Learning Dynamic Similarity by Bidirectional Hierarchical Sliding Semantic Probe for Efficient Text Video Retrieval

Yang Liu[1,2], **Shudong Huang**[1,2*], **Deng Xiong**[3], **Jiancheng Lv**[1,2]

[1] College of Computer Science, Sichuan University, Chengdu, 610065, China
[2] Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, Chengdu, China
[3]Department of Mechanical Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA
liuyyy111@gmail.com, huangsd@scu.edu.cn, dxiong@stevens.edu, lvjiancheng@scu.edu.cn

Abstract

Text-video retrieval is a foundation task in multi-modal research which aims to align texts and videos in the embedding space. The key challenge is to learn the similarity between videos and texts. A conventional approach involves directly aligning video-text pairs using cosine similarity. However, due to the disparity in the information conveyed by videos and texts, i.e., a single video can be described from multiple perspectives, the retrieval accuracy is suboptimal. An alternative approach employs cross-modal interaction to enable videos to dynamically acquire distinct features from various texts, thus facilitating similarity calculations. Nevertheless, this solution incurs a computational complexity of $O(n^2)$ during retrieval. To this end, this paper proposes a novel method called Bidirectional Hierarchical Sliding Semantic Probe (BiHSSP), which calculates dynamic similarity between videos and texts with $O(n)$ complexity during retrieval. We introduce a hierarchical semantic probe module that learns semantic probes at different scales for both video and text features. Semantic probe involves a sliding calculation of the cross-correlation between semantic probes at different scales and embeddings from another modality, allowing for dynamic similarity computation between video and text descriptions from various perspectives. Specifically, for text descriptions from different angles, we calculate the similarity at different locations within the video features and vice versa. This approach preserves the complete information of the video while addressing the issue of unequal information between video and text without requiring cross-modal interaction. Additionally, our method can function as a plug-and-play module across various methods, thereby enhancing the corresponding performance. Experimental results demonstrate that our BiHSSP significantly outperforms the baseline.

Introduction

Representation learning that integrates both visual and linguistic modalities offers significant potential for a wide range of cross-modal tasks (Liu et al. 2022b), including image-text matching (Qin et al. 2022; Lee et al. 2018; Liu et al. 2023b; Qin et al. 2024), text-video retrieval (Liu et al. 2019; Gabeur et al. 2020; Bain et al. 2021), and video-question answering(Lei et al. 2018). Among these tasks, text-video retrieval (TVR), a fundamental aspect of multi-modal learning, has garnered considerable attention due to the rapid expansion of short video platforms. TVR focuses on retrieving semantically relevant videos based on user-provided textual queries. The primary challenge in TVR arises from the disparate initial distributions of multiple uni-modal features, necessitating an effective method for calculating the similarity between video and text in order to accurately identify semantically related video-text pairs.

The field of visual-language learning has recently witnessed significant advancements, largely driven by the success of contrastive learning techniques (He et al. 2020), such as CLIP (Radford et al. 2021). These techniques project image and text features into a shared latent space by leveraging the cosine similarity (dot product) of image-text pairs. This cross-modal contrastive approach enables networks to learn discriminative video-language representations, prompting many researchers to incorporate CLIP into the TVR task.

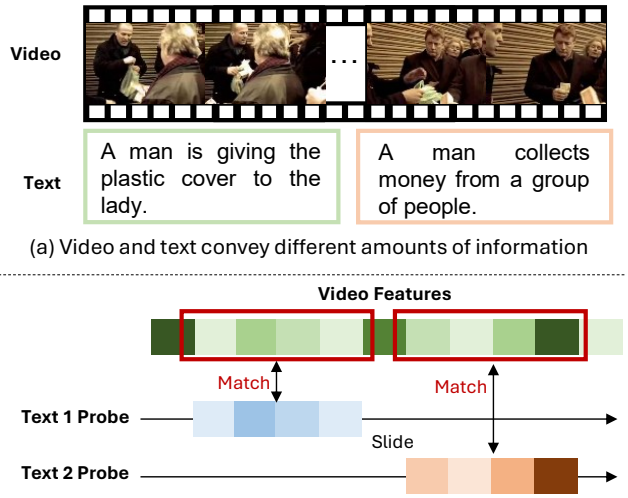

Figure 1: (a) Illustration of the information unequal problem between videos and texts, in this case, video contain more information than texts. (b) Taking text to video as an example, we illustrate how our method dynamically calculates similarity.

*Corresponding Author.

Such methods either directly align the complete features of video and text or fine-grainedly align video and text features across multiple scales. Compared to retrieval methods (Ji et al. 2022) that do not utilize CLIP, these approaches have shown substantial improvements.

Nevertheless, the amount of information conveyed by videos and texts is inherently unequal. As illustrated in Figure 1(a), a video often comprises multiple clips, each corresponding to a different description, thereby potentially offering more detailed information than a single text. Conversely, texts can also contain information not present in videos, such as references to specific landmarks, cultural relics, and other details imbued with historical context or background. Consequently, the method of directly projecting video and text into a common space to learn their similarity for contrastive learning may result in incomplete or inaccurate feature representation, thereby reducing retrieval accuracy. In response to technological advancements, cross-modal interaction-based methods (Gorti et al. 2022; Wang et al. 2024) have emerged. These methods extract features dynamically according to different text queries, allowing for more adaptive feature extraction. Although this approach addresses some of the aforementioned issues, it poses challenges for the practical application of retrieval due to its computational complexity. Specifically, the computational complexity of extracting dynamic features in retrieval tasks increases quadratically, with an order of $O(n^2)$, as the number of video-text pairs grows.

