Action-and-object Aware Alignment for Partially Relevant Video Retrieval

Chuanshen Chen[1,2*], **Kai Zhou**[1*], **Zhiquan Wen**[1†], **Zeng You**[1,2],
Yirui Li[1], **Tianhang Xiang**[1], **Mingkui Tan**[1,2]

[1]South China University of Technology
[2]Peng Cheng Laboratory
{chuanshenc888, kayjoe0723, sewenzhiquan, zengyou.yz, jac.bigwood}@gmail.com,
sexiangtianhang@mail.scut.edu.cn, mingkuitan@scut.edu.cn

Abstract

Partially Relevant Video Retrieval (PRVR) aims to retrieve untrimmed videos containing relevant moments for a given text query. This task is extremely challenging, as untrimmed videos often include numerous actions and objects unrelated to the query. However, existing methods usually struggle with fine-grained action-object modeling, limiting their retrieval performance. To tackle this challenge, we introduce Action-and-object Aware Alignment for Partially Relevant Video Retrieval (A^3PRVR), a dual-branch framework designed to enhance retrieval by improving the modeling of action-object relationships. Specifically, we propose a Query-specific Deformable Temporal Attention (Q-DTA) module to effectively capture action-relevant object information in video features, while filtering out irrelevant content. Additionally, we propose an action-and-object aware alignment module to enable fine-grained textual understanding and video-text alignment. It uses action- and object-aware contrastive losses to enhance the model's sensitivity to action-object distinctions in the text query. Compared to state-of-the-art methods, A^3PRVR achieves an average relative gain of 6.5% in SumR across the Charades-STA, ActivityNet-Caption, and TVR datasets.

Code — https://github.com/chuanshen-chen/A3PRVR

1 Introduction

With the rapid growth of online media, millions of videos are uploaded daily, driving demand for effective video retrieval from large corpora. Text-to-video retrieval (T2VR) has gained research interest due to applications like content recommendation (Hanu et al. 2022) and surveillance analysis (Irene, Prakash, and Uthariaraj 2024; Yuan et al. 2024). Most T2VR methods (Bain et al. 2021; Chen et al. 2020; Jin et al. 2021; Wang et al. 2022b; Li et al. 2020; Song et al. 2021) assume videos are pre-trimmed, short, and fully relevant to the query. However, real-world videos are usually untrimmed, with complex scenes where only parts are query-relevant. To handle this, Dong et al. (Dong et al. 2022a) proposed Partially Relevant Video Retrieval (PRVR), focusing on retrieving untrimmed videos that contain moments relevant to a given text query. The challenge of this task is that untrimmed videos typically contain numerous actions and objects unrelated to the text query. As a result, the PRVR task requires models to develop a fine-grained perception and understanding of actions, action-relevant objects, and their nuanced relationships, enabling precise retrieval in complex scenes.

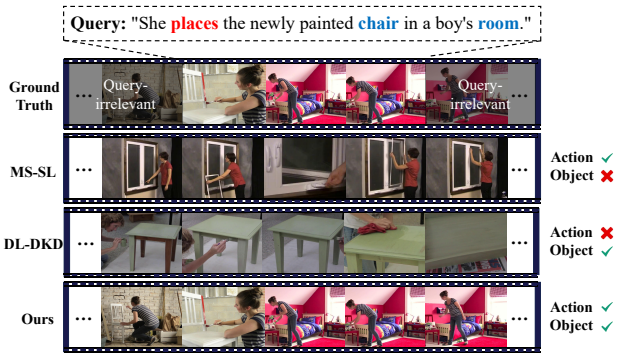


Figure 1: Comparison between PRVR methods. MS-SL (Dong et al. 2022a) captures only the action "places" without identifying the correct object ("chair"), while DL-DKD (Dong et al. 2023) focuses on the "chair" but misses the action "places". Our method focuses on action and object in the query, thus accurately retrieving the video.

Existing partially relevant video retrieval methods (Dong et al. 2022a; Jiang et al. 2023; Wang et al. 2024b; Dong et al. 2023) rely on contrastive learning to align visual and textual features, facilitating the capture of partial relationships between text and video. The perception of actions and objects differs across different methods. One group of methods (Dong et al. 2022a; Jiang et al. 2023; Wang et al. 2024b; Cho et al. 2025; Jun et al. 2025) rely on pre-trained action recognition models (e.g., I3D (Carreira and Zisserman 2017)) to extract video features. While effective in capturing action-related information, these methods struggle to capture detailed object information fully. To fill this gap, (Dong et al. 2023; Song et al. 2025) uses large-scale vision-language models like CLIP (Radford et al. 2021) to guide object-focused learning. This method employs separate branches to extract both action patterns and object-level information. However, since it does not model the relation-

*These authors contributed equally.
†Corresponding Author.

ship between actions and interacting objects, this method may focus on action-irrelevant object information, leading to potential interference. The visualization results in Fig. 1 also support this. For example, for the query "She places the newly painted chair in a boy's room.", MS-SL (Dong et al. 2022a) retrieves a Top-1 video with the action "places" but a "window" instead of the "chair" in the query, capturing the action while overlooking the object. Conversely, the Top-1 result of DL-DKD (Dong et al. 2023) includes a "chair" but misses the action "places", focusing on an action-irrelevant object rather than the correct action-object relationship in the query. To sum up, existing methods lack a fine-grained visual understanding of actions, objects, and their relationships, which limits retrieval performance.

Beyond the visual understanding issues, most current methods (Dong et al. 2022a; Jiang et al. 2023; Wang et al. 2024b; Dong et al. 2023; Song et al. 2025) also struggle with the text query side. They typically form positive and negative sample pairs by directly matching text queries with video segments within the same batch. However, the noticeable action and object distinctions between text queries lead the model to rely on obvious semantic cues, rather than developing a fine-grained textual understanding of actions and relevant objects. This approach fails to establish the fine-grained alignment between videos and text queries, which is crucial for the PRVR task.

In this paper, we propose a framework called Action-and-object Aware Alignment for Partially Relevant Video Retrieval (A^{3}PRVR), enabling fine-grained action-and-object aware video-text alignment for precise retrieval. To effectively capture the relationships between actions and relevant objects, we propose a Query-specific Deformable Temporal Attention (Q-DTA) module with dual-branch video feature extraction. One branch extracts action-focused features (e.g., using I3D (Carreira and Zisserman 2017)), while the other extracts object-focused features (e.g., using CLIP (Radford et al. 2021)). Our deformable attention mechanism allows each video segment to attend to nearby but flexible segments rather than all segments (Vaswani 2017), as temporally distant ones often lack meaningful action-object relationships. Additionally, our query-specific approach allows each query video segment to focus on its own action-object relationships, rather than on shared attention regions (Xia et al. 2022). Moreover, to achieve fine-grained textual understanding and video-text alignment, we propose an action-and-object aware alignment module. This module integrates temporal-level video-text matching with action- and object-aware contrastive losses in their respective alignment branches, where hard negative text samples are generated by replacing actions (verbs) or objects (nouns) in the text queries. This enables fine-grained video-text alignment in an action- and object-aware manner.

