This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Progressive Spatio-Temporal Prototype Matching for Text-Video Retrieval

Pandeng Li[1]*, Chen-Wei Xie[2], Liming Zhao[2], Hongtao Xie[1]†, Jiannan Ge[1],
Yun Zheng[2], Deli Zhao[2], Yongdong Zhang[1]
[1] University of Science and Technology of China, [2]DAMO Academy, Alibaba Group
{lpd, gejn}@mail.ustc.edu.cn, {htxie, zhyd73}@ustc.edu.cn
{eniac.xcw, lingchen.zlm, zhengyun.zy}@alibaba-inc.com, zhaodeli@gmail.com

Abstract

The performance of text-video retrieval has been significantly improved by vision-language cross-modal learning schemes. The typical solution is to directly align the global video-level and sentence-level features during learning, which would ignore the intrinsic video-text relations, i.e., a text description only corresponds to a spatio-temporal part of videos. Hence, the matching process should consider both fine-grained spatial content and various temporal semantic events. To this end, we propose a text-video learning framework with progressive spatio-temporal prototype matching. Specifically, the matching process is decomposed into two complementary phases: object-phrase prototype matching and event-sentence prototype matching. In the object-phrase prototype matching phase, the spatial prototype generation mechanism predicts key patches or words, which are aggregated into object or phrase prototypes. Importantly, optimizing the local alignment between object-phrase prototypes helps the model perceive spatial details. In the event-sentence prototype matching phase, we design a temporal prototype generation mechanism to associate intra-frame objects and interact inter-frame temporal relations. Such progressively generated event prototypes can reveal semantic diversity in videos for dynamic matching. Validated by comprehensive experiments, our method consistently outperforms the state-of-the-art methods on four video retrieval benchmark.[1]

1. Introduction

Understanding multimodal information [13, 37, 40, 76, 6, 79, 41, 12, 53, 54] is an essential way for humans to perceive the world. As a fundamental task in multimodal learning [25, 62, 60], Text-Video Retrieval (TVR) [22, 70, 68, 24] has garnered huge interest with the rapid development of short video platforms. TVR [45, 21] aims to search semantically relevant videos based on user-entered text queries. Unfortunately, the inherent modality gap phenomenon [42, 61, 38, 3, 67] increases the difficulty of associating multimodal data. Towards such a concern, pioneering works [47, 65, 73] usually exploit multiple unimodal pre-trained models to extract features, and then use metric learning strategies [14] to strengthen the modality alignment in the joint space. However, there are large differences in the initial distributions of multiple unimodal offline features, which inevitably brings the feature fusion challenge to affect the retrieval results.

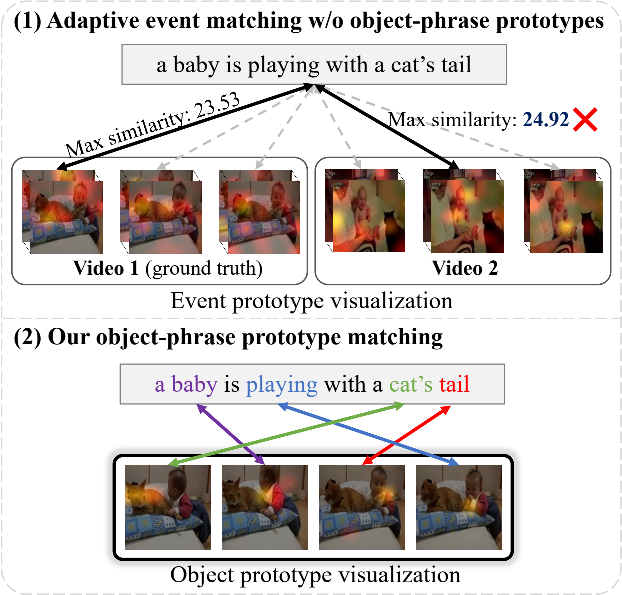
(1) Adaptive event matching w/o object-phrase prototypes
a baby is playing with a cat's tail
Max similarity: 23.53
Max similarity: 24.92
Video 1 (ground truth)
Video 2
Event prototype visualization
(2) Our object-phrase prototype matching
a baby is playing with a cat's tail
Object prototype visualization

Figure 1. Visualization of event and object prototypes learned by 1) adaptive event matching w/o object-phrase prototypes and 2) object-phrase prototype matching. Object prototypes can focus on local patches (*e.g.*, tail) and complement event-level matching.