In view of the above problems, a natural question arises: Is there an algorithm that can dynamically calculate the similarity between videos and different texts with $O(n)$ complexity? As shown in Figure 1(b), this paper proposes a solution called Bidirectional Hierarchical Sliding Semantic Probe (BiHSSP), which calculates dynamic similarity between videos and texts with $O(n)$ complexity during retrieval. Taking text-to-video as an example, the core idea of this method is to retain the complete features of the video as much as possible, further extract key information from the text features, and obtain the Text Probe. Finally, use the Probe to slide on the original video features to calculate the cross-correlation index as the basis for the similarity between the video and the text. Specifically, we begin by extracting hierarchical text probes from the original textual features. These hierarchical probes, which vary in size, encapsulate the most critical components of the initial features. The hierarchical text probes are then employed to perform sliding probes across the video features, during which the cross-correlation indices are calculated. This sliding mechanism allows the BiHSSP framework to pinpoint the segments of the video that are most relevant to the corresponding text. Given the bidirectional nature of our method, sliding probe is also performed in the reverse direction, i.e., from video to text. This bidirectional sliding probe yields a comprehensive set of cross-correlation indices, which are further refined through a graph-based cross-correlation reasoning module. This refinement process aggregates the indices, ultimately producing a more nuanced and precise final similarity score. Additionally, our method can function as a plug-and-play module across various methods, thereby enhancing their performance. Experimental results demonstrate that our BiHSSP significantly outperforms the baseline.

The main contributions are as follows:

- We introduce a novel Bidirectional Hierarchical Sliding Semantic Probe (BiHSSP) framework for text-based video retrieval, which effectively addresses the issue of information imbalance between video and text. This framework achieves $O(n)$ computational complexity and enables dynamic similarity computation between video and text without relying on cross-modal interaction.
- We propose a plug-and-play similarity calculation method that can be seamlessly integrated into existing approaches to enhance their performance. This method leverages hierarchical probe and sliding mechanisms to dynamically align relevant video clips with corresponding textual descriptions. The similarity measurement is further refined through the integration of a graph-based cross-correlation reasoning module.
- We conduct extensive experimental validation on multiple benchmark datasets to empirically demonstrate the efficacy and efficiency of the proposed BiHSSP framework. Our approach consistently surpasses existing baseline methods in terms of both retrieval accuracy and computational efficiency.

Related Work

Text-Video Retrival. JSFusion (Yu, Kim, and Kim 2018) was among the first to investigate hierarchical similarities between video and text through the use of a convolutional decoder, thereby establishing a foundational benchmark for the task of text-video retrieval. Transformer-based methods (Dosovitskiy et al. 2021; Vaswani 2017; Deng et al. 2023; Dzabraev et al. 2021; Gabeur et al. 2020; Gao et al. 2021; Huang et al. 2023; Ji et al. 2023; Li et al. 2023a), which abstract multimodal data features via cross-attention mechanisms, have since led to significant performance improvements. Recent advancements have further leveraged CLIP (Radford et al. 2021) for semantic extraction in text-video retrieval tasks (Gorti et al. 2022; Lei et al. 2021; Luo et al. 2022; Wu et al. 2023; Xu et al. 2021; Xue et al. 2023; Zhao et al. 2022). For instance, CLIP4Clip (Luo et al. 2022) explores the transferability of the pre-trained CLIP model to text-video retrieval. To address domain gaps, CLIP-ViP (Xue et al. 2023) employs video post-pretraining, achieving state-of-the-art results. Similarly, Cap4Video (Wu et al. 2023) enhances retrieval by introducing additional captions, thereby demonstrating the value of augmented data. TEACHTEXT (Croitoru et al. 2021) improves retrieval performance through the integration of multiple text encoders. Additionally, DiffusionRet (Jin et al. 2023) advances the field by incorporating diffusion models into text-video retrieval tasks. The inclusion of additional modalities, such as audio (Akbari et al. 2021; Ibrahimi et al. 2023; Lin et al. 2022; Liu et al. 2022a; Miech, Laptev, and Sivic 2018), has also garnered increased attention. The proposed method focuses on learning expressive and robust text embeddings, achieving substantial improvements without requiring the

