Appearance-Motion Decomposed Alignment for Text-Video Retrieval

Meng Meng[1], **Zichang Tan**[1], **Yong Zhang**[1], **Xu Zhou**[1]*

[1]Innovation Research Institute, Sangfor Technologies Inc., Shenzhen, China
meng18@mail.ustc.edu.cn, tanzichang@foxmail.com, yongzhang@link.cuhk.edu.cn, zhouxu@sangfor.com.cn

Abstract

Text-video retrieval aims to bridge vision and language areas, which is a crucial task in multi-modal intelligence. The core idea is to learn video and textual features to quantify their semantic relevance. A common limitation in current approaches is the oversimplification of video content, where complex spatiotemporal structures are compressed into a single global representation. Consequently, these methods struggle to fully capture dynamic visual variations and discriminative appearance inside a video, further complicating cross-modal alignment. To alleviate these issues, we introduce a novel decoupling approach that independently processes appearance and motion cues, capitalizing on their complementary nature for more expressive video modeling. Specifically, we propose an appearance-motion decomposed network (AMD-Net) to decouple spatial-level appearance and temporal-level motion understanding via the discriminative appearance learning and multi-scale motion learning modules. The proposed model enjoys several merits. First, the designed discriminative appearance learning module with a Singular Value Decomposition (SVD) based prototype initialization can effectively reduce redundant appearance information, and a high-order cross-aggregation mechanism enhances prototype resilience and facilitates comprehensive video understanding. Second, the proposed multi-scale motion learning (MML) module can capture motion features at varying temporal scales, which are complementary to appearance features for accurate text-video retrieval. Extensive experiments on five standard benchmarks demonstrate that our method performs favorably against state-of-the-art methods.

Introduction

Understanding multimodal information is crucial for humans to perceive the world effectively. As a fundamental task in multimodal learning and with the rapid growth of short video platforms, vision-language retrieval has gained significant research attention, which aims to identify the most relevant images or videos given a textual query. With the success of image-text pre-trained models, image retrieval methods (Huang et al. 2024) are prompted to utilize CLIP (Radford, Kim et al. 2021) to encode visual and textual features for semantic alignment, and achieve superior performance on various benchmarks. Unlike static images, videos inherently possess a more complex structure due to the introduction of an additional temporal dimension, which poses significant challenges (Deng, Chen et al. 2023).

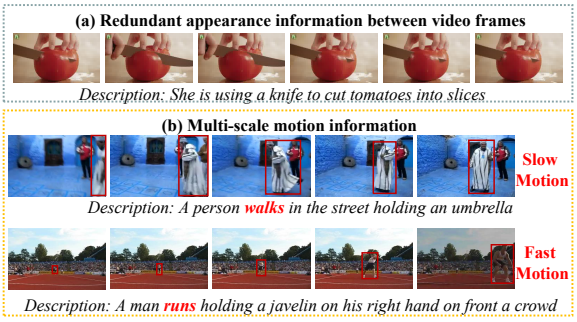

Figure 1: Illustration of our motivation. Video frames often exhibit appearance redundancy, and inherent multi-scale motion patterns, ranging from slow motions (e.g., *walking*) to fast motions (e.g., *running*).

Current approaches (Luo, Ji et al. 2022; Tian, Zhao et al. 2024) in text-video retrieval typically oversimplify the complex nature of video by reducing it to a single holistic embedding. For example, CLIP4Clip (Luo, Ji et al. 2022) conducts a preliminary study by aggregating frame-level features into a video-level representation through mean pooling. However, the mean-pooling of frame representations may lose some essential semantic details of the video and hinder the retrieval performance (Deng, Chen et al. 2023). In pursuit of learning better video representation, several techniques have been proposed, which can be mainly classified into two categories. The first category (Ibrahimi, Sun et al. 2023; Zhuang, Li et al. 2024) seeks to enhance video representation by incorporating textual information through cross-modal interaction. These architectures limit scalability in large-scale retrieval systems (Miech, Alayrac et al. 2021) because their coupled video-text pipelines hinder parallel feature extraction. The other category (Wang, Sung et al. 2023; Li, Xie et al. 2023) enhances video representation through multi-level feature integration (e.g., segment and frame features) to better capture semantics. However, these multi-level features fail to fully capture dynamic visual

*Corresponding author.

variations and discriminative appearance, which exacerbates alignment difficulties (Wang, Sun et al. 2024).

Considering that frame-level features struggle to understand motion cues, and static appearances may hinder temporal perception by overshadowing video motion signals, an ideal solution is to decouple spatial-level appearance and temporal-level motion understanding for better video representation, where appearance and motion cues perform their complementary roles. Here, the appearance focuses on learning potential static visual features, where motion cues are not necessary, and the motion cues are used to match the target sentences by aligning them with textual features. To achieve this goal, two critical aspects require careful consideration: (1) **Concise Appearance Learning.** As shown in Figure 1 (a), video frames frequently contain redundant appearance information, necessitating the development of efficient mechanisms to learn representative yet discriminative appearance features. (2) **Multi-level Motion Learning.** As shown in Figure 1 (b), video data inherently exhibits motion patterns at varying temporal scales. For instance, these patterns range from slow motions, such as *walking*, to fast motions, such as *running*. Consequently, it is imperative for the framework to simultaneously capture both fast and slow motions for comprehensive representation.

Motivated by the above discussion, we propose an end-to-end appearance-motion decomposed network **(AMD-Net)** to decouple spatial-level appearance and temporal-level motion understanding via a discriminative appearance learning **(DAL)** module and a multi-scale motion learning **(MML)** module. In the **DAL** module, it is challenging to represent discriminative object details (*e.g., small children*) and global scene context (*e.g., soccer field*) inherent in videos simultaneously. Thus, we propose object and scene appearance prototypes for capturing different granularity of video appearance. To acquire concise yet discriminative appearance prototypes, we introduce **Singular Value Decomposition (SVD) based prototype initialization** for prototype learning. In this way, redundant appearance information can be reduced by controllable information decay. Then, we propose a **high-order cross-aggregation mechanism** that facilitates dynamic interactions between these prototypes and video representations. This mechanism enhances the robustness and representation capacity of object/scene prototypes, thereby extracting discriminative and robust appearance features. In the **MML** module, we introduce a dual-pathway architecture that captures both short-term and long-term temporal dynamics: (1) a **fast motion encoder** modeling immediate frame-to-frame variations representing fast motion pattern, and (2) a **slow motion encoder** analyzing distant-frame disparities to encode slow motion patterns. Subsequently, a **parameter generator** is proposed to amplify distinctive channels in video representation that are sensitive to the fast/slow motion pattern, and a **motion modulator** is designed to fuse the fast and slow motion strengthened video representation to acquire comprehensive motion feature. Finally, both appearance and motion features are used to find the most relevant text query for accurate matching.