Recently, encouraged by the success of vision-language pre-training [50, 66, 34, 27, 72, 35, 4], a series of canonical works [51, 20, 44, 16] are proposed by transferring knowl-

*Interns at DAMO Academy, Alibaba Group

†Corresponding author

[1] Code is available at https://github.com/IMCCretrieval/ProST.

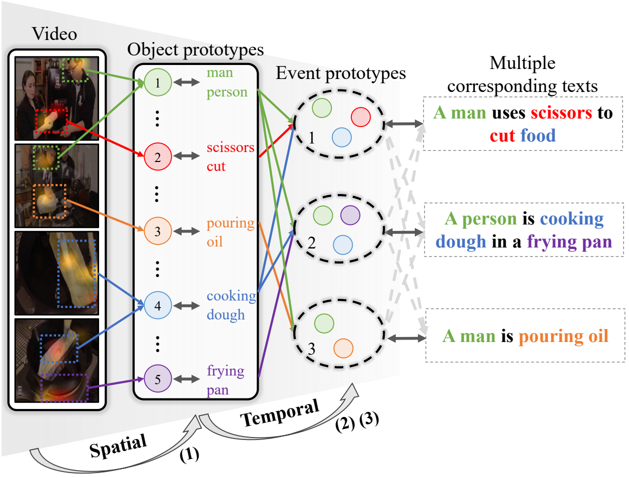

Figure 2. The progressive prototype generation process: 1) focus on patch-level spatial details; 2) aggregate intra-frame object prototypes; 3) interact inter-frame relations temporally as event prototypes, corresponding to multiple semantics.

edge from cross-modal pre-trained models (*e.g.*, CLIP [57]) to the TVR task. Among them, some methods [51, 1, 28] learn to encode text or video data as a single embedding and perform global alignment between the sentence and the whole video. Subsequent approaches [48, 78] employ frame-level or segment-level matching by taking into account the different importance of video content. These methods try to encode the video into a single semantic content described by corresponding text, establishing one-to-one relations. However, videos contain rich visual elements [39] and a text description may only correspond to a spatio-temporal part of the video, dynamic and fine-grained matching is needed beyond the global video-level or frame-level matching. Specifically, feature encoding should characterize different semantic information to support dynamic partial alignment during the matching process. Recently, Lin *et al.* [43] attempt to address the dynamic matching problem by temporally segmenting videos into multiple semantic events, so that different text descriptions can adaptively match event prototypes. Nevertheless, various event prototypes are still inadequate to model the relations between fine-grained objects and content words, lacking the perception of local details. As shown in Fig. 1(1), the correct video corresponding to the text should focus on "tail", but the event prototype corresponding to the maximum similarity may suffer from scene bias [8, 63] and ignore this small object. Based on the above observation, an effective event-sentence matching process needs to emphasize fine-grained local spatial matching. This also suggests that the generation of event prototypes should be bootstrapped from the semantic combination of important objects. Therefore, we should construct event prototypes progressively, *i.e.*, explicitly preserve the local information learned in the spatial matching stage, and then perform temporal aggregation.

To this end, we propose a novel Progressive Spatio-Temporal Prototype Matching (ProST) framework, which decomposes the matching process into complementary object-phrase and event-sentence prototype alignments. In the object-phrase prototype matching phase, a spatial prototype generation mechanism is first developed to focus on key image patches or word tokens. To prevent massive video redundancy, we use sparse weights to filter irrelevant background interference and aggregate important patch or word tokens into object or phrase prototypes. As shown in Fig. 1(2), optimal local responses can be found by the maximum similarity between object and phrase prototypes to help the model perceive spatial details.

In the event-sentence prototype matching phase, a temporal prototype generation mechanism is designed to decode various event prototypes for dynamic semantic alignment. Specifically, since the spatial details should not be ignored in the event generation process, we send object prototypes to the frame and event decoder for progressive prototype learning. Multiple different event prototypes are generated to reveal the semantic diversity in videos and support dynamic matching. Finally, as shown in Fig. 2, different text sentences can match the most appropriate event prototype learned from the video. Overall, ProST encodes the rich video content by generating object-level, frame-level and event-level prototypes in a progressive manner, and constructs dynamic and fine-grained spatio-temporal matching strategies. ProST not only strengthens the interaction of local spatial content, but also considers the intrinsic relations where video and text are only partially aligned.