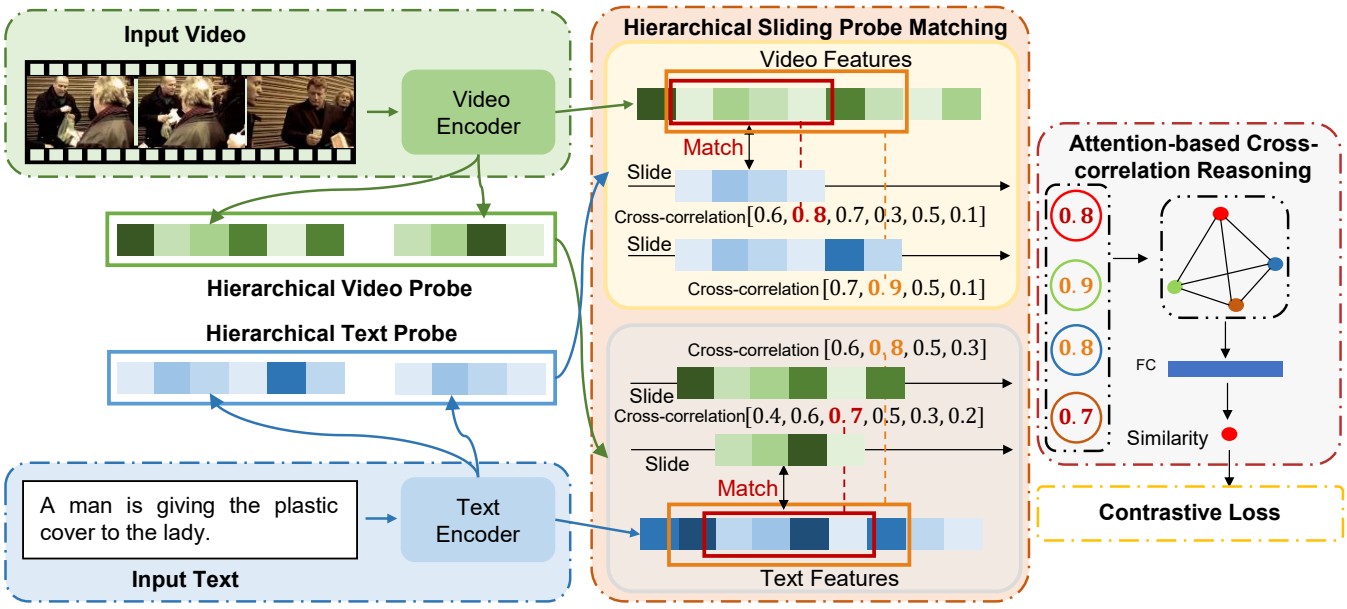

Figure 2: Overview of our Bidirectional Hierarchical Sliding Semantic Probe (BiHSSP) framework for text-video retrieval. As our method is bidirectional, we describe the process of text-to-video matching, with the video-to-text process being fully symmetrical. Given a video-text pair (v, t), we first utilize CLIP-based encoders to extract visual and textual features. We then introduce a hierarchical semantic probes learning module that generates multi-scale text probes—vectors of varying scales that encapsulate the textual features. Next, the Hierarchical Sliding Probe Matching module computes the cross-correlation between the original video features and the hierarchical text probes by sliding them across each other. Finally, we consolidate all cross-correlation indices and refine them through the Attention-based Cross-correlation Reasoning module to determine the final similarity score.

post-pretraining of CLIP on additional video data. Remarkably, the proposed approach even outperforms prior methods that rely on post-processing techniques (Bogolin et al. 2022; Cheng et al. 2021).

Text and Video Representation Learning. This work builds upon the CLIP model (Radford et al. 2021), recognized for its effective semantic extraction capabilities. Leveraging CLIP, existing approaches have primarily focused on refining video and text representations for retrieval tasks (Dong et al. 2022; Fang et al. 2022; Gorti et al. 2022; Guan et al. 2023; Han et al. 2022; Jin et al. 2022; Li et al. 2023b; Liu et al. 2019; Pei et al. 2023). For instance, TS2-Net (Liu et al. 2022c) models fine-grained temporal visual features, demonstrating promising results. X-Pool (Gorti et al. 2022) enhances semantic similarity by utilizing text-conditioned feature fusion across frames, resulting in more cohesive embeddings. Fine-grained semantic modeling is further explored in methods like PIDRo (Guan et al. 2023) and ProST (Li et al. 2023b), both of which achieve notable performance improvements. Additionally, UATVR (Fang et al. 2023) introduces an innovative approach by recognizing and modeling uncertainties within both modalities.

Preliminaries

Pipeline of Text-Video Retrieval. We denote the text as t and the raw video clip as v. The task of text-video retrieval involves learning embeddings for both text and video in a shared space, represented as $\mathbf{t}, \mathbf{v} \in \mathbb{R}^d$, where d is the feature dimension. A similarity function $s(\mathbf{t}, \mathbf{v})$, such as cosine similarity, is then used to measure relevance. Given a dataset with K text-video pairs, $\mathcal{D} = (t_k, v_k)_{k=1}^K$, a commonly used symmetric cross-entropy loss function minimizes the distance between relevant pairs and maximizes the distance between irrelevant ones. The loss functions for text-to-video and video-to-text retrieval are defined as:

$$\begin{aligned} \mathcal{L}_{t\to v} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\mathbf{t}_i,\mathbf{v}_i)\cdot\lambda}}{\sum_j e^{s(\mathbf{t}i,\mathbf{v}j)\cdot\lambda}}, \\ \mathcal{L}_{v\to t} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\mathbf{t}_i,\mathbf{v}_i)\cdot\lambda}}{\sum_j e^{s(\mathbf{t}_j,\mathbf{v}_i)\cdot\lambda}}, \end{aligned} \tag{1}$$

where N is the batch size, and λ is a learnable scaling factor. The embeddings, $\mathbf{t}_i$ and $\mathbf{v}_i/\mathbf{v}j$, are produced by feature extractors with learnable parameters. The overall loss function $\mathcal{L}ce$ is defined as:

$$\mathcal{L}ce = \frac{1}{2}(\mathcal{L}_{t\to v} + \mathcal{L}_{v\to t}). \tag{2}$$

The loss function reaches zero when all text-video pairs in a batch are perfectly aligned, i.e., $s(\mathbf{t}_i, \mathbf{v}_i) = 1$ for relevant pairs and $s(\mathbf{t}_i, \mathbf{v}_j) = 0$ for irrelevant pairs where $i \neq j$.

As shown in Figure 1, the information content in videos

and texts is often unequal, making accurate similarity computation critical. Achieving this is challenging and depends heavily on the quality of embeddings ($\mathbf{t}$ and $\mathbf{v}$) and how to measure the similarity between video and text embeddings.

Feature Extraction. The recent advancements in CLIP have significantly influenced text-video retrieval methods, leading us to primarily focus on CLIP-based approaches in this study. Given a video consisting of T frames, denoted as $v = [f_1, \ldots, f_T]$, a common protocol is to sample T' frames and process them through CLIP, yielding T' distinct frame embeddings $\mathbf{f}_i$, where $i = [1, \ldots, T']$.

Let Φ_v and Φ_t represent the image and text encoders of CLIP, respectively. The feature extraction process is defined as follows:

$$\mathbf{f}_i = \Phi_v(f_i), i \in [1, \ldots, T']; \quad \mathbf{t} = \Phi_t(t), \tag{3}$$

where $\mathbf{f}_i \in \mathbb{R}^d$. Based on the frame embeddings $[\mathbf{f}_i, \ldots, \mathbf{f}_{T'}]$, various strategies have been developed in previous works to compute the final video embedding $\mathbf{V}$ for similarity measurement:

$$\mathbf{v} = \Psi([\mathbf{f}_i, \ldots, \mathbf{f}_{T'}], t), \tag{4}$$

where Ψ denotes the feature fusion module that captures video semantics through frame-text interaction at different granularities or via temporal modeling.