Contributions: 1) We propose A^{3}PRVR, a dual-branch framework that enables fine-grained action-and-object aware alignment for precise retrieval. Extensive evaluations on three benchmarks show A^{3}PRVR achieves an average relative gain of **6.5%** in SumR over SOTA methods. **2)** We propose the Q-DTA module to effectively capture action-object relationships in video features. Empirical results demonstrate that capturing action-relevant object information is crucial for PRVR, as shown in **Tab. 4**. **3)** We propose the action-and-object-aware alignment module to enable fine-grained textual understanding and video-text alignment. Beyond quantitative results, we provide visualizations in the supplementary[1] showing how this module enhances sentence feature sensitivity to action–object distinctions.

2 Related Works

Partially Relevant Video Retrieval (PRVR). PRVR aims to retrieve untrimmed videos containing relevant moments for a given query. Existing methods (Dong et al. 2022a; Jiang et al. 2023; Wang et al. 2024b; Jun et al. 2025; Cho et al. 2025; Zhang et al. 2025; Li et al. 2025) typically use pre-trained action recognition models like I3D (Carreira and Zisserman 2017) to extract video features, while AMDNet (Song et al. 2025) uses CLIP (Radford et al. 2021) to extract object features. DL-DKD (Dong et al. 2023), leverages I3D and CLIP for action-focused and object-focused features but struggles to model the relationship between actions and relevant objects. Furthermore, the large differences in action and object content between positive and negative sample pairs hinder fine-grained alignment. Our A^{3}PRVR overcomes these limitations by enhancing a fine-grained understanding of actions, objects, and their relationships in both visual and textual domains.

Deformable Attention. Introduced by DCN (Dai et al. 2017; Zhu et al. 2019), adjusts the sampling locations of convolutional kernels, enhancing adaptability to varying scales and shapes. Deformable Attention (Xia et al. 2022) extends this idea to Transformers, enabling flexible attention regions and reducing computational complexity. In video tasks, Deformable Video Attention (Wang and Torresani 2022) operates in 3D space ($T \times H \times W$), but is less effective for downstream tasks with pre-trained, frozen video features, as the spatial dimensions are discarded during feature extraction, leaving only the temporal dimension (T). To address this, we propose a Query-specific Deformable Attention (Q-DTA) module that operates along the temporal dimension (T). Unlike prior methods (Xia et al. 2022; Wang and Torresani 2022) that use shared deformable points, our module selects deformable points independently for each query, improving attention precision and reducing noise. This approach is well-suited for PRVR tasks with pre-trained encoders, retaining the benefits of deformable mechanisms for temporal modeling.

Action and Object Understanding. In video-language alignment tasks, understanding actions and objects is essential. To improve action comprehension, (Wang et al. 2023; Momeni et al. 2023) propose action-aware losses during pre-training. For object understanding, (Wang et al. 2022a) enhances object awareness during pre-training, while (Li et al. 2022; Han et al. 2022) uses object detection results for downstream tasks. In text-to-video retrieval, methods (Li et al. 2023; Song, Chen, and Jiang 2023) prioritize action- and object-related information through spatio-temporal

[1] https://github.com/chuanshen-chen/A3PRVR

modeling to deepen video-text relations. However, for partially relevant video retrieval, where untrimmed videos contain irrelevant moments, extensive spatio-temporal modeling becomes inefficient. We focus more on the temporal aspect to identify relevant moments matching the text queries. To address this, we propose a novel action-and-object aware alignment module that combines temporal-level video-text matching with action- and object-aware contrastive losses.

3 Method

Problem Formulation and Method Overview

Given a text query, partially relevant video retrieval aims to retrieve untrimmed videos containing moments semantically relevant to the query. Let the video corpus, consisting of N_V videos, be denoted as $\mathcal{V} = \{V_1, V_2, \ldots, V_{N_V}\}$, where each video V_i consists of T segments, represented as $V_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,T}\}$ with $v_{i,j}$ being the j-th segment of the i-th video. Let the set of text queries, consisting of N_U queries, be denoted as $\mathcal{U} = \{U_1, U_2, \ldots, U_{N_U}\}$, where each text query U_i with R tokens, denotes as $U_i = \{u_{i,1}, u_{i,2}, \ldots, u_{i,R}\}$ with $u_{i,j}$ being the j-th token of i-th text query. To simplify notation, and where no ambiguity arises, we use V and v_j to denote a video and its j-th segment, respectively. Similarly, U and u_j are used to represent a text query and its j-th token. The task of PRVR is extremely challenging, as untrimmed videos often include numerous actions and objects unrelated to the query. However, existing PRVR methods usually struggle with fine-grained action-object modeling, limiting their retrieval performance.

To tackle this, we propose a framework called Action-and-object Aware Alignment for Partially Relevant Video Retrieval (A^{3}PRVR) to model action-and-object aware video-text alignment with two parts. **First,** we construct a dual-branch framework to extract visual features, with the two branches utilizing I3D-based (Carreira and Zisserman 2017) and CLIP-based (Radford et al. 2021) visual encoders to capture action-aware and object-aware features, respectively. We then introduce a *Query-specific Deformable Temporal Attention (Q-DTA)* module, enabling action-centric queries to focus on relevant object features through deformable attention, thereby enhancing action-object interaction understanding in video features. **Second,** we propose an *action-and-object aware alignment module* that integrates temporal-level video-text matching with action- and object-aware contrastive losses for respective alignment branches. The framework of our A^{3}PRVR is illustrated in Fig. 2.

Query-specific Deformable Temporal Attention

We first present a dual-branch video feature extraction framework using I3D- and CLIP-based encoders to extract action- and object-focused features, respectively. Subsequently, we propose a *Query-specific Deformable Temporal Attention (Q-DTA)* module that enhances the action-focused features by integrating action-relevant object features, leading to a richer visual representation.

