generalizable features and highlights the essential semantics in-video for distinguishability, which provides more explicit semantics in visual feature extraction and improves the performance. 3) With features from more layers, as shown in SE-DETR+, our method shows the best result. Multi-layer features provide more fine-grained information and AFF effectively fuses the features and brings performance gains.

Video Moment Retrieval Tasks In this section, we experiment with the methods for the Video Moment Retrieval task on TACoS [33] and Charades-STA [5], as shown in Table 2. We have the following observations: 1) Our method shows a significant improvement compared to traditional methods like 2D-TAN [49] and VSLNet [47], and DETR-based methods like Moment-DETR [13], which indicates the effectiveness of our designs for explicit semantics exploration. 2)

Table 5: Ablation study on different modules. We experiment on QVHighlight [13] *val* split. Checked (✓) item means the corresponding module is utilized.

SPA	TSM	AFF	VMR			HD	
			R1@0.5	R1@0.7	mAP	mAP	HIT@1
			62.19	46.26	42.15	37.40	61.35
✓			67.74	50.77	45.29	40.02	65.41
	✓		64.97	49.15	44.86	38.68	63.47
		✓	63.47	47.61	43.18	37.47	61.09
✓	✓		68.80	52.68	47.63	40.08	66.02
✓		✓	68.73	51.65	46.72	39.88	65.51
	✓	✓	64.26	49.03	45.00	38.05	62.39
✓	✓	✓	**69.51**	**53.61**	**48.29**	**40.27**	**66.97**

Table 6: Ablation study on different numbers of proxies C_N. We experiment on QVHighlight [13] *val* split.

Number of Proxies	VMR			HD	
C_N	R1@0.5	R1@0.7	mAP	mAP	HIT@1
100	68.90	51.16	46.90	39.72	66.00
250	69.23	52.58	47.75	40.01	66.31
500	69.53	**53.85**	47.95	**40.25**	66.52
1000 **(Ours)**	**69.51**	53.61	**48.29**	40.27	**66.97**
2000	66.97	51.68	46.92	39.72	65.61

SE-DETR shows comparable results to previous methods that have large numbers of parameters [17] or utilize multi-layer features [21], which shows the effectiveness of the proposed method for balancing model parameters and input feature size. 3) Compared to previous methods with transformer architecture, like R^2-Tuning [21] and UniVTG [17], SE-DETR+ shows improvement on both datasets. Previous methods mainly focus on extracting meaningful query-related visual features via a vanilla cross-attention module without any refined multi-modality feature alignment. However, our SE-DETR+ explores extracting effective word-aggregated visual features that are both generalizable and distinguishable. In this way, the semantics in visual features are more explicit to the corresponding query, which leads to a better performance.

Highlight Detection and Video Summarization Tasks We conduct the experiment for HD on Youtube-HL [40] (Table 3) and VS on TVSum [37] (Table 4). Our method shows state-of-the-art results compared to previous methods, including methods with vision-only input and methods with multi-modality input for both vision and audio. SE-DETR is also effective for small-scale datasets. Our design for generalizable and distinguishable semantic feature extraction improves the model to focus more on the essential concepts and brings an effective result.

4.4 Ablation Study

Study on Different Modules In this study, we show the effectiveness of different modules in Table 5 on QVHighlight *val* split [13]. For unchecked items, we remove the module and corresponding losses. When AFF is removed, features are fused by average pooling. We have the following observations: 1) SPA shows a great contribution even when used alone. The reason is that SPA aligns the multimodal features via semantic proxies, which brings generalizable features and is the basis for effective grounding. 2) Equipped with TSM and AFF, the performance improves. TSM modulates word-aggregated visual features and highlights the features that scarcely appear for distinguishability. AFF adaptively fuses the multi-layer features to ensure effectiveness.

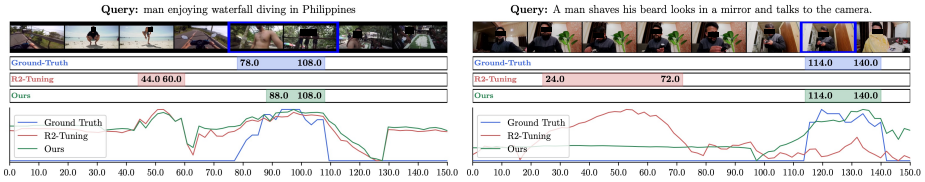

Fig. 4: Visualization of the VMR and HD result on QVHighlight [13] *val* split.

Fig. 5: Visualization of the attention map for visual-text features on QVHighlight [13] *val* split.