Based on the choice of fusion module Ψ, previous methods can be broadly categorized into two types: Temporal Fusion-based Methods and Cross-modal Fusion-based Methods. Temporal fusion-based methods are advantageous due to their high efficiency, as the computational complexity increases linearly with the number of video-text pairs, following $O(n)$. However, these methods rely on fixed video features, making it challenging to align related text descriptions from varying perspectives using only cosine similarity. On the other hand, cross-modal fusion-based methods enable dynamic cross-modal interactions, allowing for the generation of adaptive video features corresponding to different texts. This method works well because it partially solves the problem of different amounts of information between video and text. However, the computational complexity for these methods scales quadratically with the number of video-text pairs, following $O(n^2)$.

Motivation. *It is intuitive to propose a method that is faster while dynamically calculating the similarity between features of different modalities.* Our analysis indicates that achieving $O(n)$ complexity primarily requires avoiding cross-modal fusion, which results in fixed video features. Consequently, previous methods relying on cosine similarity are unable to dynamically compute the similarity between video and text. To address this, we propose to develop a new similarity calculation approach that can dynamically select different parts of the video features for similarity assessment. This would enable us to realize the proposed idea effectively.

Methods

We main focus to design a new framework for calculating the similarity between video and text in a retrieval context while maintaining $O(n)$ computational complexity. The overview of our BiHSSP is illustrated in Figure.2. In this paper, we focus on introducing the process of text-to-video matching. Since our method is bidirectional, the process from video to text is completely symmetrical.

Hierarchical Semantic Probe Learning

Given a video-text pair (v, t), we first extract the video feature $\mathbf{v}$ and text feature $\mathbf{t}$ by video encoder Φ_v and text encoder Φ_t respectively. In order to dynamically calculate the similarity between video and text while avoiding cross-modal fusion, we want to dynamically select different parts of the video features to participate in the calculation when calculating the similarity of different texts. Therefore, we need to get some shorter text vectors, which we call Text Probe. So, we propose the Hierarchical Semantic Probe Learning module to get a set of multi-scale text probe $[\mathbf{t}_{p_1}, ..., \mathbf{t}_{p_n}]$ from text features $\mathbf{t}$. The network consists of a fully connection layer for each probe.

$$\mathbf{t}_{p_i} = Linear(\mathbf{t}), i \in [1, ..., n], \tag{5}$$

where $\mathbf{t}_{p_i} \in \mathbb{R}^{d^i}$, d^i is smaller than d and various for different probe $\mathbf{t}_{p_i}$. n indicate the number of the text probes.

Symmetrically, we can also get a set of hierarchical video probes $[\mathbf{v}_{p_1}, ..., \mathbf{v}_{p_m}]$ for videos.

Hierarchical Sliding Probe Matching

The Hierarchical Sliding Probe Matching (HSPM) Module is a critical component of our proposed BiHSSP framework. This module is designed to address the challenge of effectively aligning video and text features by leveraging a dynamic and hierarchical matching strategy.

Sliding Probe Mechanism. Central to the HSPM module is the sliding probe mechanism, which systematically matches the hierarchical text probes against the video features. We start with a single probe.

In this process, the text probe $\mathbf{t}_{p_i}$ slides along the sequence of video features $\mathbf{v}$. At each sliding position k, the cross-correlation is computed to measure the similarity between the text probe and the corresponding segment of video features. Formally, the cross-correlation at the k^{th} position is given by:

$$C_{p_i}^{(k)} = \mathrm{g}^{t \to v}(\mathbf{t}_{p_i} \cdot \mathbf{v}^{(k)}) = \sum_{l=1}^{d^i} t_{p_i}^{(l)} \cdot v^{(k,l)}, \tag{6}$$

where $g(\cdot)$ is the cross-correlation function, and d^i is the dimension of the text probe, and $t_{p^i}^{(l)}$ and $v^{(k,l)}$ represent the elements of the text and video feature vectors, respectively. This cross-correlation $C_{p_i}^{(k)}$ quantifies the alignment between the text probe and the video features at each sliding position. As the text probe slides across the video features, a series of cross-correlation scores $\mathbf{C}_{p_i}^{t \to v} = [C_{p_i}^{(1)}, C_{p_i}^{(2)}, \ldots, C_{p_i}^{(K)}]$ is generated, capturing the degree of match at various locations within the video. We take the maximum value of $\mathbf{C}_{p_i}^{t \to v}$ as the similarity between the text probe $\mathbf{t}_{p_i}$ and the video feature $\mathbf{v}$. This mechanism allows the module to dynamically select and emphasize the most