The contributions of this paper are threefold: 1) we propose a novel framework, ProST, to decompose the matching process into complementary object-phrase and event-sentence prototype alignments; 2) two prototype generation mechanisms are developed to learn sparse spatial and dynamic temporal information respectively, which can fully explore fine-grained local details and video semantic diversity; 3) extensive experiments demonstrate that ProST outperforms state-of-the-art methods on MSRVTT, DiDeMo, VATEX and LSMDC datasets.

2. Related Work

Text-video retrieval. Existing works can be divided into two categories: offline feature based [47, 11, 77, 26] and end-to-end training [1, 51, 48, 43, 31, 29, 30, 28] methods.

Early methods [69, 11] extract video-text features offline, and then perform embedding alignment in a common space. To enhance representation quality, CE [47], MMT [17] and HiT [46] employ multiple unimodal experts (OCR, speech, etc.) to enrich video features. Teach-Text [10] adopts multiple text encoders to gather text information and obtain more credible alignments. Besides, HGR [5] and T2vlad [65] build semantic graphs or shared centers to explore global-local relations. However, these methods employ additional analysis tools or experts to preprocess data, which increases the complexity and limits flexibility. Unlike the global-local alignment explored on multiple experts [65, 46], we consider finer-grained patch-level visual content without requiring additional experts.

Current mainstream works benefit from large-scale vision-language pre-training [57, 19]. ClipBERT [33] and Frozen [1] first propose efficient end-to-end pre-training schemes. MCQ [18] builds cross-modal associations by predicting verb or noun features. Recently, CLIP4clip [51] uses CLIP [57] as the backbone network, significantly improving TVR performance and inspiring a series of works [48, 15, 20, 78, 28]. Centerclip [78] conducts segment-level clustering, which reduces token redundancy and computational overhead. TS2Net [48] designs the token shift module to realize the perception of local movement between frames. Xpool [20] uses text as a condition to guide the aggregation of video tokens. But the limitation is that the video encoding process must involve both modal data.

Adaptive cross-modal retrieval. Recently, the correspondence ambiguity problem (*i.e.*, a text may only correspond to a part of the video or image) has raised concerns in the cross-modal retrieval community [9, 59, 43]. In text-image retrieval, PVSE [59] employs the self-attention mechanism to explore different local parts of instances and generates K candidate features for adaptive matching. PCME [9] samples the text-image probabilistic embeddings multiple times from the learned Gaussian distribution to explore one-to-many multiplicity relationships. In text-video retrieval, TVMM [43] utilizes fully connected layers to directly aggregate all tokens into multiple event prototypes for text-adaptive event matching. However, fine-grained object-level information is not well exploited in the event matching process of TVMM, while the proposed progressive spatio-temporal matching framework could simultaneously consider local spatial details and dynamic temporal relations.

3. Method

Given a dataset consisting of n videos $\mathcal{V} = \{\boldsymbol{v}_i\}_{i=1}^n$ and their corresponding m captions $\mathcal{T} = \{\boldsymbol{t}_i\}_{i=1}^m$, Text-Video Retrieval (TVR) aims to learn a function $\boldsymbol{s}(\boldsymbol{t}_i, \boldsymbol{v}_i)$ to effectively measure the similarity between modalities. For a text query $\boldsymbol{t}_i$, we can rank all videos in the dataset according to the similarity scores. Ideally, the cross-modal similarity (*e.g.*, cosine similarity) between paired cross-modal instances should be greater than that of unpaired instances: $\boldsymbol{s}(\boldsymbol{t}_i, \boldsymbol{v}_i) > \boldsymbol{s}(\boldsymbol{t}_i, \boldsymbol{v}_j)$. This requires the model to learn powerful text encoding network $\boldsymbol{f}_t : \mathcal{T} \rightarrow \mathcal{Y}$ and video encoding network $\boldsymbol{f}_v : \mathcal{V} \rightarrow \mathcal{X}$, which generate high-quality features and ultimately enable efficient matching.