The contributions of our model could be summarized as follows: (1) We propose an end-to-end appearance-motion decomposed network **(AMD-Net)** to decouple spatial-level appearance and temporal-level motion understanding, where appearance and motion cues perform their complementary roles for better video representation. (2) The discriminative appearance learning **(DAL)** module is designed with a **Singular Value Decomposition (SVD) based prototype initialization** to reduce redundant appearance information, and a **high-order cross-aggregation mechanism** to enhance prototype resilience and facilitate comprehensive video understanding. (3) We design a multi-scale motion learning **(MML)** module to capture motion features at varying temporal scales, which are complementary to appearance features for accurate text-video matching. (4) Extensive experimental results with two backbones on five challenging benchmarks demonstrate that our proposed AMD-Net performs favorably against the current state-of-the-art methods.

Related Work

Text-video retrieval, a fundamental task in video-language understanding, has garnered significant attention with the rapid growth of short video platforms. Existing approaches can be categorized into two main methods: (1) cross-modal interaction, and (2) multi-grained representation learning.

Cross-modal Interaction: As a representative method of CLIP variants (Liu, Xiong et al. 2022; Ma, Xu et al. 2022; Luo, Ji et al. 2022), CLIP4Clip aggregates frame-level features given by pretained models into a video-level representation through mean pooling. However, representing an entire video as a single aggregated representation can be over-abstraction and misleading to match the common subcues depicted by multiple corresponding texts (Tian, Cheng et al. 2024). To learn better video representation, recent works (Ibrahimi, Sun et al. 2023; Ma, Xu et al. 2022; Liu, Fan et al. 2021; Zhuang, Li et al. 2024) pay attention to adaptive video features with different text interaction mechanisms. X-CLIP (Ma, Xu et al. 2022), TEFAL (Ibrahimi, Sun et al. 2023), TS2-Net (Liu, Xiong et al. 2022) and KDProR (Zhuang, Li et al. 2024) design a heavy interaction block for learning joint text-conditioned video representation. X-Pool (Gorti, Vouitsis et al. 2022) enhances performance through optimized interaction blocks that efficiently compute similarity measures between frame features and query sentence features with less learnable parameters. However, these methods require tightly coupled video-text feature extraction, limiting parallel processing capabilities (Miech, Alayrac et al. 2021).

Multi-grained Representation Learning: To improve video representation, some recent approaches incorporate multi-grained features through sentence-frame (Gorti, Vouitsis et al. 2022; Lin, Wu et al. 2022; Reddy, Martin et al. 2025), word-frame (Wang, Zhang et al. 2022), and hierarchical-level interactions (Wang, Sung et al. 2023; Wang, Sun et al. 2024; Zhang, Ren et al. 2023). Notable examples include UCOFIA (Wang, Sung et al. 2023), which jointly learns video-sentence, frame-sentence, and patch-word alignments to capture multi-grained text-video similarities, and ProST (Li, Xie et al. 2023), which performs object-phrase and event-sentence prototype matching. However, these methods fail to handle semantic redundancy and

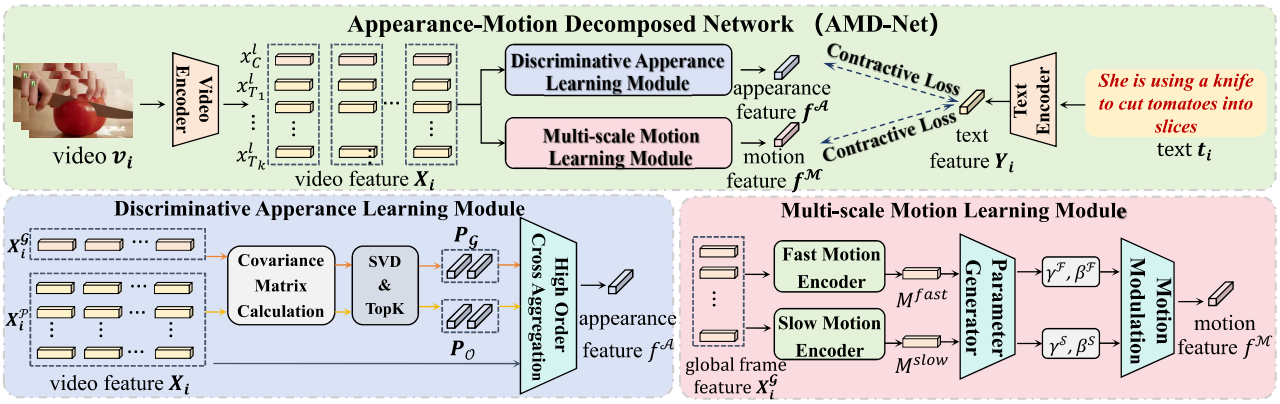

Figure 2: The architecture of our AMD-Net, which decouples spatial-level appearance and temporal-level motion understanding to enhance video representation, and is implemented through two novel modules: a discriminative appearance learning module and a multi-scale motion learning module. For more details, please refer to the paper.

complex motion patterns in high-dimensional video data, with their multi-grained learning modules are inadequate for these challenges (Wang, Sun et al. 2024). Building on these insights, we propose an end-to-end appearance-motion decomposed network **(AMD-Net)** to decouple spatial-level appearance and temporal-level motion understanding for video representation enhancement, and implemented through two key innovations: (1) concise appearance modeling for efficient feature extraction and (2) multi-scale motion modeling for a comprehensive understanding of temporal dynamics.