method	MSRVTT Retrieval					DiDeMo Retrieval				
	R@1↑	R@5↑	R@10↑	MdR↓	MeanR↓	R@1↑	R@5↑	R@10↑	MdR↓	MeanR↓
CLIP-ViT-B/32										
X-pool (Gorti et al. 2022)	46.9	72.8	82.2	**2.0**	14.3	44.6	73.2	8**2.0**	**2.0**	15.4
TS2-Net(Liu et al. 2022c)	47.0	74.5	83.8	-	13.0	41.8	71.6	8**2.0**	**2.0**	14.8
PIDRo (Guan et al. 2023)	48.2	74.9	83.3	**2.0**	12.6	-	-	-	-	-
ProST (Li et al. 2023b)	48.2	74.6	83.4	**2.0**	12.4	44.9	72.7	82.7	**2.0**	13.7
DiffusionRet(Jin et al. 2023)	**49.0**	**75.2**	82.7	**2.0**	12.1	46.7	**74.7**	82.7	**2.0**	14.3
UATVR(Fang et al. 2023)	47.5	73.9	83.5	**2.0**	12.3	43.1	71.8	82.3	**2.0**	15.1
CLIP$_{finetune}$(Radford et al. 2021)	43.1	70.4	80.8	**2.0**	16.2	43.4	70.2	80.6	**2.0**	17.5
CLIP + BiHSSP	45.8	73.1	83.8	**2.0**	14.0	45.3	74.1	82.3	**2.0**	14.5
STAN(Liu et al. 2023a)	46.6	7**2.0**	82.1	**2.0**	-	46.2	70.4	80.0	**2.0**	-
STAN + BiHSSP	48.1	74.0	**84.1**	**2.0**	**12.1**	**47.9**	73.6	**83.0**	**2.0**	**12.6**
CLIP-ViT-B/16										
X-pool(Gorti et al. 2022)	48.2	73.7	82.6	2.0	12.7	47.3	74.8	82.8	2.0	14.2
UATVR(Fang et al. 2023)	45.8	76.3	**85.5**	1.0	12.4	45.8	73.7	83.3	2.0	13.5
ProST(Li et al. 2023b)	49.5	75.0	84.0	2.0	11.7	47.5	75.5	84.4	2.0	12.3
CLIP$_{finetune}$	45.3	73.3	83.0	2.0	13.0	44.8	75.1	83.2	2.0	13.0
CLIP + BiHSSP	47.3	**77.9**	84.2	2.0	12.2	47.8	77.5	85.3	2.0	12.6
STAN (Liu et al. 2023a)	50.0	75.2	84.1	1.5	-	49.4	74.9	84.5	1.0	-
STAN + BiHSSP	**50.8**	75.9	84.4	**1.0**	**11.0**	**49.8**	**77.8**	**85.9**	**1.0**	**11.8**

Table 1: **Text-to-video comparison** on MSRVTT and DiDeMo. Bold denotes the best performance. "-": result is unavailable.

relevant segments of the video that align with the text description.

Hierarchical Matching Strategy. The Hierarchical Sliding Probe Matching module leverages hierarchical probes to ensure that both localized details and global context are incorporated into the similarity computation. This approach mitigates the risk of overly similar small local fragments influencing the overall similarity assessment.

Consequently, the complete cross-correlation scores can be computed as $C^{t\to v} = [\text{Max}(\mathbf{C}_{p_1}^{t\to v}), \ldots, \text{Max}(\mathbf{C}_{p_n}^{t\to v})]$. Similarly, the scores for the reverse direction, $C^{v\to t} = [\text{Max}(\mathbf{C}_{p_1}^{v\to t}), \ldots, \text{Max}(\mathbf{C}_{p_m}^{v\to t})]$, are calculated in an analogous manner.

Graph-based Cross-correlation Reasoning

Graph Construction. To facilitate more comprehensive cross-correlation reasoning, we construct a similarity graph to propagate cross-correlation information among potential alignments in both the text-to-video and video-to-text processes. Specifically, we concatenate all the cross-correlation scores as graph nodes, denoted as $\mathcal{N} = C^{t\to v}||C^{v\to t}$. The edge between two nodes, $\mathcal{N}_p$ and $\mathcal{N}_q$, is computed as:

$$e(\mathcal{N}_p, \mathcal{N}_q; \mathbf{W}_{in}, \mathbf{W}_{out}) = \frac{\exp((\mathbf{W}_{in}\mathcal{N}_p)(\mathbf{W}_{out}\mathcal{N}_q))}{\sum_q \exp((\mathbf{W}_{in}\mathcal{N}_p)(\mathbf{W}_{out}\mathcal{N}_q))} \tag{7}$$

where $\mathbf{W}_{in} \in \mathbb{R}^{m\times m}$ and $\mathbf{W}_{out} \in \mathbb{R}^{m\times m}$ are linear transformations applied to the incoming and outgoing nodes, respectively. The directed edges between nodes $\mathcal{N}_p$ and $\mathcal{N}_q$ enable efficient and complex information propagation for similarity reasoning.

Graph Reasoning. With the graph nodes and edges constructed, we perform similarity graph reasoning by iteratively updating the nodes and edges. The update process is defined as follows:

$$\hat{\mathcal{N}}_p^n = \sum_q e(\mathcal{N}_p^n, \mathcal{N}_q^n; \mathbf{W}_{in}^n, \mathbf{W}_{out}^n) \cdot \mathcal{N}_q^n \tag{8}$$

$$\mathcal{N}_p^{(n+1)} = \text{RELU}(\mathbf{W}_r^n \mathcal{N}_p^n) \tag{9}$$

where $\mathcal{N}_p^0$ and $\mathcal{N}_q^0$ are initialized from $\mathcal{N}$ at step $n = 0$, and $\mathbf{W}_r^n$, $\mathbf{W}_{in}^n$, and $\mathbf{W}_{out}^n$ are learnable parameters at each step. After each reasoning step, node $\mathcal{N}_p^n$ is replaced with $\mathcal{N}_p^{n+1}$.

This reasoning process is repeated for N steps, with the output of the global node at the final step serving as the refined similarity representation. This representation is then passed through a fully connected layer to infer the final similarity score. The reasoning model enhances the propagation of information between cross-correlations, capturing more comprehensive interactions and improving the accuracy of similarity predictions.

Experiment

Experimental Settings

Datasets. We adopt five benchmark datasets for the evaluation, including (1) **MSRVTT** (Xu et al. 2016) that contains 10K video clips, where each has 20 captions. We follow the 1K-A testing split (Liu et al. 2019). (2) **LSMDC** (Rohrbach et al. 2015) incorporating 118081 clips from 202 movies, where each one is paired with a text description. Following (Gabeur et al. 2020; Gorti et al. 2022), we adopt