Dual-branch video feature extraction. Given an untrimmed video $V = \{v_1, v_2, \ldots, v_T\}$, we extract action-focused features $A \in \mathbb{R}^{T\times d}$ and object-focused features $O \in \mathbb{R}^{T\times d}$ using the I3D encoder F_I and the CLIP visual encoder F_C, respectively. Each uses a one-layer Transformer and a ReLU-activated FC layer for dimension reduction, denoted as ψ_I and ψ_C. Specifically:

$$\begin{aligned} A &= [a_1; a_2; \ldots; a_T] = \text{Transformer}(\psi_I(F_I(V))), \\ O &= [o_1; o_2; \ldots; o_T] = \text{Transformer}(\psi_C(F_C(V))), \end{aligned} \tag{1}$$

where $F_I(V) = [F_I(v_1); F_I(v_2); \ldots; F_I(v_T)]$ and $F_C(V) = [F_C(f_1); F_C(f_2); \ldots; F_C(f_T)]$, with f_j as the middle frame of segment v_j. Here, a_j and o_j denote the action- and object-focused features for segment v_j.

Query-specific deformable temporal attention module. Intuitively, feature interaction between the two branches naturally suits cross-attention (Vaswani 2017). However, in PRVR, a text is generally relevant to specific contiguous segments of a video, with weak relationships between temporally distant segments. Consequently, applying standard attention calculations to segment features that are temporally distant may introduce noise instead of yielding useful information. To allow a segment to attend to adjacent yet deformable segments, we propose a Query-specific deformable temporal attention mechanism. Unlike deformable attention in DAT (Xia et al. 2022) learns shared deformed points for all queries, our method learns query-specific deformed temporal points. This better suits PRVR, as each query segment often represents distinct actions and requires its own action-relevant objects to model diverse temporal relations. Ablation results in Tab. 3 confirm its effectiveness.

To illustrate our Query-specific Deformable Temporal Attention, we use action-focused segment features $A = [a_1; a_2; \ldots; a_T]$ as input of queries (Q) and object-focused segment features $O = [o_1; o_2; \ldots; o_T]$ as input of keys (K) and values (V). The query matrix is computed as $A_q = AW_q$, with $W_q \in \mathbb{R}^{d\times d}$ as a learnable projection. The calculation of keys and values requires a sampling operation for object-focused segment features. We need to generate deformable sampling points to sample features O. Given a uniform grid of points $P = [p_1, p_2, \ldots, p_T] \in \mathbb{R}^T$ as references, the values of these reference points are linearly spaced 1D coordinates $\{1, 2, \ldots, T\}$. The values of the reference points can be regarded as the indices of queries. To obtain multiple deformed points for each reference point, offsets are added to each reference. These offsets are generated by feeding A_q into a lightweight sub-network $\theta_{\text{offset}}(\cdot)$:

$$\Delta P = [\Delta p_1; \Delta p_2; \ldots; \Delta p_T] = \theta_{\text{offset}}(A_q), \tag{2}$$

where $\Delta p_j \in \mathbb{R}^n$ denotes n offsets of reference point p_j, and $\Delta P \in \mathbb{R}^{T\times n}$ is an offsets matrix. The deformed points for each reference point can be expressed as $\delta_j = \text{Vec}(p_j) + \Delta p_j$, where $\delta_j \in \mathbb{R}^n$ represents n deformed sampling points of p_j, and $\text{Vec}(\cdot)$ denotes the vectorization operation. The object-focused features are then sampled at the locations of the deformed points, expressed as $\hat{o}_j = \Phi(O, \delta_j)$, where $\hat{o}_j \in \mathbb{R}^{n\times d}$ denotes the sampled features obtained using the deformed points δ_j. We set the sampling function $\Phi(\cdot,\cdot)$ to a linear interpolation to make it differentiable:

$$\begin{aligned} \Phi(O, x) &= \left[(1 - \epsilon(x_i))o_{\lfloor x_i \rfloor} + \epsilon(x_i)o_{\lfloor x_i \rfloor + 1}\right]_{i=1}^{n}, \\ &\text{with} \quad \epsilon(x_i) = x_i - \lfloor x_i \rfloor, \end{aligned} \tag{3}$$

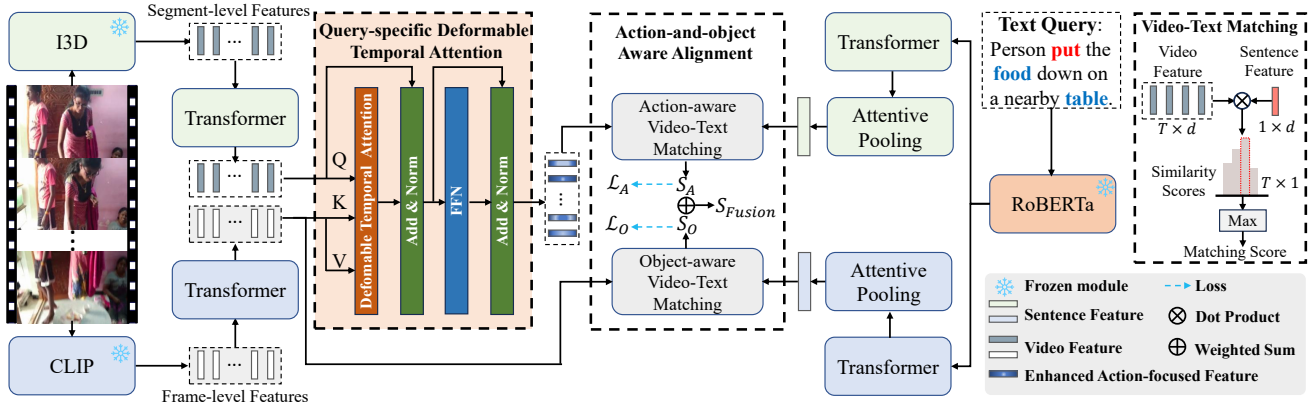


Figure 2: Framework of our A^{3}PRVR. The video encoder uses I3D and CLIP to extract action- and object-focused video features, refined by the Query-specific Deformable Temporal Attention module to model action-object relationships *(Note: "Query-specific" refers to query (Q) in the attention mechanism, not the text query)*. Text features are extracted with RoBERTa (Liu 2019). Fine-grained video-text alignment is achieved in the Action-and-object Aware Alignment module, with video-text matching scores S_A, S_O in the action and object branches, respectively. The final score S_{Fusion} combines S_A and S_O with weighted sum. Action- and object-aware contrastive losses $\mathcal{L}_A, \mathcal{L}_O$ enforce action- and object-level alignment, respectively.