Given a dataset consisting of n videos $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^{n}$ and their corresponding m texts $\mathcal{T} = \{\mathbf{t}_i\}_{i=1}^{m}$, the text-video retrieval (TVR) aims to learn a function $\mathcal{S}(\mathbf{v}_i, \mathbf{t}_i)$ that effectively quantifies the similarity between these two modalities. Formally, for a given text query $\mathbf{t}_i$, the objective is to rank all videos according to their similarity to the query. This optimization enforces that the similarity score between correctly paired instances exceeds that of mismatched pairs: $\mathcal{S}(\mathbf{v}_i, \mathbf{t}_i) > \mathcal{S}(\mathbf{v}_i, \mathbf{t}_j)$. This requires the model to learn powerful textual and video representations.

Framework

Figure 2 provides an overview of our end-to-end **(AMD-Net)**, which includes a text encoder, a video encoder, a discriminative appearance learning module, and a multi-scale motion learning module.

Text Encoder: Given a text query $\mathbf{t}_i$, we first prepend the identifiers [CLS] and [SEP] to the sentence, and then utilize the text encoder of CLIP $f_t : \mathcal{T} \rightarrow \mathcal{Y}$ to encode the text representation. The output text token features can be defined as $\mathbf{Y}_i = \{\mathbf{y}_S, \mathbf{y}_{T_1}, \cdots, \mathbf{y}_{T_M}, \mathbf{y}_E\} \in \mathbb{R}^{(M+2)\times D}$, where M and D are the number of words and dimensions.

Video Encoder: For each video $\mathbf{v}_i$, we uniformly select L frames as key frames, and employ the transformer-based encoder of CLIP $f_v : \mathcal{V} \rightarrow \mathcal{X}$ to extract features $\mathbf{X}_i = \{\mathbf{x}_C^l, \mathbf{x}_{T_1}^l, \cdots, \mathbf{x}_{T_k}^l\}_{l=1}^{L} \in \mathbb{R}^{L\times(K+1)\times D}$, where $\mathbf{x}_C^l$ is the global frame feature ([CLS]) and K is the patch number.

Discriminative Appearance Learning Module

It is non-trivial to represent diverse semantic appearances (both object details and global scene context inherent in videos. To achieve this goal, we propose object and scene prototypes to capture video appearance at different granularity. In order to enable object and scene prototypes to represent diverse semantic clues and reduce redundant information in the video, we resort to Singular Value Decomposition (SVD) to implement principal component analysis on the basis of global frame features and patch features.

SVD-based Prototype Generation: We first collect all global frame features $\mathbf{X}_i^{\mathcal{G}} = \{\mathbf{x}_C^l\}_{l=1}^{L} \in \mathbb{R}^{L\times D}$ and compute the covariance matrix $\mathbf{W}^{\mathcal{G}} \in \mathbb{R}^{L\times L}$. Meanwhile, we collect all patch features $\mathbf{X}_i^{\mathcal{P}} = \{\mathbf{x}_{T_1}^l, ..., \mathbf{x}_{T_k}^l\}_{l=1}^{L} \in \mathbb{R}^{L\times K\times D}$ and calculate the covariance matrix $\mathbf{W}^{\mathcal{P}} \in \mathbb{R}^{LK\times LK}$. Then, we decompose the original correlation matrix $\mathbf{W}^{\mathcal{G}}$ and $\mathbf{W}^{\mathcal{P}}$ via SVD and only keep the largest I singular values:

$$\mathbf{W}^{\mathcal{G}} \xrightarrow[\text{Top - K}]{\text{SVD}} \mathbf{V}_{\mathcal{G}}\mathbf{\Sigma}\mathbf{V}_{\mathcal{G}}^{\top}, \quad \mathbf{W}^{\mathcal{P}} \xrightarrow[\text{Top - K}]{\text{SVD}} \mathbf{V}_{\mathcal{P}}\mathbf{\Sigma}\mathbf{V}_{\mathcal{P}}^{\top}, \quad (1)$$

where $\mathbf{V}_{\mathcal{G}} \in \mathbb{R}^{L\times I}$, $\mathbf{V}_{\mathcal{P}} \in \mathbb{R}^{LK\times I}$, $\mathbf{\Sigma} \in \mathbb{R}^{I\times I}$, and $\top$ denotes the transpose operation. Since the singular value decreases rapidly, preserving the largest I singular values is sufficient to preserve discriminative appearance information and reduce redundant information. For the singular matrix $\mathbf{V}_{\mathcal{G}} \in \mathbb{R}^{L\times I}$ and $\mathbf{V}_{\mathcal{P}} \in \mathbb{R}^{L\times I}$ obtained from SVD decomposition, we explicitly map global frame features and patch features through a linear transformation (multiplied by the orthogonal bases) to attain object prototypes $\mathbf{P}_{\mathcal{O}} \in \mathbb{R}^{I\times D}$ and scene prototypes $\mathbf{P}_{\mathcal{S}} \in \mathbb{R}^{I\times D}$ in the orthotropic space, which can retain informative appearance information:

$$\mathbf{P}_{\mathcal{G}}^{\top} = (\mathbf{X}_i^{\mathcal{G}})^{\top}\mathbf{V}_{\mathcal{G}}, \quad \mathbf{P}_{\mathcal{O}}^{\top} = (\mathbf{X}_i^{\mathcal{O}})^{\top}\mathbf{V}_{\mathcal{P}}. \quad (2)$$

High-order Cross-aggregation Mechanism: To capture discriminative appearance information robustly, we introduce a high-order interaction mechanism that models dynamic interactions between video features and prototypes to enhance the robustness and representation capacity of object

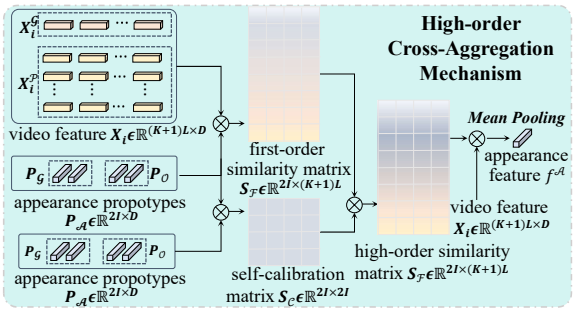

Figure 3: Overview of high-order cross-aggregation mechanism, which enables dynamic video feature-prototype interactions, enhancing robustness and representation capacity.