method	MSVD Retrieval					ActivityNet Retrieval				
	R@1↑	R@5↑	R@10↑	MdR↓	MeanR↓	R@1↑	R@5↑	R@10↑	MdR↓	MeanR↓
CLIP-ViT-B/32										
X-pool(Gorti et al. 2022)	47.2	77.4	**86.0**	**2.0**	9.3	-	-	-	-	-
X-CLIP(Ma et al. 2022)	47.1	77.8	-	-	9.5	44.3	74.1	-	-	7.9
EMCL-Net(Jin et al. 2022)	42.1	71.3	81.1	**2.0**	17.6	41.2	72.7	83.6	**2.0**	8.6
PIDRo(Guan et al. 2023)	47.5	77.5	**86.0**	**2.0**	9.2	44.9	74.5	86.1	**2.0**	6.4
DiffusionRet(Jin et al. 2023)	46.6	75.9	84.1	**2.0**	15.6	45.8	75.6	86.3	**2.0**	6.5
CLIP$_{finetune}$(Radford et al. 2021)	46.2	76.1	84.6	**2.0**	10.0	40.5	72.4	84.1	**2.0**	7.4
CLIP + BiHSSP	47.9	**78.3**	85.9	**2.0**	9.2	43.1	70.5	81.2	**2.0**	12.4
STAN(Liu et al. 2023a)	46.7	72.5	83.6	**2.0**	-	43.8	73.9	86.2	**2.0**	6.5
STAN + BiHSSP	**48.1**	74.1	84.4	**2.0**	**9.1**	**45.8**	74.6	**86.9**	**2.0**	**6.4**

Table 2: **Text-to-video** comparison on MSVD and ActivityNet. Bold denotes the best performance. "-": result is unavailable.

method	R@1↑	R@5↑	R@10↑	MdR↓	MeanR↓
CLIP-ViT-B/32					
CLIP4Clip	42.7	70.9	80.6	**2.0**	11.6
CenterCLIP	42.8	71.7	82.2	**2.0**	10.9
X-Pool	44.4	73.3	84.0	**2.0**	9.0
TS2-Net	45.3	74.1	83.7	**2.0**	9.2
DiffusionRet	47.7	73.8	**84.5**	**2.0**	8.8
UATVR	46.9	73.8	83.8	**2.0**	8.6
CLIP	43.1	70.5	81.2	**2.0**	12.4
CLIP + BiHSSP	45.2	73.2	84.1	**2.0**	9.0
STAN	46.6	72.0	82.1	**2.0**	-
STAN + BiHSSP	**48.0**	**74.1**	83.5	**2.0**	9.0
CLIP-ViT-B/16					
X-Pool	46.4	73.9	84.1	**2.0**	8.4
TS2-Net	46.6	75.9	84.9	**2.0**	8.9
CenterCLIP	47.7	75.0	83.3	**2.0**	10.2
UATVR	48.1	76.3	85.4	**2.0**	8.0
CLIP	44.8	73.2	82.2	**2.0**	9.6
CLIP + BiHSSP	48.8	**76.8**	84.4	**2.0**	8.1
STAN	49.3	75.1	83.9	**2.0**	-
STAN + BiHSSP	**50.3**	75.5	**84.5**	**1.5**	**7.8**

Table 3: **Video-to-text** comparison on MSRVTT dataset. Bold denotes the best performance.

the testing data with 1000 videos. (3) **DiDeMo** (Anne Hendricks et al. 2017) consists of 10642 clips and 40543 captions in total. Following (Luo et al. 2022), all caption descriptions of a video are concatenated as a query to evaluate all methods. (4) **ActivityNet Captions**(Krishna et al. 2017) cconsists of densely annotated temporal segments of 20K YouTube videos. We use the 10K training split to train the model and report the performance on the 5K "val1" split.

Metrics. Recall at rank 1, 5, 10 (R@1, R@5, and R@10), Median Rank (MdR), and Mean Rank (MnR) are adopted to evaluate the retrieval performance.

Implementation Details. For the training, we set the batch size as 128 for both backbones and different datasets. The CLIP model is fine-tuned with a learning rate of 2e-7. We train the models for epochs with the AdamW (Loshchilov and Hutter 2019) optimizer. Following CLIP, we employ a cosine schedule (Loshchilov and Hutter 2017) with a warm-up proportion of 0.1. We uniformly sample 12 frames from the video clips upon different datasets. All the frames are resized to 224×224. We perform experiments on 4 3090 GPUS. We keep it consistent for our method by using batch size as 128 and frame number as 12 for all datasets. Our method can also be plug-and-play integrated into previous methods by replacing cosine similarity. *More details, results, and discussions are provided in supplementary.*

Why use cross-correlation? Cross-correlation serves as a fundamental metric for measuring the similarity between two sets of data, such as hierarchical text probes and video features in our framework. However, the process of sliding a window across the data to compute this similarity can be computationally expensive, especially when dealing with large-scale video-text retrieval tasks. To address this, we leverage the close relationship between cross-correlation and convolution. In essence, convolution can be viewed as a specific form of cross-correlation where the kernel is not flipped. Given that convolution operations are highly optimized in various deep learning frameworks, we implement cross-correlation using convolution operations. This approach allows us to take advantage of these optimizations, significantly enhancing the efficiency of the sliding window process. Although we use convolution in practice, the operation remains, in essence, a cross-correlation, ensuring that the similarity metrics are accurately computed with greater computational efficiency.

Permformance Comparison

We compare the text-to-video retrieval performance of BiHSSP with previous methods on four benchmark dataset. We find that our BiHSSP not only improves the baseline CLIP (Radford et al. 2021) and STAN (Liu et al. 2023a) by a large margin on all metrics, but also achieves state-of-the-art performance compared with most recent methods. As shown in Table 1, our methods improves DiffusionRet (Jin et al. 2023) 1.2% on DiDeMo datset. In Table 2, we show the results of out method on MSVD and Activ-

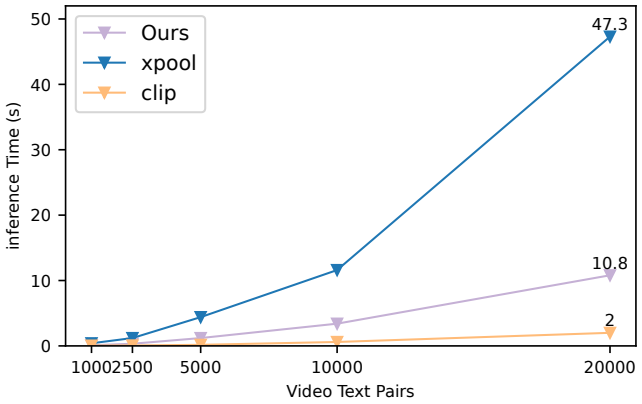

Figure 3: Inference time for text-video retrieval on single GPU with different video candidates (lower the better).

ity Captions datasets. Our model outperforms recently proposed state-of-the-art methods on multiple metrics on text-to-video retrieval tasks. Table 3 shows the results of our method for video-to-text task on MSRVTT dataset. Massive experiments on text-video retrieval tasks demonstrate the superiority and flexibility of our method.