where $x \in \mathbb{R}^n$ is an indices vector, and $x_i \in \mathbb{R}$ denotes the i-th element of x. The overall sampled features, keys, and values are obtained as follows:

$$\begin{aligned} \hat{O} &= [\hat{o}_1; \ldots; \hat{o}_T] = [\Phi(O, \delta_1); \ldots; \Phi(O, \delta_T)], \\ \hat{O}_k &= \hat{O}W_k, \hat{O}_v = \hat{O}W_v, \end{aligned} \tag{4}$$

where $\hat{O}, \hat{O}_k, \hat{O}_v \in \mathbb{R}^{(T\times n)\times d}$ are the sampled features, keys, and values, and $W_k, W_v \in \mathbb{R}^{d\times d}$ are learnable projections. Our deformable attention is computed as:

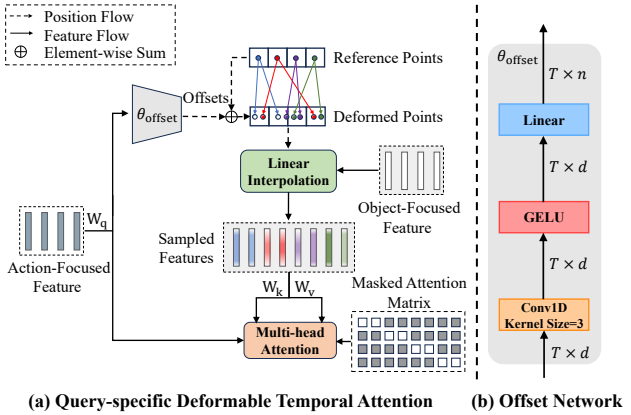
$$\begin{aligned} \text{Att}(A_q, \hat{O}_k, \hat{O}_v, M) &= \text{softmax}(\frac{A_q \hat{O}_k^\top}{\sqrt{d}} + M)\hat{O}_v, \\ \text{with} \quad M_{ij} &= \begin{cases} 0, & \text{if } (i-1)n < j \le in, \\ -\infty, & \text{otherwise}, \end{cases} \end{aligned} \tag{5}$$

where Att denotes the attention operation, and $M \in \mathbb{R}^{T\times(T\times n)}$ is a mask matrix with $M_{ij} = -\infty$ indicating masked positions and $M_{ij} = 0$ indicating active ones. The mask ensures each query attends only to its deformed keys. To reduce computation, we avoid computing the full attention matrix. Our attention has complexity $\mathcal{O}(Tn)$ (with $n \ll T$), significantly lower than the standard $\mathcal{O}(T^2)$.

Following (Vaswani 2017), we apply the multi-head attention (MHA) mechanism, denoted as $\text{MHA}(A_q, \hat{O}_k, \hat{O}_v, M)$, with details omitted for brevity. Our deformable temporal cross-attention module is formulated as follows:

$$\begin{aligned} A' &= \text{LN}(\text{MHA}(A_q, \hat{O}_k, \hat{O}_v, M) + A), \\ A'' &= \text{LN}(\text{FFN}(A') + A'), \end{aligned} \tag{6}$$

where LN is Layer Normalization (Ba 2016), and FFN is a two-layer Feed-Forward Network with ReLU activation. Here, $A', A'' \in \mathbb{R}^{T\times d}$, where $A'' = \text{Q-DTA}(A, O)$ is the enhanced action-focused feature from our Query-specific Deformable Temporal Attention module. Alternatively, object-focused features can serve as queries (Q) and action-focused features as keys/values to obtain $O'' = \text{Q-DTA}(O, A)$. However, replacing O with O'' for subsequent video-text matching does not yield optimal performance (see Tab. 4). For each text query, actions typically involve specific objects (e.g., a person using a tool), while objects in the background are not always related to the action (e.g., scene objects). Forcing all object information to interact with action features may introduce noise. Thus, we use A'' to replace A, but not O'' to replace O, for video-text matching.



(a) Query-specific Deformable Temporal Attention (b) Offset Network

Figure 3: Framework of the Query-specific Deformable Temporal Attention (Q-DTA) module. (a) Q-DTA lets action-centric queries attend to relevant object features via deformable attention, with the Masked Attention Matrix ensuring each query focuses on its point. (b) The Offset Network generates query-specific offsets using 1D convolution, GELU, and a linear layer for flexible temporal focus.

Offset generation and grouping. A sub-network generates offsets by consuming query features and producing offset values for reference points. The sub-network consists of two linear layers with a nonlinear activation, as shown in Fig.

3(b). To stabilize the training process, we normalize the values of reference points to the range $[-1, +1]$. We then apply a corresponding scaling to offsets as $\Delta P \leftarrow \gamma \tanh(\Delta P)$, where γ is a scale factor. We implement index mapping operations during sampling features, which are omitted here for brevity. To encourage diversity, we follow DAT (Xia et al. 2022) by splitting features into g groups, each using a shared sub-network to generate offsets. The number of attention heads h is a multiple of g, ensuring multiple heads are assigned to one group of deformed keys and values.

Action-and-object Aware Alignment

We first present a dual-branch sentence feature extraction framework that encodes the text query using RoBERTa (Liu 2019). Specifically, we extract two sentence features: one corresponding to the action-focused video feature and the other to the object-focused video feature, for subsequent video-text matching. We then propose an action-and-object aware alignment module, which incorporates action- and object-aware contrastive losses for each branch to enable more precise and fine-grained video-text alignment.

Dual-branch sentence feature extraction. Given a text query $U = \{u_1, \ldots, u_R\}$, we extract action-sensitive and object-sensitive features $z^A, z^O \in \mathbb{R}^{1\times d}$, corresponding to the action- and object-focused video features, using RoBERTa (Liu 2019) followed by a one-layer Transformer:

$$\begin{aligned} z^A &= \text{AttPool}(\text{Transformer}(\psi_A(F_{Ro}(U)))), \\ z^O &= \text{AttPool}(\text{Transformer}(\psi_O(F_{Ro}(U)))), \end{aligned} \tag{7}$$

where ψ_A, ψ_O are fully connected layers with ReLU activation. $\psi_A(F_R(U)), \psi_O(F_R(U)) \in \mathbb{R}^{R\times d}$ denote the token-level features. $\text{AttPool}(\cdot)$ denotes the attentive pooling operation, formulated as $\text{AttPool}(y) = \text{softmax}(wy^\top)y$, where $y \in \mathbb{R}^{R\times d}$ represents a 2D feature matrix, and $w \in \mathbb{R}^{1\times d}$ is a learnable vector.