3.1. Framework

Fig. 3 sheds light on our end-to-end trainable architecture, which consists of a backbone network and two prototype matching schemes. For a fair comparison with recent methods [48, 51, 20], we adopt CLIP [57] as the backbone network, which exhibits strong performance in downstream tasks. Given an input video $\boldsymbol{v}_i$, we uniformly select L frames as keyframes to extract sequential features $\boldsymbol{X}_i = \{\boldsymbol{x}_C^l, \boldsymbol{x}_{T_1}^l, \cdots, \boldsymbol{x}_{T_K}^l\}_{l=1}^L \in \mathbb{R}^{L\times(K+1)\times D}$, where $\boldsymbol{x}_C^l$ is the global frame token feature ([CLS]) and K is the patch number. For each query text $\boldsymbol{t}_i$, we add the [SOT] and [EOT] tokens to indicate the start and the end of the text. The output text token features can be defined as $\boldsymbol{Y}_i = \{\boldsymbol{y}_S, \boldsymbol{y}_{T_1}, \cdots, \boldsymbol{y}_{T_M}, \boldsymbol{y}_E\} \in \mathbb{R}^{(M+2)\times D}$, where $\boldsymbol{y}_E$ is the global text token feature([EOT]), M and D are the number of words and dimensions, respectively.

Apart from the backbone network, we decompose the matching process into two spatio-temporal complementary parts: 1) Object-Phrase Prototype Matching (Sec. 3.2) aligns the visual object prototypes and text phrase prototypes generated by Spatial Prototype Generation (SPG) to emphasize fine-grained spatial information; 2) Event-Sentence Prototype Matching (Sec. 3.3) exploits event prototypes progressively generated by Temporal Prototype Generation (TPG) to learn dynamic semantic alignment, which explores intrinsic one-to-many video-text relations. The object-phrase and event-sentence prototype matching schemes determine the final cross-modal similarity.

3.2. Object-Phrase Prototype Matching

The text of the video caption subjectively describes a certain event composed of different objects, actions and temporal activities across the video sequence. This results in pairs of video and text that are often partially matched. To solve this problem [9, 59], existing method [43] learns a single patch-event projection to aggregate multiple event prototypes for text-adaptive matching. However, these prototypes are still inadequate to perceive the local details. Our intuition is that decoupling the spatio-temporal modeling process in a divide-and-conquer manner is more beneficial than a single projection. Therefore, we conduct progressive spatio-temporal prototype matching for text-video retrieval. In this section, we illustrate how to model spatial matching. We first perform patch-object and word-phrase spatial prototype aggregation to reveal key local details, and then introduce the fine-grained prototype alignment.

Spatial prototype generation. For videos, we aggregate the patch tokens into object prototypes to represent fine-grained spatial information, such as object instance, object part, and action region. In the prototype generation process, not all patch tokens are aggregated. Although the patch tokens contain important spatial information, they also bring redundancy. For example, some background to-

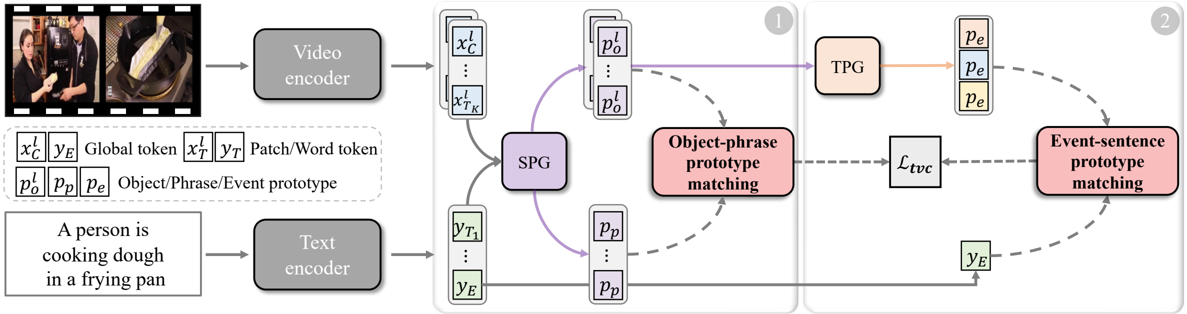
Video encoder
Text encoder
x_C^l y_E Global token
x_T^l y_T Patch/Word token
p_o^l p_p p_e Object/Phrase/Event prototype
A person is cooking dough in a frying pan
SPG
1
2
TPG
Object-phrase prototype matching
Event-sentence prototype matching
$\mathcal{L}_{tvc}$

Figure 3. Our ProST framework. We decompose the matching process into two spatio-temporal complementary parts: 1) Object-Phrase Prototype Matching aligns the visual object prototypes and text phrase prototypes generated by Spatial Prototype Generation (SPG) to emphasize fine-grained spatial information; 2) Event-Sentence Prototype Matching exploits event prototypes progressively generated by Temporal Prototype Generation (TPG) to learn dynamic semantic alignment, which explores intrinsic one-to-many video-text relations.