Ablation Study

Inference efficiency analysis. In evaluating the performance of models used in search engines with large-scale databases for video or text queries, the efficiency at the inference stage is of equal importance as the accuracy of caption or image retrieval. As shown in Figure 3, we conducted a comparative analysis of the inference time for image-text retrieval using a single GPU, evaluating two representative methods: CLIP(Radford et al. 2021), which employs a temporal fusion approach, and X-Pool(Gorti et al. 2022), which utilizes a cross-modal fusion technique. Both methods are considered fundamental and fastest within their respective categories. It is obvious that temporal fusion based methods are much faster than cross modal fusion based methods. When the number of candidate videos is small (typically 1000), X-Pool(Gorti et al. 2022) seems to have a similar retrieval time to temporal fusion based methods. However, the time advantage of temporal fusion based methods increases with the number of candidate images. With an increase in the number of candidate images (e.g. 20,000), the retrieval time of our proposed method decreases by 26.5 seconds compared to XPool, and as the number of video-text pairs increases, the gap will be even greater. This is a significant advantage of our proposed method over other methods.

Effective of multi-caption retrieval. Since the test sets in Table 1 and Table 2 have only one text description for each image, in order to verify that our method can effectively retrieve scenes with different descriptions of the same video, we conducted corresponding ablation experiments. On our existing MSRVTT dataset, we selected all videos in the MSRVTT-1kA test set and the 20 text descriptions corresponding to each video, and used 20,000 video-text pairs for

method	R@1↑	R@5↑	R@10↑	MeanR↓
CLIP$_{finetune}$	40.6	68.7	78.4	17.9
+BiHSSP	**44.8**$^{+4.2}$	**72.5**$^{+3.8}$	**82.9**$^{+3.8}$	**14.6**$^{-3.3}$

Table 4: Effect of multi-caption retrieval in MSRVTT-1kA test dataset. We measure the performance using all 20,000 video-text pairs, i.e., 20 text descriptions per video.

$t \rightarrow v$ Sliding	$v \rightarrow t$ Sliding	Graph Reasoning	Text→video R@1↑	R@5↑	MeanR↓
✓			42.9	70.3	16.2
	✓		42.6	70.1	16.4
✓	✓		44.6	71.0	15.3
✓	✓	✓	**45.8**	**73.1**	**14.0**

Table 5: Ablation study of sliding probe matching module and graph reasoning module on MSRVTT text-to-video retrieval task.

experiments. The experimental results are shown in Table 4. Compared with CLIP which uses cosine similarity, using our framework can increase the retrieval results by 4.4% in R@1, which demonstrates our method can effectively solve the problem of unequal information in video and text, and can better retrieve scenes with multiple descriptions.

Ablation study of modules. We conduct the ablation study of the core module of the proposed methods. In this experiment, when the graph reasoning module is not used, we just sum all the cross-correlation scores as the final similarity. As shown in Table 5 , when we only use unidirectional sliding probe matching, the performance of R@1 is 42.9% and 4.26%, respectively, but when we use bidirectional probe matching, R@1 reaches 44.6. In addition, graph reasoning will further improve the effect of our method.

Conclusion

In this paper, we introduced the Bidirectional Hierarchical Sliding Semantic Probe (BiHSSP) framework, a novel solution for text-based video retrieval that addresses the challenges of information imbalance and computational inefficiency in existing methods. The BiHSSP framework efficiently achieves $O(n)$ complexity while dynamically calculating the similarity between video and text without necessitating cross-modal interaction. A key innovation of our approach is the introduction of hierarchical text probes, which, through a sliding mechanism, effectively match video clips with their corresponding text descriptions. A significant advantage of the BiHSSP framework is its *plug-and-play* design, allowing it to be easily integrated into existing retrieval methods. This modularity not only simplifies the adoption of our approach but also improves the performance of current systems by incorporating our dynamic similarity calculation method. Extensive experimental evaluations on multiple benchmark datasets validate the effectiveness and efficiency of the proposed BiHSSP framework.

Acknowledgments

This work was partially supported by the National Major Scientific Instruments and Equipments Development Project of National Natural Science Foundation of China under Grant 62427820, the National Science Foundation of China under Grant 62376175, the 111 Project under Grant B21044, the Science Fund for Creative Research Groups of Sichuan Province Natural Science Foundation under Grant 2024NSFTD0035 and the Sichuan Science and Technology Program under Grant 2021ZDZX0011.

References

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; and Gong, B. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS*, 34: 24206–24221.

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *ICCV*, 5803–5812.

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 *ICCV*, 1728–1738.

Bogolin, S.-V.; Croitoru, I.; Jin, H.; Liu, Y.; and Albanie, S. 2022. Cross modal retrieval with querybank normalisation. In *CVPR*, 5194–5205.

Cheng, X.; Lin, H.; Wu, X.; Yang, F.; and Shen, D. 2021. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*.

Croitoru, I.; Bogolin, S.-V.; Leordeanu, M.; Jin, H.; Zisserman, A.; Albanie, S.; and Liu, Y. 2021. Teachtext: Cross-modal generalized distillation for text-video retrieval. In *ICCV*, 11583–11593.