Action-and-object aware alignment module. Given a video-text pair (V, U), with enhanced action-focused video feature $A'' = [a''_1; \ldots; a''_T]$, object-focused video feature $O = [o_1; \ldots; o_T]$, action-sensitive sentence feature z^A, and object-sensitive sentence feature z^O, we compute the video-text matching scores, formulated as:

$$\begin{aligned} S_A(V,U) &= \text{Max}([\cos(a''_1, z^A), \ldots, \cos(a''_T, z^A)]), \\ S_O(V,U) &= \text{Max}([\cos(o_1, z^O), \ldots, \cos(o_T, z^O)]), \end{aligned} \tag{8}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity. Max denotes max-pooling, and S_A, S_O are the action and object branch matching scores, respectively. Without training, video-text matching scores are low due to feature misalignment. Aligning them in an action-and-object-aware manner is challenging.

To tackle this, we propose action-aware and object-aware contrastive losses for respective branches. While the distinct pre-trained visual encoders emphasize actions or objects, RoBERTa lacks inherent sensitivity to either. We address this by constructing action-aware and object-aware negative video-text pairs for contrastive learning. As shown in Fig. 4, we use the Stanford Stanza tool (Qi et al. 2020) to substitute actions (verbs) or objects (nouns) in a text query, generating action-aware or object-aware negative text samples. We extract verbs or nouns from the text query and replace them with others from the dataset to create negative text samples, excluding subject-related terms like "person" and "woman". Each negative is formed by replacing a single verb or noun, enabling the model to learn fine-grained contextual differences. The action-aware contrastive loss $\mathcal{L}_A$ is defined as:

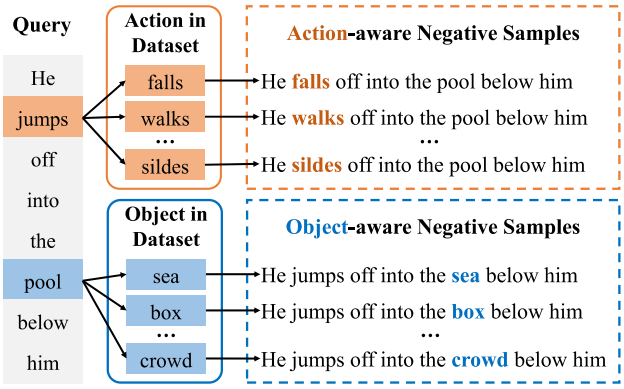


Figure 4: Hard negative sample generation for $\mathcal{L}_A$ and $\mathcal{L}_O$. We generate action-aware and object-aware negative samples by substituting verbs (e.g., "jumps" to "falls") and nouns (e.g., "pool" to "sea") in a text query, enabling the model to learn fine-grained distinctions.

$$\begin{aligned} \mathcal{L}_A &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_A(V_i, U_i)/\tau)}{\exp(S_A(V_i, U_i)/\tau) + \Sigma_{i,A}}, \\ \text{with} \quad \Sigma_{i,A} &= \sum_{U_i^A \in \mathcal{N}_{i,A}} \exp(S_A(V_i, U_i^A)/\tau), \end{aligned} \tag{9}$$

where B is the batch size, τ is a learnable temperature parameter, and $\mathcal{N}_{i,A}$ represents all action-aware negative text samples of U_i. $U_i^A \in \mathcal{N}_{i,A}$ indicates an action-aware negative sample of U_i. Similarly, the object-aware contrastive loss $\mathcal{L}_O$ shares the same structure, with S_O, $\mathcal{N}_{i,O}$, and $U_i^O \in \mathcal{N}_{i,O}$ denoting the corresponding object-aware terms.

Overall Optimization

Beyond action-aware and object-aware contrastive losses, it is crucial for the model to capture diverse overall semantics in text queries for effective video-text alignment. Following MS-SL (Dong et al. 2022a), we also jointly use the triplet ranking loss (Dong et al. 2021; Faghri et al. 2017) and InfoNCE loss (Miech et al. 2020; Zhang et al. 2021a), which are widely used in retrieval tasks. Given a positive video-text pair (V_i, U_i) in a batch, all other video and text samples serve as negative examples. We denote the triplet ranking losses for the action and object branches as $\mathcal{L}_A^t$ and $\mathcal{L}_O^t$, and the InfoNCE losses as $\mathcal{L}_A^I$ and $\mathcal{L}_O^I$. *Detailed formulas are provided in the* ***Supplementary***. Finally, the model is trained by minimizing the overall loss function defined as follows:

$$\mathcal{L} = \frac{\lambda(\mathcal{L}_A + \mathcal{L}_O) + \eta(\mathcal{L}_A^t + \mathcal{L}_O^t) + \sigma(\mathcal{L}_A^I + \mathcal{L}_O^I)}{2}, \tag{10}$$

where λ, η, σ are positive hyper-parameters, respectively.