and scene prototypes. Specifically, we concatenate object and scene prototypes $\mathbf{P}_{\mathcal{A}} = [\mathbf{P}_{\mathcal{G}}; \mathbf{P}_{\mathcal{O}}]$ to acquire the overall appearance prototypes $\mathbf{P}_{\mathcal{A}} = \{\mathbf{p}_1; \mathbf{p}_2; \ldots; \mathbf{p}_{2I}\}, \mathbf{p}_i \in \mathbb{R}^D$. Inspired by the transformer architecture (Vaswani et al. 2017), we utilize queries, keys, and values to implement a high-order cross-aggregation mechanism. Formally, queries $\mathbf{Q} = \{\mathbf{q}_1; \mathbf{q}_2; \ldots; \mathbf{q}_{2I}\}$ arise from the appearance prototypes, keys $\mathbf{K} = \{\mathbf{k}_1; \mathbf{k}_2; \ldots; \mathbf{k}_{L(K+1)}\}$ and values $\mathbf{V} = \{\mathbf{v}_1; \mathbf{v}_2; \ldots; \mathbf{v}_{L(K+1)}\}$ arise from the reshaped video features $\mathbf{X}_{\mathcal{R}} = \{\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_{L(K+1)}\}, \mathbf{x}_i \in \mathbb{R}^D$ as

$$\mathbf{q}_i = \mathbf{p}_i \mathbf{W}^{\mathcal{Q}},\ \mathbf{k}_j = \mathbf{x}_j \mathbf{W}^{\mathcal{K}},\ \mathbf{v}_j = \mathbf{x}_j \mathbf{W}^{\mathcal{V}}, \tag{3}$$

where $i = 1, 2, \ldots, 2I; j = 1, 2, \ldots, L(K+1)$ and $\mathbf{W}^{\mathcal{Q}} \in \mathbb{R}^{D \times D}$, $\mathbf{W}^{\mathcal{K}} \in \mathbb{R}^{D \times D}$, $\mathbf{W}^{\mathcal{V}} \in \mathbb{R}^{D \times D}$ are linear projections. Then, we calculate the attention weight between the i-th query $\mathbf{q}_i$ and the j-th key $\mathbf{k}_j$ as

$$s_{i,j}^{\text{first}} = \frac{exp(\beta_{i,j})}{\sum_{i=1}^{l} exp(\beta_{i,j})},\ \beta_{i,j} = \frac{\mathbf{q}_i\, \mathbf{k}_j^{\mathsf{T}}}{\sqrt{D}}. \tag{4}$$

In this way, we obtain the first-order similarity matrix $S_{\mathcal{F}} \in \mathbb{R}^{2I \times (K+1)L}$ by ensembling all attention weights $s_{i,j}^{\text{first}}$. The attention weights between queries are computed as:

$$s_{i,j}^{\text{self}} = \frac{exp(\beta_{i,j})}{\sum_{i=1}^{l} exp(\beta_{i,j})},\ \beta_{i,j} = \frac{\mathbf{q}_i\, \mathbf{q}_j^{\mathsf{T}}}{\sqrt{D}}. \tag{5}$$

Similarly, we acquire a self-calibration matrix $S_{\mathcal{C}} \in \mathbb{R}^{2I \times 2I}$ with a concatenation of all attention weights $s_{i,j}^{\text{self}}$. The high-order similarity matrix $S_{\mathcal{H}} \in \mathbb{R}^{2I \times (K+1)L}$ is obtained by multiplying $S_{\mathcal{F}} \in \mathbb{R}^{2I \times (K+1)L}$ and $S_{\mathcal{C}} \in \mathbb{R}^{2I \times 2I}$ to rectify the first-order similarity matrix.

$$S_{\mathcal{H}} = S_{\mathcal{F}} \cdot S_{\mathcal{C}} \tag{6}$$

Here, we use $s_{i,j}^{\text{high}}$ to denote the i-th row and j-th column of $S_{\mathcal{H}}$. Finally, the discriminative yet robust appearance feature $\mathbf{f}^{\mathcal{A}} \in \mathbb{R}^D$ is obtained via the weighted sum over all values:

$$\mathbf{f}^{\mathcal{A}} = \sum_{i=1}^{2l} \sum_{j=1}^{(K+1)L} s_{i,j}^{\text{high}} \mathbf{v}_j. \tag{7}$$

Multi-scale Motion Learning Module

To capture motion features at varying temporal scales, we propose a multi-scale motion learning module, which is complementary to appearance features. The computation of motion features is fundamentally grounded in the temporal differences between consecutive frames, which have been theoretically established to correlate with optical flow patterns and serve as effective approximations for motion representation (Wang, Tong et al. 2021). Building upon this theoretical foundation, we formulate the motion feature extraction process through a dual-pathway architecture that captures both short-term and long-term temporal dynamics.

Fast Motion Encoder and Slow Motion Encoder: Specifically, given a sequence of global frame features $\mathbf{X}_i^{\mathcal{G}} = \{\mathbf{x}_C^l\}_{l=1}^{L} \in \mathbb{R}^{L \times D}$, where L denotes the total number of frames and D represents the feature dimension, we compute the short-term and long-term temporal differences as:

$$M^{fast} = \frac{1}{L-1} \sum x_C^{(l+1)} - x_C^l, l = 1, 2, ..., L-1, \tag{8}$$

$$M^{slow} = \frac{1}{L-h} \sum x_C^{(l+h)} - x_C^l, l = 1, 2, ..., L-h, \tag{9}$$

where M^{fast} captures immediate frame-to-frame variations representing fast motion patterns, while M^{slow} encodes slow motion patterns using differences across five frame intervals. h is a hyper-parameter and we set $h = 5$ though experiments. Then, we input M^{fast} and M^{slow} into the fast motion encoder ϕ^{fast} and the slow motion encoder ϕ^{slow}, respectively. Both encoders contain two fully-connected layers (output dimension D) that transform the inputs. The encoded slow/fast motion features are computed as:

$$\hat{M}^{fast} = \phi^{fast}(M^{fast}),\ \hat{M}^{slow} = \phi^{slow}(M^{slow}). \tag{10}$$