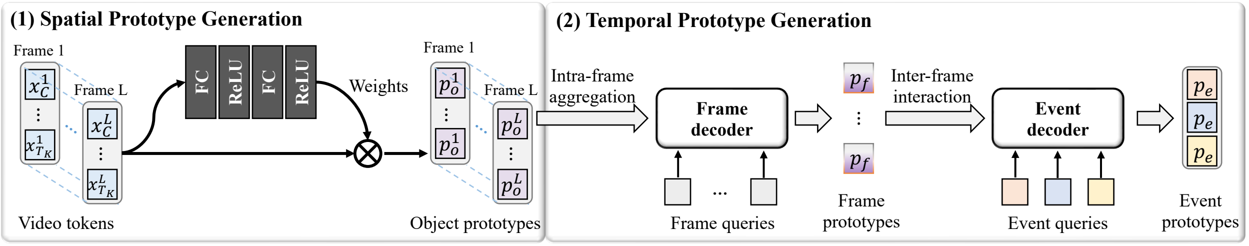

Figure 4. Details of the Spatial Prototype Generation (SPG) and Temporal Prototype Generation (TPG) mechanisms.

kens may interfere with cross-modal alignment. Hence, we hope to filter out retrieval-superfluous information and generate object-level prototypes in a sparse aggregation manner. For simplicity, two Fully Connected (FC) layers and ReLU functions are used to predict sparse weights $\boldsymbol{W}_o^l \in \mathbb{R}^{(K+1)\times N_o}$, where N_o is the number of object prototypes. This way prevents object prototypes from being affected by redundant patches. For each frame $\boldsymbol{X}^l \in \mathbb{R}^{(K+1)\times D}$, SPG can be defined as:

$$\boldsymbol{P}_o^l = \boldsymbol{W}_o^{l^T} \cdot \boldsymbol{X}^l \in \mathbb{R}^{N_o \times D}. \tag{1}$$

Ideally, each object prototype can adaptively aggregate the corresponding object-related or action-related patches.

For text, we draw on the SPG mechanism and design a similar network structure to aggregate word tokens, which generates phrase prototypes $\boldsymbol{P}_p \in \mathbb{R}^{N_p \times D}$. In this way, we explore spatially important information based on fine-grained patch tokens and word tokens. Then, the prototypes are optimized by spatial object-phrase prototype matching.

Object-phrase matching. Different from the existing text-image token-wise interaction [74], we propose a text-video prototype-wise interaction designed from a spatial perspective. Specifically, we first compute the maximum similarity of object-phrase prototypes within each frame. This associates the most similar phrase prototypes to each object prototype, reflecting cross-modal fine-grained matching. Then, for the multi-frame object similarity matrix, we find the largest similarity score across the frame sequences, which gives a more confident probability of object-phrase matching. Finally, the object-phrase matching scores are summed to obtain the final similarity s_{op}. The prototype-wise interaction process is defined as:

$$s_{op} = \frac{1}{N_o} \sum_{j=1}^{N_o} \max_{l=1}^{L} \max_{i=1}^{N_p} [\boldsymbol{P}_p \cdot \boldsymbol{P}_o^{l^T}]_{ij}, \tag{2}$$

where N_o, N_p and L are the number of object prototypes, phrase prototypes and frames, respectively.

3.3. Event-Sentence Prototype Matching

In this section, we illustrate how to model dynamic temporal matching. We first perform progressive object-event prototype aggregation to reveal the video semantic diversity, and then introduce dynamic prototype matching.

Temporal prototype generation. A naive solution to obtain video-level features based on global frame features is by mean pooling [51, 65, 5], or by adding temporal encoder layers [11]. However, this leads to two issues: 1) failure to perceive local details [43, 78, 48] and ignoring important objects will exacerbate the bias of video feature learning; 2) these strategies generate a single video-level feature, which can only quantify one-to-one relations. Therefore, we investigate how to incorporate key fine-grained objects and dynamic temporal changes into diverse event prototypes.