Method	Visual Backbone	Text Backbone	Charades-STA					ActivityNet-Caption					TVR				
			R@1	R@5	R@10	R@100	SumR	R@1	R@5	R@10	R@100	SumR	R@1	R@5	R@10	R@100	SumR
CE	I	R	1.3	4.5	7.3	36.0	49.1	5.5	19.1	29.9	71.1	125.6	3.7	12.8	20.1	64.5	101.1
DE++	I	R	1.7	5.6	9.6	37.1	54.1	5.3	18.4	29.2	68.0	121.0	8.8	21.9	30.2	67.4	128.3
RIVRL	I	R	1.6	5.6	9.4	37.7	54.3	5.2	18.0	28.2	66.4	117.8	9.4	23.4	32.2	70.6	135.6
XML	I	R	1.6	6.0	10.1	46.9	64.6	5.3	19.4	30.6	73.1	128.4	10.0	26.5	37.3	81.3	155.1
ReLoCLNet	I	R	1.2	5.4	10.0	45.6	62.3	5.7	18.9	30.0	72.0	126.6	10.0	26.5	37.3	81.3	155.1
CONQUER	I	R	1.8	6.3	10.3	47.5	66.0	6.5	20.4	31.8	74.3	133.1	11.0	28.9	39.6	81.3	160.8
MS-SL	I	R	1.8	7.1	11.8	47.7	68.4	7.1	22.5	34.7	75.8	140.1	13.5	32.1	43.4	83.4	172.4
GMMFormer	I	R	2.1	7.8	12.5	50.6	72.9	8.3	24.9	36.7	76.1	146.0	13.9	33.3	44.5	84.9	176.6
PEAN	I	R	2.7	8.1	13.5	50.3	74.7	7.4	23.0	35.5	75.9	141.8	13.5	32.8	44.1	83.9	174.2
HLFormer	I	R	2.6	8.5	13.7	54.0	78.7	8.7	27.1	40.1	79.0	154.9	15.7	37.1	48.5	86.4	187.7
ARL	I	R	-	-	-	-	-	8.3	24.6	37.4	78.0	148.3	15.6	36.3	47.7	86.3	185.9
MGAKD	I	R	-	-	-	-	-	7.9	25.7	38.3	77.8	149.6	16.0	37.8	49.2	87.5	190.5
Jun *et al.*	I	R	-	-	-	-	-	9.1	27.3	40.4	79.8	156.6	17.4	39.7	51.4	87.9	196.4
CLIP4Clip	CV	CT	1.8	6.5	10.9	44.2	63.4	5.9	19.3	30.4	71.6	127.3	9.9	24.3	34.3	72.5	141.0
InternVideo2	CV	Llama-2	1.9	7.5	12.3	49.2	70.9	7.5	23.4	36.1	76.5	143.5	13.8	32.9	44.4	84.2	175.3
MS-SL	CV	CT	-	-	-	-	-	11.3	30.7	43.5	81.7	167.2	17.8	39.4	50.7	88.2	196.1
GMMFormer	CV	CT	-	-	-	-	-	10.6	29.5	42.6	79.7	162.4	18.1	40.2	51.7	89.0	199.1
AMDNet	CV	CT	-	-	-	-	-	12.3	32.5	45.9	82.1	172.8	19.7	42.4	54.1	88.9	205.1
DL-DKD	I+CV	R	-	-	-	-	-	8.0	25.0	37.5	77.1	147.6	14.4	34.9	45.8	84.9	180.0
MS-SL*	I+CV	R	2.3	8.4	13.7	52.3	76.8	10.7	31.5	44.8	82.1	169.1	25.4	50.0	61.4	92.1	228.9
GMMFormer*	I+CV	R	2.0	7.4	12.4	51.6	73.4	11.3	31.2	44.5	80.8	167.8	24.9	49.6	61.2	92.3	228.0
AMDNet*	I+CV	R	1.3	6.4	11.3	48.3	67.3	8.3	25.6	37.4	77.3	148.6	17.0	39.0	50.0	87.1	193.0
A^{3}PRVR (Ours)	I+CV	R	**2.9**	**9.6**	**15.5**	**54.3**	**82.3**	<u>11.8</u>	**32.9**	**46.2**	**82.3**	**173.2**	**27.0**	**52.4**	**63.6**	**92.3**	**235.3**

Table 1: Comparison with state-of-the-art methods on the Charades-STA, ActivityNet-Caption, and TVR datasets. The best performance is highlighted in bold, while the second-best of our results is underlined. "-" indicates that the corresponding results are unavailable in the original publication. "*" indicates our reimplementation using dual visual features and RoBERTa-Base text features. In this table, "I", "CV", "CT", and "R" denote the I3D visual encoder, CLIP-B/32 visual encoder, CLIP-B/32 text encoder, and RoBERTa-Base text encoder, respectively.

Model inference. The video-text matching score is the weighted sum of action and object branch scores:

$$S_{Fusion}(V,U) = \omega S_A(V,U) + (1-\omega)S_O(V,U), \quad (11)$$

where $\omega \in [0,1]$ balances the two matching scores. For each text query, videos are ranked by descending matching score.

4 Experiments

Datasets: We evaluate A^{3}PRVR on Charades-STA (Gao et al. 2017), ActivityNet-Caption (Krishna et al. 2017), and TVR (Lei et al. 2020) *(dataset details in the supplementary)*.
Evaluation Setting: Following (Dong et al. 2022a), we use rank-based metrics R@K (K = 1, 5, 10, 100), reporting the percentage of queries with correct results in the top K. Sum of Recalls (SumR) is also used to assess overall performance, with higher values indicating better retrieval.
Implementation Details: Our method follows (Dong et al. 2022a), adopting the same data preprocessing, optimizer, learning rate, number of epochs, batch size, and loss weights η and σ (for triplet ranking and InfoNCE), as well as early stopping. The main difference lies in our use of I3D and CLIP-B/32 as the visual backbone. In the Q-DTA module, we set the number of attention heads to $h = 8$, offset groups to $g = 8$, offsets per query to $n = 8$, and the scale factor to $\gamma = 64$. For the losses $\mathcal{L}_A$ and $\mathcal{L}_O$, we use 12 action- and object-aware negative samples with a loss weight of $\lambda = 0.1$. The final matching scores from both branches are fused with equal weights, using $\omega = 0.5$ in Eq. (11).

Comparison with the State-of-the-Art

We compare our method with state-of-the-art methods in three categories: Text-to-Video Retrieval (T2VR), Video Corpus Moment Retrieval (VCMR, without moment localization), and Partially Relevant Video Retrieval (PRVR). **1) T2VR:** We compare with CE (Liu et al. 2019), DE++ (Dong et al. 2021), RIVRL (Dong et al. 2022b), CLIP4Clip (Luo et al. 2022), and InternVideo2 (Wang et al. 2024a). **2) VCMR:** We include XML (Lei et al. 2020), ReLoCLNet (Zhang et al. 2021b), and CONQUER (Hou, Ngo, and Chan 2021). **3) PRVR:** We evaluate against MS-SL (Dong et al. 2022a), PEAN (Jiang et al. 2023), GMMFormer (Wang et al. 2024b), DL-DKD (Dong et al. 2023), Jun *et al.* (Jun et al. 2025), MGAKD (Zhang et al. 2025), ARL (Cho et al. 2025), HLFormer (Li et al. 2025), and AMDNet (Song et al. 2025).

As shown in Tab. 1, our A^{3}PRVR outperforms existing methods across all datasets. Compared to state-of-the-art PRVR methods HLFormer and AMDNet, A^{3}PRVR achieves an average improvement of **6.5%** in SumR, with relative gains of **+4.6%** (82.3 *vs.* 78.7), **+0.2%** (173.2 *vs.* 172.8), and **+14.7%** (235.3 *vs.* 205.1) on Charades-STA, ActivityNet-Caption and TVR datasets, respectively.