Deng, C.; Chen, Q.; Qin, P.; Chen, D.; and Wu, Q. 2023. Prompt switch: Efficient clip adaptation for text-video retrieval. In *ICCV*, 15648–15658.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022. Partially relevant video retrieval. In *ACM MM*, 246–257.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Dzabraev, M.; Kalashnikov, M.; Komkov, S.; and Petiushko, A. 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. In *CVPR*, 3354–3363.

Fang, B.; Wu, W.; Liu, C.; Zhou, Y.; Song, Y.; Wang, W.; Shu, X.; Ji, X.; and Wang, J. 2023. Uatvr: Uncertainty-adaptive text-video retrieval. In *ICCV*, 13723–13733.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2022. Clip2video: Mastering video-text retrieval via image clip. *IEEE T-MM*.

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal transformer for video retrieval. In *ECCV*, 214–229. Springer.

Gao, Z.; Liu, J.; Chen, S.; Chang, D.; Zhang, H.; and Yuan, J. 2021. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*, 1(2): 6.

Gorti, S. K.; Vouitsis, N.; Ma, J.; Golestan, K.; Volkovs, M.; Garg, A.; and Yu, G. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 5006–5015.

Guan, P.; Pei, R.; Shao, B.; Liu, J.; Li, W.; Gu, J.; Xu, H.; Xu, S.; Yan, Y.; and Lam, E. Y. 2023. Pidro: Parallel isomeric attention with dynamic routing for text-video retrieval. In *ICCV*, 11164–11173.

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

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

Huang, J.; Li, Y.; Feng, J.; Wu, X.; Sun, X.; and Ji, R. 2023. Clover: Towards a unified video-language alignment and fusion model. In *CVPR*, 14856–14866.

Ibrahimi, S.; Sun, X.; Wang, P.; Garg, A.; Sanan, A.; and Omar, M. 2023. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *ICCV*, 12054–12064.

Ji, K.; Liu, J.; Hong, W.; Zhong, L.; Wang, J.; Chen, J.; and Chu, W. 2022. Cret: Cross-modal retrieval transformer for efficient text-video retrieval. In *SIGIR*, 949–959.

Ji, Y.; Tu, R.; Jiang, J.; Kong, W.; Cai, C.; Zhao, W.; Wang, H.; Yang, Y.; and Liu, W. 2023. Seeing what you miss: Vision-language pre-training with semantic completion learning. In *CVPR*, 6789–6798.

Jin, P.; Huang, J.; Liu, F.; Wu, X.; Ge, S.; Song, G.; Clifton, D.; and Chen, J. 2022. Expectation-maximization contrastive learning for compact video-and-language representations. *NeurIPS*, 35: 30291–30306.

Jin, P.; Li, H.; Cheng, Z.; Li, K.; Ji, X.; Liu, C.; Yuan, L.; and Chen, J. 2023. Diffusionret: Generative text-video retrieval with diffusion model. In *ICCV*, 2470–2481.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *ICCV*, 706–715.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *ECCV*, 201–216.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 7331–7341.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. *EMNLP*.

Li, L.; Gan, Z.; Lin, K.; Lin, C.-C.; Liu, Z.; Liu, C.; and Wang, L. 2023a. Lavender: Unifying video-language understanding as masked language modeling. In *CVPR*, 23119–23129.

Li, P.; Xie, C.-W.; Zhao, L.; Xie, H.; Ge, J.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023b. Progressive spatio-temporal prototype matching for text-video retrieval. In *ICCV*, 4100–4110.

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

Liu, R.; Huang, J.; Li, G.; Feng, J.; Wu, X.; and Li, T. H. 2023a. Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In *CVPR*, 6555–6564.

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

Liu, Y.; Chen, H.; Huang, L.; Chen, D.; Wang, B.; Pan, P.; and Wang, L. 2022a. Animating images to transfer clip for video-text retrieval. In *SIGIR*, 1906–1911.

Liu, Y.; Liu, H.; Wang, H.; and Liu, M. 2022b. Regularizing visual semantic embedding with contrastive learning for image-text matching. *IEEE Signal Processing Letters*, 29: 1332–1336.

Liu, Y.; Liu, H.; Wang, H.; Meng, F.; and Liu, M. 2023b. BCAN: Bidirectional correct attention network for cross-modal retrieval. *IEEE T-NNLS*.

Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; and Jin, Q. 2022c. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, 319–335. Springer.

Loshchilov, I.; and Hutter, F. 2017. Sgdr: Stochastic gradient descent with warm restarts. *ICLR*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. *ICLR*.

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.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, 638–647.

Miech, A.; Laptev, I.; and Sivic, J. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

Pei, R.; Liu, J.; Li, W.; Shao, B.; Xu, S.; Dai, P.; Lu, J.; and Yan, Y. 2023. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *CVPR*, 18983–18992.

Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *CVPR*, 27197–27206.

Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *ACM MM*, 4948–4956.

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 *ICML*, 8748–8763. PMLR.

Rohrbach, A.; Rohrbach, M.; Tandon, N.; and Schiele, B. 2015. A dataset for movie description. In *CVPR*, 3202–3212.

Vaswani, A. 2017. Attention is all you need. *NeurIPS*.

Wang, J.; Sun, G.; Wang, P.; Liu, D.; Dianat, S.; Rabbani, M.; Rao, R.; and Tao, Z. 2024. Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval. In *CVPR*, 16551–16560.

Wu, W.; Luo, H.; Fang, B.; Wang, J.; and Ouyang, W. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *CVPR*, 10704–10713.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *EMNLP*.

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

Xue, H.; Sun, Y.; Liu, B.; Fu, J.; Song, R.; Li, H.; and Luo, J. 2023. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *ICLR*.

Yu, Y.; Kim, J.; and Kim, G. 2018. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 471–487.

Zhao, S.; Zhu, L.; Wang, X.; and Yang, Y. 2022. Centerclip: Token clustering for efficient text-video retrieval. In *SIGIR*, 970–981.