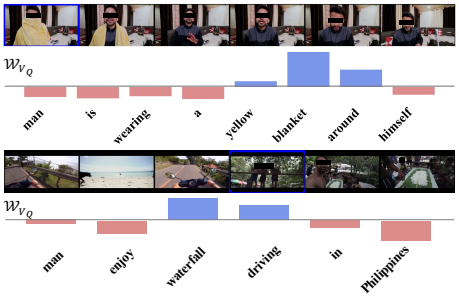

Fig. 6: Visualization of the weights of different word-aggregated visual features $\mathcal{W}_{V_Q}$ in TSM on QVHighlight [13] *val* split.

Study on Number of Proxies C_N We study the effectiveness of the number of semantic proxies C_N in SPA. We select C_N from 100 to 2000 and test on QVHighlight *val* split [13]. In Table 6, we have the following observations: 1) As the number increases, the performance first increases until $C_N = 1000$. More prototypes encourage finer-grained feature learning and yield a more generalizable semantic space distribution. 2) When the number increases extremely ($C_N = 2000$), the performance suddenly degrades. One potential reason is that too many proxies cannot be fully learned; only some are optimized from uneven assignment.

4.5 Visualization

Visualization on Predict Result We visualize the predicted result of R^2-Tuning [21] and SE-DETR in Figure 4 on QVHighlight *val* split. Visualization demonstrates the effectiveness of SE-DETR on both VMR and HD. R^2-Tuning tends to retrieve the moment with partial concepts, *e.g.* "*man*" that are present in the overall video. SE-DETR focuses more on sparse semantics like "*waterfall diving*" and "*shaves his beard*". Also, our prediction on the highlight is more accurate and stable. More visualizations of the prediction results are listed in the supplementary material.

Visualization on Attention Map To show generalizability, we visualize multi-modality attention maps in Figure 5. We show the average heatmap of all layers and word tokens. SE-DETR accurately highlights the visual region corresponding to the provided queries, such as "*yellow mask*" shown in the upper video and

the "*A man dressed in a black suit*", "*takes notes*" in the lower videos. This indicates SE-DETR is generalizable to aggregate the visual features that meet query semantics. In supplementary materials, we list more attention maps of different queries.

Visualization on Semantics Weight To show distinguishability, we illustrate the layer-wise average weight $\mathcal{W}_{V_Q}$ in Figure 6 on QVHighlight *val* split. [SOS], [EOS], and punctuations are omitted. The blue bar means the weight is higher than the average, the red bar means the weight is lower than the average. The bar chart plots the difference in weight and the average. $\mathcal{W}_{V_Q}$ modulates the semantics by temporal sparsity, emphasizing sparse but critical semantics. For instance, the upper video emphasizes prominent semantics like "*yellow blanket*". In the lower video, our method can locate the correct moment with a rare word "*waterfall*" (0.2% queries in training set), which indicates proxies associate the common concepts with these rare vocabularies that can improve the generalization. More visualizations of the semantics weight are illustrated in the supplementary material.

5 Conclusion

In this work, we propose SE-DETR, a unified framework for Video Temporal Grounding that explicitly models semantic properties for effective video-language alignment. Our method introduces explicit mechanisms to capture both semantic generalizability across videos and semantic distinguishability within videos. Specifically, we design Semantics-Proxied Alignment (SPA) to align multimodal features through learnable semantic proxies, enabling concept-level generalization across different lexical expressions. We further introduce Temporal Sparsity Modulation (TSM) to quantify the temporal sparsity of semantic components and emphasize informative yet sparse concepts. Extensive experiments demonstrate that SE-DETR achieves competitive performance in various VTG tasks.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 72192821), the National Natural Science Foundation of China (No. 62472282), the Fundamental Research Funds for the Central Universities (project number: YG2023QNA35), YuCaiKe [2023] Project (Project Number: 231111310300). All faces in demonstrated frames sampled in public benchmark datasets are masked to avoid portrait infringement.

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017)

2. Badamdorj, T., Rochan, M., Wang, Y., Cheng, L.: Joint visual and audio learning for video highlight detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8127–8137 (2021)
3. Escorcia, V., Soldan, M., Sivic, J., Ghanem, B., Russell, B.: Temporal localization of moments in video collections with natural language. arXiv preprint arXiv:1907.12763 (2019)
4. Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting clip's image representation via text-based decomposition. arXiv preprint arXiv:2310.05916 (2023)
5. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
6. Gygli, M., Song, Y., Cao, L.: Video2gif: Automatic generation of animated gifs from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1001–1009 (2016)
7. Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. JSTOR: Applied Statistics **28**(1), 100–108 (1979)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
9. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
10. Hong, F.T., Huang, X., Li, W.H., Zheng, W.S.: Mini-net: Multiple instance ranking network for video highlight detection. In: European Conference on Computer Vision. pp. 345–360. Springer (2020)
11. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14271–14280 (2024)
12. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
13. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems **34**, 11846–11858 (2021)
14. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvr: A large-scale dataset for video-subtitle moment retrieval. In: European Conference on Computer Vision. pp. 447–463. Springer (2020)
15. Lepori, M.A., Tartaglini, A.R., Vong, W.K., Serre, T., Lake, B.M., Pavlick, E.: Beyond the doors of perception: Vision transformers represent relations between objects. Advances in Neural Information Processing Systems **37**, 131503–131544 (2024)
16. Li, J., Huang, Y., Wu, M., Zhang, B., Ji, X., Zhang, C.: CLIP-SP: vision-language model with adaptive prompting for scene parsing. Computational Visual Media **10**(4), 741–752 (2024)
17. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)