The core idea is to progressively aggregate spatial object prototypes into frame prototypes and then perform inter-frame interaction to generate various event prototypes. In Fig. 4, a frame decoder is first designed to incorporate all object prototypes $\boldsymbol{P}_o \in \mathbb{R}^{(L \times N_o) \times D}$ into frame-level prototypes $\boldsymbol{P}_f \in \mathbb{R}^{L \times D}$, which implies fine-grained inter-object spatial relations. To learn frame-level object relations, we define the masked attention as:

$$\boldsymbol{P}_f = \operatorname{softmax}\left(\boldsymbol{M}_f + \boldsymbol{Q}_f \boldsymbol{K}_o^{\mathrm{T}}\right) \boldsymbol{V}_o + \boldsymbol{Q}_f, \tag{3}$$

where $\boldsymbol{Q}_f \in \mathbb{R}^{L \times D}$ refers to frame queries (*i.e.*, a set of randomly initialized learnable tokens), $\boldsymbol{K}_o$ and $\boldsymbol{V}_o$ are the features after the linear transformation of object prototypes $\boldsymbol{P}_o$. The attention mask $\boldsymbol{M}_f \in \mathbb{R}^{L \times (L \times N_o)}$ is:

$$\boldsymbol{M}_f(i,j) = \begin{cases} 0 & \text{if } i \cdot N_o \leq j < (i+1) \cdot N_o \\ -\infty & \text{otherwise} \end{cases}. \tag{4}$$

Inspired by [9], we add frame prototype $\boldsymbol{p}_f^l$ and original global feature $\boldsymbol{x}_C^l$ of corresponding frames to enhance the robustness of the model:

$$\boldsymbol{p}_f^l = (\boldsymbol{p}_f^l + \boldsymbol{x}_C^l)/2. \tag{5}$$

Next, a dynamic event decoder is developed to learn the inter-frame relations in $\boldsymbol{P}_f$, which can obtain different event prototypes $\boldsymbol{P}_e \in \mathbb{R}^{N_e \times D}$ to illustrate the rich information of videos. Our dynamic attention is formulated as:

$$\boldsymbol{P}_e = \operatorname{softmax}\left(\boldsymbol{Q}_e \boldsymbol{K}_f^{\mathrm{T}}\right) \boldsymbol{V}_f + \boldsymbol{Q}_e, \tag{6}$$

where $\boldsymbol{Q}_e \in \mathbb{R}^{N_e \times D}$ refers to event queries, $\boldsymbol{K}_f$ and $\boldsymbol{V}_f$ are the linear transformation features of frame prototypes $\boldsymbol{P}_f$. During training, each event query learns how to adaptively focus on video frame prototypes, while multiple queries implicitly guarantee a certain event diversity.

Differently, since the same video often corresponds to multiple text semantic descriptions, we directly use the global text representation $\boldsymbol{y}_E$ as a sentence prototype to align with the event prototypes $\boldsymbol{P}_e$.

Event-sentence matching. Similar to [43], this event-sentence prototype matching process can be expressed as:

$$s_{es} = \max_{i=1}^{N_e} \left\langle \boldsymbol{y}_E, \boldsymbol{P}_{e_i} \right\rangle. \tag{7}$$

We look for the closest event prototype to the text representation for dynamic matching.

3.4. Training and Inference

The InfoNCE loss [55, 75] is employed to optimize the prototype matching within a batch. We treat text-video pairs as positive examples, while considering all other pairwise combinations in the batch as negative examples:

$$\mathcal{L}_{tvc} = (\mathcal{L}_{t2v}(\boldsymbol{S}_{op}) + \mathcal{L}_{v2t}(\boldsymbol{S}_{op}) + \mathcal{L}_{t2v}(\boldsymbol{S}_{es}) + \mathcal{L}_{v2t}(\boldsymbol{S}_{es}))/4, \tag{8}$$

$$\mathcal{L}_{t2v}(\boldsymbol{S}) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\boldsymbol{S}^{ii}/\sigma\right)}{\sum_{j=1}^{B} \exp\left(\boldsymbol{S}^{ij}/\sigma\right)}, \tag{9}$$

$$\mathcal{L}_{v2t}(\boldsymbol{S}) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\boldsymbol{S}^{ii}/\sigma\right)}{\sum_{j=1}^{B} \exp\left(\boldsymbol{S}^{ji}/\sigma\right)}, \tag{10}$$

where the σ is a learnable temperature parameter, $\boldsymbol{S}_{op}$ and $\boldsymbol{S}_{es}$ are object-phrase and event-sentence prototype matching similarity matrices in a batch of size B.

During the inference stage, we directly weight the spatio-temporal matching scores for the final similarity matching: $s = s_{es} + \beta s_{op}$, where β is the spatial matching factor.