Parameter Generator and Motion Modulator: To effectively integrate fast and slow motion features for comprehensive motion representation, we propose the parameter generators g_γ and g_β conditioned on $\hat{M}^{fast}$ and $\hat{M}^{slow}$:

$$\gamma_{\mathcal{F}} = g_\gamma(\hat{M}^{fast}),\ \beta_{\mathcal{F}} = g_\beta(\hat{M}^{fast}) \tag{11}$$

$$\gamma_{\mathcal{S}} = g_\gamma(\hat{M}^{slow}),\ \beta_{\mathcal{S}} = g_\beta(\hat{M}^{slow}) \tag{12}$$

where each parameter generator consists of one linear layer (output dimension D), followed by a ReLU activation function. Under the guidance of fast/slow motion features, $\gamma_{\mathcal{F}}/\beta_{\mathcal{F}}$ and $\gamma_{\mathcal{S}}/\beta_{\mathcal{S}}$ can strengthen the distinctive channels that are sensitive to the fast/slow motion patterns. Then, we employ a motion modulator $\Pi^{\mathcal{M}}$ to fuse the fast and slow motion strengthened video representation. The modulator consists of some modulation layers, each applying an affine transformation to adapt the video features in the corresponding encoder layer. Specifically, given the global frame features $\mathbf{X}_i^{\mathcal{G}} = \{\mathbf{x}_C^l\}_{l=1}^{L}$ for video i , we obtain the comprehensive motion feature $\mathbf{f}_{\mathcal{M}}$ through the motion modulator $\Pi^{\mathcal{M}}$ as:

$$\mathbf{f}_{\mathcal{M}} = \Pi^{\mathcal{M}}(\gamma_{\mathcal{F}} \mathbf{X}_i^{\mathcal{G}} + \beta_{\mathcal{F}} + \gamma_{\mathcal{S}} \mathbf{X}_i^{\mathcal{G}} + \beta_{\mathcal{S}}) \tag{13}$$

Methods	MSRVTT Retrieval R@1	R@5	R@10	MnR (↓)	DiDeMo Retrieval R@1	R@5	R@10	MnR (↓)	ActivityNet Retrieval R@1	R@5	R@10	MnR (↓)
CLIP-ViT-B/32												
EMCL (Jin, Huang et al. 2022)	46.8	73.1	83.1	-	-	-	-	-	-	-	-	-
X-Pool (Gorti, Vouitsis et al. 2022)	46.9	72.8	82.2	14.3	44.6	73.2	82.0	15.4	-	-	-	-
TS2-Net (Liu, Xiong et al. 2022)	47.0	74.2	83.3	13.6	41.8	71.6	82.0	14.8	-	-	-	-
CLIP-ViP* (Xue, Sun et al. 2022)	50.1	74.8	84.6	-	48.6	77.1	84.4	-	51.1	78.4	88.3	-
UATVR (Fang, Wu et al. 2023)	47.5	73.9	83.5	12.3	43.1	71.8	82.3	15.1	-	-	-	-
ProST (Li, Xie et al. 2023)	48.2	74.6	83.4	12.4	44.9	72.7	82.7	13.7	-	-	-	-
HBI (Jin, Huang et al. 2023)	48.6	74.6	83.4	12.0	46.9	74.9	82.7	12.1	43.9	73.0	84.6	6.6
Cap4Video* (Wu, Luo et al. 2023)	49.3	74.3	83.8	12.0	52.0	**79.4**	**87.5**	10.5	-	-	-	-
UCOFIA* (Wang, Sung et al. 2023)	49.4	72.1	-	12.9	46.5	74.8	-	13.4	45.7	76.0	-	6.6
T-MASS (Wang, Sun et al. 2024)	50.2	75.3	85.1	11.9	50.9	77.2	85.3	12.1	-	-	-	-
ProxyNet (Xiao, Hu et al. 2025)	52.3	77.8	85.8	11.1	50.6	76.9	86.0	11.5	53.0	80.9	89.6	-
VIDEO-COLBERT (Reddy, Martin et al. 2025)	48.1	74.9	83.9	-	48.2	75.1	83.7	-	45.5	74.6	85.5	-
AMD-Net (ours)	**55.2**	**78.7**	**85.3**	**9.5**	**53.5**	78.9	86.9	**9.1**	**57.8**	**82.0**	**90.1**	**5.6**
CLIP-ViT-B/16												
X-Pool (Gorti, Vouitsis et al. 2022)	48.2	73.7	82.6	12.7	47.3	74.8	82.8	14.2	-	-	-	-
ProST (Li, Xie et al. 2023)	49.5	75.0	84.0	11.7	47.3	74.8	82.8	14.2	-	-	-	-
UATVR (Fang, Wu et al. 2023)	50.8	76.3	85.5	12.4	45.8	73.7	83.3	13.5	-	-	-	-
CLIP-ViP* (Xue, Sun et al. 2022)	54.2	77.2	84.8	-	50.5	78.4	87.1	-	53.4	81.4	90.0	-
Cap4Video* (Wu, Luo et al. 2023)	51.4	75.7	83.9	12.4	-	-	-	-	-	-	-	-
T-MASS (Wang, Sun et al. 2024)	52.7	77.1	85.6	10.5	53.3	80.1	87.7	9.8	-	-	-	-
ProxyNet (Xiao, Hu et al. 2025)	55.2	80.4	86.8	9.3	52.1	80.4	**88.7**	8.7	56.7	83.1	91.1	-
VIDEO-COLBERT (Reddy, Martin et al. 2025)	51.5	76.3	85.5	-	51.7	76.1	84.8	-	45.8	76.3	86.7	-
AMD-Net (ours)	**59.0**	**80.8**	**87.4**	**7.9**	**55.1**	**80.9**	86.7	**8.3**	**58.3**	**82.5**	**91.5**	**5.4**

Table 1: Comparison results of the cross-modal retrieval on the MSRVTT, DiDeMo and ActivityNet test set in terms of Recall@K(R@K). **CLIP-ViT-B/32** and **CLIP-ViT-B/16** represent using ViT-B/32 and ViT-B/16 of CLIP (Luo, Ji et al. 2022) to extract video features, respectively. Bold denotes the best performance. '$*$' denotes the method using extra data.