19. Liu, S., Quan, W., Zhou, M., Chen, S., Kang, J., Zhao, Z., Yan, K., Chen, C., Yan, D.M.: M2hf: Multi-branch multi-modal hybrid fusion for text-video retrieval. Computational Visual Media **12**(2), 449–471 (2026)
20. Liu, W., Mei, T., Zhang, Y., Che, C., Luo, J.: Multi-task deep visual-semantic embedding for video thumbnail selection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3707–3715 (2015)
21. Liu, Y., He, J., Li, W., Kim, J., Wei, D., Pfister, H., Chen, C.W.: R2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In: European Conference on Computer Vision. pp. 421–438. Springer (2024)
22. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3042–3051 (2022)
23. 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: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 4514–4523 (2025)
24. Ma, J., Huang, P.Y., Xie, S., Li, S.W., Zettlemoyer, L., Chang, S.F., Yih, W.T., Xu, H.: Mode: Clip data experts via clustering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26354–26363 (2024)
25. Ma, K., Zang, X., Feng, Z., Fang, H., Ban, C., Wei, Y., He, Z., Li, Y., Sun, H.: Llavilo: Boosting video moment retrieval via adapter-based multimodal modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2798–2803 (2023)
26. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 202–211 (2017)
27. Moon, W., Hyun, S., Lee, S., Heo, J.P.: Correlation-guided query-dependency calibration for video temporal grounding. arXiv preprint arXiv:2311.08835 (2023)
28. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23023–23033 (2023)
29. Ning, M., Zhu, B., Xie, Y., Lin, B., Cui, J., Yuan, L., Chen, D., Yuan, L.: Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. Computational Visual Media **12**(1), 71–84 (2026)
30. Pujol-Perich, D., Escalera, S., Clapés, A.: Sparse-dense side-tuner for efficient video temporal grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21515–21524 (2025)
31. 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: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
32. Ran, R., Wei, J., Cai, X., Guan, X., Zou, J., Yang, Y., Shen, H.T.: Cdtr: Semantic alignment for video moment retrieval using concept decomposition transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6684–6692 (2025)
33. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics **1**, 25–36 (2013)

34. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14313–14323 (2024)
35. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27**(3), 379–423 (1948)
36. Song, Y., Redi, M., Vallmitjana, J., Jaimes, A.: To click or not to click: Automatic selection of beautiful thumbnails from videos. In: Proceedings of the 25th ACM international on conference on information and knowledge management. pp. 659–668 (2016)
37. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5179–5187 (2015)
38. Sun, H., Zhou, M., Chen, W., Xie, W.: Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection (2024)
39. Sun, L., Cao, J., Xie, J., Jiang, X., Pang, Y.: Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23199–23209 (2025)
40. Sun, M., Farhadi, A., Seitz, S.: Ranking domain-specific highlights by analyzing edited videos. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 787–802. Springer (2014)
41. 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: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7473–7481 (2025)
42. Wang, L., Liu, D., Puri, R., Metaxas, D.N.: Learning trailer moments in full-length movies with co-contrastive attention. In: European Conference on Computer Vision. pp. 300–316. Springer (2020)
43. Xiong, B., Kalantidis, Y., Ghadiyaram, D., Grauman, K.: Less is more: Learning highlight detection from video duration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1258–1267 (2019)
44. Xu, M., Wang, H., Ni, B., Zhu, R., Sun, Z., Wang, C.: Cross-category video highlight detection via set-based learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7970–7979 (2021)
45. Yang, H., Wang, B., Lin, S., Wipf, D., Guo, M., Guo, B.: Unsupervised extraction of video highlights via robust recurrent auto-encoders. In: Proceedings of the IEEE international conference on computer vision. pp. 4633–4641 (2015)
46. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7950–7959 (2021)
47. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6543–6554. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.acl-main.585
48. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: European conference on computer vision. pp. 766–782. Springer (2016)

49. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12870–12877 (2020)
50. Zhao, K., Yuan, W., Wang, Z., Li, G., Zhu, X., Fan, D.P., Zeng, D.: Open-vocabulary camouflaged object segmentation with cascaded vision language models. Computational Visual Media **12**(2), 473–492 (2026)