4. Experiments

We carry out experiments on the standard text-video retrieval datasets of MSRVTT [71], DiDeMo [23], VATEX [64] and LSMDC [58]. Standard metrics in information retrieval [48, 17] are adopted to measure the retrieval performance, including Recall@1/5/10 (R@1/5/10), Median Rank (MdR), and Mean Rank (MnR).

4.1. Experimental Settings

Datasets. (1) **MSRVTT** [71] consists of 10k videos and 200k human-labeled descriptions, where many videos have diverse captions. To thoroughly compare the existing methods, we follow [20] to train on 7k or 9k train+val videos and test on a test set of 1k text-video pairs. (2) **DiDeMo** [23] contains 10K Flickr videos and 40K captions. Following [43], all caption descriptions of a video are concatenated as a query to evaluate all methods. (3) **VATEX** [64] is comprised of $34,991$ video clips. According to [5], we select $25,991$, $1,500$ and $1,500$ videos as the training, validation and test sets. (4) **LSMDC** [58] consists of 118,081 movie clips each paired with a single caption description. We choose $101,079$, $7,408$ and $1,000$ videos as the training, validation and test sets.

Implementation details. We initialize the backbone with the pre-trained model CLIP (ViT-B/32) [57] following [48]. Our decoders consist of two transformer layers with a single attention head. To reduce the computing overhead, all videos are resized to 224×224 with random cropping and flipping, and the frame rate is 3. We set the max text length as 32 and video frame number as 12 except for DiDeMo (set to 64 and 64) as its videos are longer. We optimize the model for 5 epochs on 4 NVIDIA Tesla A100 GPUs, employ Adam optimizer [32] with 0.2 weight decay and fix the batch size as 128. The learning rate is set to 1e-7 for CLIP-initialized weights and 1e-4 for all other parameters. Following [57], we decay the learning rate using a warmup cosine schedule [49]. The spatial matching factor β is 1.5. By default, the prototype number N_o, N_p, and N_e are 12, 28 and 3. The dimension of prototypes is 512.

Table 1. Retrieval results on MSRVTT-9k. We reproduce TVMM* [43] using the same backbone network CLIP (ViT-B/32) [57]. All results in this table do not use additional post-processing techniques [2, 7].

Method	Reference	Text → Video					Video → Text				
		R@1 ↑	R@5 ↑	R@10↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10↑	MdR ↓	MnR ↓
CE [47]	BMVC19	20.6	50.3	64.0	5.3	-	20.9	48.8	62.4	6.0	-
MMT [17]	ECCV20	26.6	57.1	69.6	4.0	24.0	27.0	57.5	69.7	3.7	21.3
T2vlad [65]	CVPR21	29.5	59.0	70.1	4.0	-	31.8	60.0	71.1	3.0	-
HiT [46]	ICCV21	27.7	59.2	72.0	2.9	-	28.8	60.3	72.3	3.0	-
TeachText [10]	ICCV21	29.6	61.6	74.2	3.0	-	32.1	62.7	75.0	3.0	-
ClipBERT [33]	CVPR21	22.0	46.8	59.9	6.0	-	-	-	-	-	-
SupportSet [56]	ICLR21	30.1	58.5	69.3	3.0	-	28.5	58.6	71.6	3.0	-
Frozen [1]	ICCV21	31.0	59.5	70.5	3.0	-	-	-	-	-	-
BridgeFormer [18]	CVPR22	37.6	64.8	75.1	-	-	-	-	-	-	-
TVMM [43]	NeurIPS22	36.2	64.2	75.7	3.0	-	34.8	63.8	73.7	3.0	-
CLIP4Clip [51]	NeurCom22	44.5	71.4	81.6	**2.0**	15.3	42.7	70.9	80.6	**2.0**	11.6
CenterCLIP [78]	SIGIR22	44.2	71.6	82.1	**2.0**	15.1	42.8	71.7	82.2	**2.0**	10.9
X-Pool [20]	CVPR22	46.9	72.8	82.2	**2.0**	14.3	-	-	-	-	-
TVMM* [43]	NeurIPS22	45.8	71.7	81.9	**2.0**	14.8	44.0	71.9	82.3	**2.0**	10.6
TS2-Net [48]	ECCV22	47.0	74.2	83.3	**2.0**	13.6	44.3	73.9	83.0	**2.0**	9.2
ProST	Proposed	**48.2**	**74.6**	**83.4**	**2.0**	**12.4**	**46.3**	**74.2**	**83.2**	**2.0**	**8.7**
ProST (ViT-B/16)	Proposed	**49.5**	**75.0**	**84.0**	**2.0**	**11.7**	**48.0**	**75.9**	**85.2**	**2.0**	**8.3**

Table 2. Text-to-Video retrieval results on MSRVTT-7k.