Training and Inference

The global textual feature is represented as $\mathbf{y}_S \in \mathbb{R}^D$. The similarity of the text-video is calculated as the inner production of the appearance-text and motion-text as:

$$s(t_i, v_i) =< \mathbf{y}_S, \mathbf{f}_{\mathcal{A}} > + < \mathbf{y}_S, \mathbf{f}_{\mathcal{M}} > . \tag{14}$$

In training, a common optimizing method is to use a symmetric cross-entropy loss in both text-to-video and video-to-text directions. Given a batch of B text-video pairs, the model updates its parameters by maximizing the sum of the main diagonal of a $B \times B$ similarity matrix:

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i}^{B} log\frac{exp(s(t_i, v_i))}{\sum_{j=1}^{B} exp(s(t_i, v_j))}, \tag{15}$$

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i}^{B} log\frac{exp(s(v_i, t_i))}{\sum_{j=1}^{B} exp(s(v_i, t_j))}, \tag{16}$$

$$\mathcal{L} = \mathcal{L}_{t2v} + \mathcal{L}_{v2t}. \tag{17}$$

During the inference stage, we directly weight the appearance-text and motion-text matching scores for the final similarity matching: $s(t_i, v_i) =< \mathbf{y}_S, \mathbf{f}_{\mathcal{A}} > +\lambda < \mathbf{y}_S, \mathbf{f}_{\mathcal{M}} >$, where λ is the motion matching factor.

Experiments

Experimental Settings

Datasets. MSRVTT (Xu, Mei et al. 2016) contains 10,000 YouTube videos, each with 20 text descriptions. **DiDeMo Dataset** (Anne Hendricks, Wang et al. 2017) is a comprehensive benchmark for text-video retrieval, consisting of 10,464 videos with 40,543 temporally-aligned textual descriptions. **ActivityNet Dataset** (Krishna, Hata et al. 2017) is a large-scale benchmark for video understanding, comprising 20,000 YouTube videos with dense temporal annotations. **MSVD Dataset** (Chen and Dolan 2011) is a widely-adopted benchmark for video captioning and retrieval tasks, consisting of 1,970 YouTube video clips with multilingual annotations. **Charades Dataset** (Sigurdsson, Varol et al. 2016) is a comprehensive benchmark for video understanding, comprising 9,848 videos with single-sentence textual descriptions. Following established evaluation protocols (Liu, Xiong et al. 2022), we adopt their standard training and testing split, and employ **Recall@K** (**R@K**, where K = 1, 5, 10) and **Mean Rank** (**MnR**) as primary metrics for quantitative assessment.

Implementation Details

Following previous works (Gorti, Vouitsis et al. 2022), we initialize both text and video encoders using pre-trained CLIP models (ViT-B/32 and ViT-B/16). For video preprocessing, we first resize all frames to a spatial resolution of 224×224 pixels. To ensure fair comparison across datasets, we uniformly sample 12 frames per video, except for ActivityNet where we extract 32 frames to accommodate its longer video durations. For model optimization, we employ the Adam optimizer with a cosine warm-up learning rate scheduling strategy (Kingma and Ba 2014). To preserve the pre-trained representations, the visual and text encoders are

Methods	MSVD R@1	R@5	R@10	MnR (↓)
CLIP-ViT-B/32				
CenterCLIP (Zhao, Zhu et al. 2022)	47.3	76.9	86.0	9.7
CLIP4Clip (Luo, Ji et al. 2022)	46.2	76.1	84.6	10.0
X-CLIP (Ma, Xu et al. 2022)	47.1	77.8	-	9.5
UCOFIA* (Wang, Sung et al. 2023)	46.5	74.8	-	13.4
TeachCLIP (Tian, Zhao et al. 2023)	46.8	74.3	-	-
AMD-Net (ours)	**56.8**	**84.0**	**90.2**	**6.4**
CLIP-ViT-B/16				
CenterCLIP (Zhao, Zhu et al. 2022)	50.6	80.3	88.4	8.4
X-CLIP (Ma, Xu et al. 2022)	50.4	80.6	-	8.4
AMD-Net (ours)	**61.6**	**85.9**	**90.8**	**5.4**

Table 2: Comparison results of the cross-modal retrieval on the MSVD test set in terms of Recall@K(R@K).

Methods	Charades R@1	R@5	R@10	MnR (↓)
CLIP-ViT-B/32				
CLIP4Clip (Luo, Ji et al. 2022)	13.9	-	-	-
ECLIPSE (Lin, Lei et al. 2022)	15.7	-	-	-
X-CLIP (Ma, Xu et al. 2022)	16.1	35.2	44.9	67.2
TEFAL (Ibrahimi, Sun et al. 2023)	18.5	37.3	48.6	60.6
AMD-Net (ours)	**19.7**	**41.3**	**51.1**	**58.6**

Table 3: Results on the test split of Charades.

optimized with a learning rate of $1e{-}7$, while other trainable modules utilize a higher learning rate of $1e-4$ for efficient adaptation. The motion modulator consists of three modulator layers, and each layer contains two fully-connected layers and a ReLU activation.

Comparison with State-of-the-art Methods

In this section, swe compare our method with current text-video retrieval methods on the MSRVTT, DiDeMo, ActivityNet, MSVD, and Charades datasets, including (1) the cross-modal interaction methods, $i.e.$, X-Pool, TS2-Net, X-CLIP, and (2) the multi-grained representation learning methods, ProST, UCOFIA, and the latest state-of-the-art methods, $i.e.$, Cap4Video, ProxyNet, etc. Unlike these methods, our model improves video representation by decoupling spatial appearance and temporal motion learning. This involves two innovations: (1) concise appearance modeling and (2) multi-scale motion modeling. For fair comparison, we group existing methods by feature extractor: those using CLIP-ViT-B/32 and CLIP-ViT-B/16.