To ensure a fairer comparison, we reimplement classic (MS-SL, GMMFormer) and recent (AMDNet) open-source PRVR methods using the same visual/text backbones as ours. As shown in Tab. 1, AMDNet performs worse than its original results, likely due to its heavier reliance on pre-aligned CLIP visual and text features. While MS-SL per-

Method	R@1	R@5	R@10	R@100	SumR
Action Branch	1.3	5.6	10.0	47.8	64.7
Object Branch	1.2	5.2	8.8	43.7	59.0
A^{3}PRVR-Base	1.9	7.0	12.3	51.9	73.1
A^{3}PRVR-Base w/ Q-DTA	2.3	8.6	13.7	53.7	78.2
A^{3}PRVR-Base w/ $\mathcal{L}_A, \mathcal{L}_O$	2.4	8.7	13.2	51.9	76.3
A^{3}PRVR	**2.9**	**9.6**	**15.5**	**54.3**	**82.3**

Table 2: Effect of the Query-specific Deformable Temporal Attention (Q-DTA) and action- and object-aware contrastive losses $\mathcal{L}_A, \mathcal{L}_O$. "Action Branch" and "Object Branch" represent predictions using only action- and object-focused features for video-text matching, respectively. "A^{3}PRVR-Base" denotes our base model, which fuses matching scores from the above two branches, the same as in Tabs. 3, 4 and 5.

Method	R@1	R@5	R@10	R@100	SumR
A^{3}PRVR-Base	1.9	7.0	12.3	51.9	73.1
+ vanilla cross-attention	2.0	6.9	12.8	53.2	74.9
+ Q-DTA w/ SDP	2.0	7.4	12.4	53.0	74.8
+ Q-DTA (Ours)	**2.3**	**8.6**	**13.7**	**53.7**	**78.2**

Table 3: Effect of different attention mechanisms in our Q-DTA. "Q-DTA w/ SDP" refers to our Q-DTA using shared deformed points (SDP) (Xia et al. 2022) instead of query-specific deformed points.

forms best among them, A^{3}PRVR still achieves improvements of **+7.2%** (82.3 *vs.* 76.8) on Charades-STA, **+2.4%** (173.2 *vs.* 169.1) on ActivityNet-Caption, and **+2.8%** (235.3 *vs.* 228.9) on TVR, demonstrating its effectiveness.

Ablation Studies

We carefully evaluate the effect of key components of our A^{3}PRVR on the most challenging dataset Charades-STA.

Effect of Q-DTA and $\mathcal{L}_A, \mathcal{L}_O$. As shown in the first three rows of Tab. 2, using only action- or object-focused features for video-text matching achieves limited performance, while our A^{3}PRVR-Base achieves superior performance through a dual-branch architecture. Rows 4–5 show that while both the Q-DTA module and $\mathcal{L}_A, \mathcal{L}_O$ enhance the model's performance, with Q-DTA contributing more. Specifically, Q-DTA improves the focus on action-relevant objects in visual features, whereas $\mathcal{L}_A$ and $\mathcal{L}_O$ increase the model's sensitivity to actions (verbs) and objects (nouns) in text. Their combination leads to a 9.2 SumR improvement over A^{3}PRVR-Base.

Effect of different attention mechanisms in our Q-DTA. Tab. 3 shows interacting video features from the action and object branches using vanilla cross-attention (Vaswani 2017) is effective (SumR: 73.1 $\rightarrow$ 74.9). By allowing each video segment to attend to nearby, flexible segments instead of distant ones with weak action-object relationships, our query-specific deformable temporal attention achieves more competitive results (SumR: 73.1 $\rightarrow$ 78.2). However, replacing our query-specific deformable points with shared deformable points for all queries (Xia et al. 2022) harms RPVR, dropping performance (SumR: 78.2 $\rightarrow$ 74.8).

Video Features for Matching	R@1	R@5	R@10	R@100	SumR
A^{3}PRVR-Base: A, O	1.9	7.0	12.3	51.9	73.1
A'', O	**2.3**	**8.6**	**13.7**	**53.7**	**78.2**
A, O''	2.0	7.1	11.7	49.0	69.9
A'', O''	1.9	7.1	11.9	49.2	70.1

Table 4: Effect of enhanced action-/object-focused features.

Method	R@1	R@5	R@10	R@100	SumR
A^{3}PRVR-Base	1.9	7.0	12.3	51.9	73.1
+ $\mathcal{L}_O$	1.9	7.6	**12.6**	52.6	74.7
+ $\mathcal{L}_A$	1.9	7.2	12.1	**53.2**	74.4
+ $\mathcal{L}_A$,$\mathcal{L}_O$ (mismatch)	1.9	7.4	12.6	52.6	74.4
+ $\mathcal{L}_A$,$\mathcal{L}_O$	**2.4**	**8.7**	13.2	51.9	**76.3**

Table 5: Effect of different combinations of action- and object-aware contrastive losses $\mathcal{L}_A, \mathcal{L}_O$ on the Action and Object branches. "mismatch" indicates $\mathcal{L}_A$ applied to the Object Branch and $\mathcal{L}_O$ to the Action Branch.

Effect of enhanced action- and object-focused video features A'', O'' in video-text matching. As mentioned in Sec. 3.2, we obtain enhanced features $A'' = \text{Q-DTA}(A, O)$ and $O'' = \text{Q-DTA}(O, A)$ by using A or O as queries. To assess their impact on video-text matching, we evaluate three combinations: $\{A'', O\}$, $\{A, O''\}$, and $\{A'', O''\}$. As shown in Tab. 4, using the enhanced action-focused feature A'' with O yields the best retrieval, highlighting the importance of capturing action-relevant object information. In contrast, incorporating O'', whether combined with A or A'', leads to a significant performance drop. This is because actions inherently involve certain objects, while not all objects in the video are directly related to the action (e.g., background objects). Forcing all object information to interact with action features may introduce unnecessary noise.

Effect of $\mathcal{L}_A, \mathcal{L}_O$ in dual branches. Tab. 5 shows that applying $\mathcal{L}_A$ to the Action Branch and $\mathcal{L}_O$ to the Object Branch yields similar gains, making sentence features "action-sensitive" and "object-sensitive," respectively. Matching each loss to its branch improves SumR from 73.1 to 76.3, while mismatching them (row 4) degrades performance, confirming I3D's action and CLIP's object strength.

5 Conclusion

We propose A^{3}PRVR, a dual-branch framework that enhances retrieval via better action-object modeling. By combining action- and object-focused video features with our Q-DTA module, A^{3}PRVR focuses on action-relevant objects and reduces the noise from irrelevant objects. Additionally, our action-and-object aware alignment module enables fine-grained textual understanding and video-text alignment. Compared to SOTA PRVR methods, our A^{3}PRVR achieves a 6.5% average SumR gain across three benchmarks.