Table 1 shows detailed comparisons on the **MSRVTT** test set, where our proposed method outperforms recent approaches in terms of most evaluation metrics. Moreover, our method consistently outperforms others across different feature extractors and achieves the best performance on MnR. When using CLIP-ViT-B/32 to extract features, we achieve 55.2%, 78.7%, 85.3%, and 9.5% for R@1, R@5, R@10 and MnR, Benefiting from the stronger feature extractor CLIP-ViT-B/16, our model can further achieve better performance. Specifically, when CLIP-ViT-B/16 is used as the feature extractor, the performance is 59.0%, 80.8%, 87.4%, and 7.9% for R@1, R@5, R@10 and MnR. X-Pool uses the text to attend to its most semantically similar frames and generates an aggregated video representation conditioned on those attended frames. Compared with this cross-modal interaction method, we achieve **10.8**$\%$ improvement in terms of R@1, demonstrating the advantages of decoupling spatial appearance and temporal motion for better video understanding. Besides, our model significantly surpasses the multi-grained representation learning method ProST by absolute **9.5%** R@1, reaching **59.0**$\%$ for R@1 metric on MSRVTT, which sets a new state-of-the-art performance.

Table 1 also lists the results on larger and complicated DiDeMo and ActivityNet datasets. For DiDeMo, our model also demonstrates competitive performance with top-performing CLIP based model Cap4Video* (Wu, Luo et al. 2023). It is worth noting that this approach employs extra training data. For **ActivityNet**, our method achieves significant improvements of **4.8%** and **1.6%** in R@1 when using CLIP-ViT-B/32 and CLIP-ViT-B/16 as backbones. Table 2 and Table 3 show results for other benchmarks. For **MSVD**, our approach with CLIP-ViT-B/16 achieves **11.2%** R@1 and **3.0%** MnR improvements compared with the current state-of-the-art method X-CLIP (Ma, Xu et al. 2022). For **Charades**, our approach achieves **2.2%** R@1 improvement for text-video retrieval. Extensive experimental results validate the superiority of our model, which stems from its ability to effectively model both discriminative appearance learning and multi-scale motion representations.

No.	DAL: w/o cross	DAL: first order	DAL: high order	MML: fast motion	MML: slow motion	R@1	R@5	R@10	MnR
0	✓					46.2	70.4	79.6	13.8
1		✓				47.1	72.5	81.3	11.2
2			✓			49.0	76.1	85.1	11.7
3			✓	✓		51.2	75.5	83.4	10.9
4			✓		✓	52.9	77.6	85.3	9.5
5			✓	✓	✓	**55.2**	**78.7**	**85.3**	**9.5**

Table 4: **Ablation studies** about the proposed appearance learning (DAL) and multi-scale motion learning (MML) .

AMD-Net	R@1	R@5	R@10	MnR
Rand Initialization	49.1	73.8	83.5	9.9
Learnable Initialization	52.8	74.9	83.7	11.6
SVD based Initialization	**55.2**	**78.7**	**85.3**	**9.5**

Table 5: **Ablation studies** about the SVD-based prototype initialization on the MSRVTT test set.

Ablation Studies and Analysis

Effectiveness of the DAL Module: In Table 4, we report the results of different dynamic interactions between video representations and prototypes: without dynamic interactions, dynamic interactions with first-order and high-order similarity matrixs.

Ablation study on High-order Cross-aggregation Mechanism. Comparing #0, #2 in Table 4, based on R@1, the

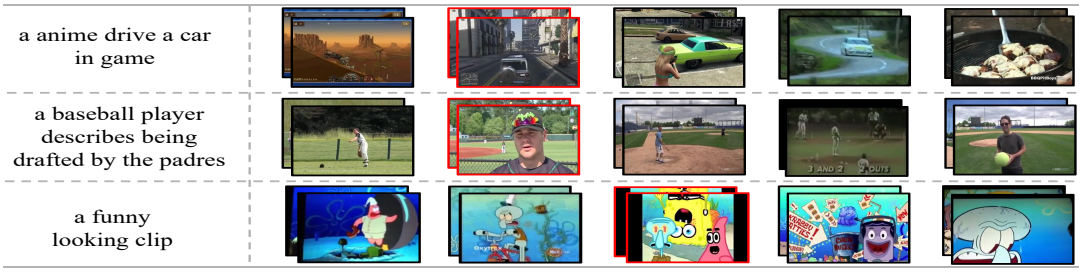

Figure 4: The failure cases on MSRVTT experiments. We report the Top-5 retrieved videos with our AMD-Net. Although our model significantly outperforms previous methods, it still confused by results that closely align with the given queries.

results show that dynamic interactions with first-order similarity matrix yields higher performance compared to the case without dynamic interactions. This observation demonstrates the critical importance of dynamic interactions between prototypes and video representations, which enables prototypes to extract informative video appearance. Comparing #1, #2, the results show that compared to dynamic interactions the first-order similarity matrix, dynamic interactions with high-order similarity matrix achieves a huge performance improvement. This stems from our high-order dynamic interactions that model prototype relationships, allowing them to become more resilient by leveraging information from other prototypes.

Ablation study on SVD-based prototype initialization. To explore the effectiveness of different ways to initialize the prototypes, we compare the performance of several intuitive initialization methods as shown in Table 5. Among these methods, *Rand Initialization* is conducted by randomly selecting I token features as prototypes from patch features and global frame features, respectively. *Learnable Initialization* means that all $2I$ prototypes are randomly initialized but trainable and shared across videos. Our SVD-based prototype initialization outperforms these schemes by a considerable margin credited to the characteristics of SVD. Compared with other methods, SVD-based prototype initialization can better reduce redundant information as well as preserve discriminative appearance.

Effectiveness of the MML Module: To explore the effectiveness of the proposed MML Module, we conduct the following ablation studies: (1) removing fast motion learning, (2) removing slow motion learning. Comparing #3 and #5 based on R@1, the performance of the model with both fast motion and slow motion learning process achieves **4.0 %** improvements over the model with only fast motion learning, indicating the effectiveness of slow motion learning. This is because the slow motion learning enables the acquisition of fine-grained motion variations, thereby enhancing feature learning. Comparing #4 and #5 based on R@1, the model with both fast motion and slow motion learning also achieves **2.3%** improvement compared to the model with only fast motion learning. The results demonstrate that fast motion learning effectively captures object dynamics during rapid changes and enhances performance by implicitly modeling relationships between adjacent frames.

Prototype Number I	R@1	R@5	R@10	MnR
1	54.5	77.2	83.9	11.1
5	54.8	77.2	84.1	10.4
10	**55.2**	**78.7**	**85.3**	**9.5**
15	55.1	78.4	85.2	9.6
20	55.0	77.4	84.5	10.2

Table 6: Performance comparison with prototype number.

Hyperparameter Evaluations

Quantitative experiments are conducted to clearly find a suitable number of prototypes I. In Table 6, we report the results of different number I on the MSRVTT test set. We can find that the performance continues to grow until I=20 and then begins to decline if I keeps increasing. We deem the main reason is too few prototypes cannot represent diverse semantic information, while too many prototypes will produce undesirable redundancy.

Limitations

Figure 4 gives some failure examples. Although our model performs much better than previous state-of-the-art methods, it still falls short in retrieving results that closely align with the given queries. For example, as seen from the first sample in Figure 4, for the given text *"a anime drive a car in game"*, the proposed model is confused by similar videos, resulting in wrong retrieval results. To alleviate this problem, future work will incorporate contrastive learning for better distinction and improve retrieval accuracy.

Conclusion

In this paper, we propose a novel appearance-motion decomposed network (AMD-Net) to decouple spatial-level appearance and temporal-level motion understanding via DAL and MML modules, where appearance and motion cues perform their complementary roles. The DAL with a Singular Value Decomposition (SVD) based prototype initialization can effectively reduce redundant appearance information, and a high-order cross-aggregation mechanism enhances prototype resilience and facilitates comprehensive video understanding. The MML can capture motion features at varying temporal scales, which are complementary to appearance features. The experiments verify its effectiveness.

Acknowledgments

This work is supported by Guangdong Provincial Key Laboratory of Cloud Security Key Technology (2022B1212020006Shenzhen Key Laboratory of Key Cloud Security Technology(No.ZDSY20200811143600002), and the China Postdoctoral Science Foundation 2025M771717.

References

Anne Hendricks, L.; Wang, O.; et al. 2017. Localizing moments in video with natural language. In *ICCV*, 5803–5812.

Chen, D.; and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 190–200.

Deng, C.; Chen, Q.; et al. 2023. Prompt switch: Efficient clip adaptation for text-video retrieval. In *ICCV*, 15648–15658.

Fang, B.; Wu, W.; et al. 2023. Uatvr: Uncertainty-adaptive text-video retrieval. In *ICCV*, 13723–13733.

Gorti, S. K.; Vouitsis, N.; et al. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 5006–5015.

Huang, H.; Nie, Z.; Wang, Z.; and Shang, Z. 2024. Cross-Modal and Uni-Modal Soft-Label Alignment for Image-Text Retrieval. In *AAAI*, volume 38, 18298–18306.

Ibrahimi, S.; Sun, X.; et al. 2023. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *ICCV*, 12054–12064.

Jin, P.; Huang, J.; et al. 2022. Expectation-maximization contrastive learning for compact video-and-language representations. *NIPS*, 35: 30291–30306.

Jin, P.; Huang, J.; et al. 2023. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *CVPR*, 2472–2482.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krishna, R.; Hata, K.; et al. 2017. Dense-captioning events in videos. In *ICCV*, 706–715.

Li, P.; Xie, C.-W.; et al. 2023. Progressive spatio-temporal prototype matching for text-video retrieval. In *ICCV*, 4100–4110.

Lin, C.; Wu, A.; et al. 2022. Text-adaptive multiple visual prototype matching for video-text retrieval. *NIPS*, 35: 38655–38666.

Lin, Y.-B.; Lei, J.; et al. 2022. Eclipse: Efficient long-range video retrieval using sight and sound. In *ECCV*, 413–430. Springer.

Liu, S.; Fan, H.; et al. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *ICCV*, 11915–11925.

Liu, Y.; Xiong, P.; et al. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, 319–335. Springer.

Luo, H.; Ji, L.; et al. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Ma, Y.; Xu, G.; et al. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, 638–647.

Miech, A.; Alayrac, J.-B.; et al. 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, 9826–9836.

Radford, A.; Kim, J. W.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Reddy, A.; Martin, A.; et al. 2025. Video-ColBERT: Contextualized Late Interaction for Text-to-Video Retrieval. *arXiv preprint arXiv:2503.19009*.

Sigurdsson, G. A.; Varol, G.; et al. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 510–526. Springer.

Tian, K.; Cheng, Y.; et al. 2024. Towards Efficient and Effective Text-to-Video Retrieval with Coarse-to-Fine Visual Representation Learning. In *AAAI*, volume 38, 5207–5214.

Tian, K.; Zhao, R.; et al. 2023. TeachCLIP: Multi-Grained Teaching for Efficient Text-to-Video Retrieval. *arXiv preprint arXiv:2308.01217*.

Tian, K.; Zhao, R.; et al. 2024. Holistic features are almost sufficient for text-to-video retrieval. In *CVPR*, 17138–17147.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NIPS*, 30.

Wang, J.; Sun, G.; et al. 2024. Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval. In *CVPR*, 16551–16560.

Wang, L.; Tong, Z.; et al. 2021. Tdn: Temporal difference networks for efficient action recognition. In *CVPR*, 1895–1904.

Wang, Q.; Zhang, Y.; et al. 2022. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*.

Wang, Z.; Sung, Y.-L.; et al. 2023. Unified coarse-to-fine alignment for video-text retrieval. In *ICCV*, 2816–2827.

Wu, W.; Luo, H.; et al. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *CVPR*, 10704–10713.

Xiao, J.; Hu, Z.; et al. 2025. Text Proxy: Decomposing Retrieval from a 1-to-N Relationship into N 1-to-1 Relationships for Text-Video Retrieval. In *AAAI*, volume 39, 8655–8663.

Xu, J.; Mei, T.; et al. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 5288–5296.

Xue, H.; Sun, Y.; et al. 2022. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *ICLR*.

Zhang, G.; Ren, J.; et al. 2023. Multi-event video-text retrieval. In *ICCV*, 22113–22123.

Zhao, S.; Zhu, L.; et al. 2022. Centerclip: Token clustering for efficient text-video retrieval. In *ACM SIGIR*, 970–981.

Zhuang, X.; Li, H.; et al. 2024. Kdpror: A knowledge-decoupling probabilistic framework for video-text retrieval. In *ECCV*, 313–331. Springer.