Acknowledgments

This work was partially supported by the Joint Funds of the National Natural Science Foundation of China (Grant No.U24A20327).

References

Ba, J. L. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1728–1738.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, S.; Zhao, Y.; Jin, Q.; and Wu, Q. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10638–10647.

Cho, C.; Moon, W.; Jun, W.; Jung, M.; and Heo, J. 2025. Ambiguity-Restrained Text-Video Representation Learning for Partially Relevant Video Retrieval. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 2500–2508. AAAI Press.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable Convolutional Networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 764–773. IEEE Computer Society.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022a. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 246–257.

Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; and Wang, M. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4065–4080.

Dong, J.; Wang, Y.; Chen, X.; Qu, X.; Li, X.; He, Y.; and Wang, X. 2022b. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE transactions on circuits and systems for video technology*, 32(8): 5680–5694.

Dong, J.; Zhang, M.; Zhang, Z.; Chen, X.; Liu, D.; Qu, X.; Wang, X.; and Liu, B. 2023. Dual learning with dynamic knowledge distillation for partially relevant video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11302–11312.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, 5267–5275.

Han, N.; Chen, J.; Zhang, H.; Wang, H.; and Chen, H. 2022. Adversarial multi-grained embedding network for cross-modal text-video retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2): 1–23.

Hanu, L.; Thewlis, J.; Asano, Y. M.; and Rupprecht, C. 2022. Vtc: Improving video-text retrieval with user comments. In *European Conference on Computer Vision*, 616–633. Springer.

Hou, Z.; Ngo, C.; and Chan, W. K. 2021. CONQUER: Contextual Query-aware Ranking for Video Corpus Moment Retrieval. In Shen, H. T.; Zhuang, Y.; Smith, J. R.; Yang, Y.; César, P.; Metze, F.; and Prabhakaran, B., eds., *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 3900–3908. ACM.

Irene, S.; Prakash, A. J.; and Uthariaraj, V. R. 2024. Person search over security video surveillance systems using deep learning methods: A review. *Image and Vision Computing*, 104930.

Jiang, X.; Chen, Z.; Xu, X.; Shen, F.; Cao, Z.; and Cai, X. 2023. Progressive Event Alignment Network for Partial Relevant Video Retrieval. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 1973–1978. IEEE.

Jin, W.; Zhao, Z.; Zhang, P.; Zhu, J.; He, X.; and Zhuang, Y. 2021. Hierarchical cross-modal graph consistency learning for video-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1114–1124.

Jun, W.; Moon, W.; Cho, C.; Jung, M.; and Heo, J. 2025. Bridging the Semantic Granularity Gap Between Text and Frame Representations for Partially Relevant Video Retrieval. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 4166–4174. AAAI Press.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 447–463. Springer.

Li, J.; Wang, J.; Tan, C.; Lian, N.; Chen, L.; Wang, Y.; Zhang, M.; Xia, S.-T.; and Chen, B. 2025. Enhancing Partially Relevant Video Retrieval with Hyperbolic Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Li, P.; Xie, C.; Zhao, L.; Xie, H.; Ge, J.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023. Progressive Spatio-Temporal Prototype Matching for Text-Video Retrieval. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 4077–4087. IEEE.

Li, X.; Zhou, F.; Xu, C.; Ji, J.; and Yang, G. 2020. SEA: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 23: 4351–4362.

Li, Y.; Yu, J.; Cai, Z.; and Pan, Y. 2022. Cross-modal Target Retrieval for Tracking by Natural Language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, 4927–4936. IEEE.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y.; Albanie, S.; Nagrani, A.; and Zisserman, A. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9879–9889.

Momeni, L.; Caron, M.; Nagrani, A.; Zisserman, A.; and Schmid, C. 2023. Verbs in Action: Improving verb understanding in video-language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 15533–15545. IEEE.

Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; and Manning, C. D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, 101–108.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Song, P.; Zhang, L.; Lan, L.; Chen, W.; Guo, D.; Yang, X.; and Wang, M. 2025. Towards Efficient Partially Relevant Video Retrieval with Active Moment Discovering. *arXiv preprint arXiv:2504.10920*.

Song, X.; Chen, J.; and Jiang, Y. 2023. Relation Triplet Construction for Cross-modal Text-to-Video Retrieval. In El-Saddik, A.; Mei, T.; Cucchiara, R.; Bertini, M.; Vallejo, D. P. T.; Atrey, P. K.; and Hossain, M. S., eds., *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 4759–4767. ACM.

Song, X.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2021. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia*, 24: 2914–2923.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, A. J.; Ge, Y.; Cai, G.; Yan, R.; Lin, X.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022a. Object-aware Video-language Pre-training for Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 3303–3312. IEEE.

Wang, J.; and Torresani, L. 2022. Deformable Video Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 14033–14042. IEEE.

Wang, Y.; Dong, J.; Liang, T.; Zhang, M.; Cai, R.; and Wang, X. 2022b. Cross-lingual cross-modal retrieval with noise-robust learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 422–433.

Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Chen, G.; Pei, B.; Zheng, R.; Xu, J.; Wang, Z.; et al. 2024a. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*.

Wang, Y.; Wang, J.; Chen, B.; Zeng, Z.; and Xia, S.-T. 2024b. GMMFormer: Gaussian-Mixture-Model Based Transformer for Efficient Partially Relevant Video Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5767–5775.

Wang, Z.; Blume, A.; Li, S.; Liu, G.; Cho, J.; Tang, Z.; Bansal, M.; and Ji, H. 2023. Paxion: Patching Action Knowledge in Video-Language Foundation Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision Transformer with Deformable Attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 4784–4793. IEEE.

Yuan, T.; Zhang, X.; Liu, B.; Liu, K.; Jin, J.; and Jiao, Z. 2024. Surveillance Video-and-Language Understanding: from Small to Large Multimodal Models. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, H.; Sun, A.; Jing, W.; Nan, G.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021a. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 685–695.

Zhang, H.; Sun, A.; Jing, W.; Nan, G.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021b. Video Corpus Moment Retrieval with Contrastive Learning. In Diaz, F.; Shah, C.; Suel, T.; Castells, P.; Jones, R.; and Sakai, T., eds., *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, 685–695. ACM.

Zhang, Q.; Yang, C.; Jiang, B.; and Zhang, B. 2025. Multi-Grained Alignment with Knowledge Distillation for Partially Relevant Video Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable ConvNets V2: More Deformable, Better Results. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 9308–9316. Computer Vision Foundation / IEEE.