Method	Text → Video				
	R@1 ↑	R@5 ↑	R@10↑	MdR ↓	MnR ↓
HowTo100M [52]	14.9	40.2	52.8	9.0	-
HERO [36]	16.8	43.4	57.7	-	-
ClipBERT [33]	22.0	46.8	59.9	6.0	-
CLIP4Clip [51]	42.1	71.9	81.4	**2.0**	15.7
X-Pool [20]	43.9	**72.5**	82.3	**2.0**	14.6
TS2-Net [48]	43.1	72.2	82.1	**2.0**	14.2
ProST	**44.5**	72.3	**82.4**	**2.0**	**13.8**
ProST (ViT-B/16)	**46.9**	**73.3**	**82.9**	**2.0**	**13.0**

Table 3. Text-to-Video retrieval results on the DiDeMo dataset.

Method	Text → Video				
	R@1 ↑	R@5 ↑	R@10↑	MdR ↓	MnR ↓
CE [47]	16.1	41.1	82.7	8.3	-
ClipBERT [33]	20.4	48.0	60.8	6.0	-
Frozen [1]	31.0	59.8	72.4	3.0	-
MCQ [18]	37.0	62.2	73.9	3.0	-
TVMM [43]	36.5	64.9	75.4	3.0	-
CLIP4Clip [51]	42.8	68.5	79.2	**2.0**	18.9
TS2-Net [48]	41.8	71.6	82.0	**2.0**	14.8
ProST	**44.9**	**72.7**	**82.7**	**2.0**	**13.7**
ProST (ViT-B/16)	**47.5**	**75.5**	**84.4**	**2.0**	**12.3**

Table 4. Text-to-Video retrieval results on the VATEX dataset.

Method	Text → Video				
	R@1 ↑	R@5 ↑	R@10↑	MdR ↓	MnR ↓
HGR [5]	35.1	73.5	83.5	2.0	-
CLIP [57]	39.7	72.3	82.2	2.0	12.8
CLIP4Clip [51]	55.9	89.2	95.0	**1.0**	3.9
QB-Norm [2]	58.8	88.3	93.8	**1.0**	-
CLIP2Video [15]	57.3	90.0	**95.5**	**1.0**	3.6
TS2-Net [48]	59.1	90.0	95.2	**1.0**	3.5
ProST	**60.6**	**90.5**	95.4	**1.0**	**3.4**
ProST (ViT-B/16)	**64.0**	**92.2**	**96.3**	**1.0**	**3.1**

Table 5. Text-to-Video retrieval results on the LSMDC dataset.

Method	Text → Video				
	R@1 ↑	R@5 ↑	R@10↑	MdR ↓	MnR ↓
Frozen [1]	15.0	30.8	39.8	20.0	-
TVMM [43]	17.8	37.1	45.9	13.5	-
CLIP4Clip [51]	21.6	41.8	49.8	11.0	61.0
X-Pool [20]	24.0	**42.9**	51.5	**9.0**	55.1
QB-Norm [2]	22.4	40.1	49.5	11.0	-
TS2-Net [48]	23.0	42.1	50.4	**9.0**	57.0
ProST	**24.1**	42.5	**51.6**	**9.0**	**54.6**
ProST (ViT-B/16)	**26.3**	**46.1**	**55.2**	**8.0**	**51.4**

4.2. Comparison with State-of-the-art Methods

To evaluate the relative benefits of ProST, we compare its performance with recent works [48, 20, 78, 51] from the literature. The results on the 9k and 7k splits of MSRVTT are shown in Tab. 1 and Tab. 2 respectively. We consistently achieve the best results for text-to-video retrieval and video-to-text retrieval across all splits and all metrics. Specifically, TS2-Net [48] integrates important patches into frame features for frame-level matching, which achieves the second best performance. Notably, for text-to-video retrieval, ProST outperforms TS2-Net [48] with 1.2% and 1.4% R@1 gains on the 9k split (47.0% *vs.* 48.2%) and the 7k split (43.1% *vs.* 44.5%). A 2.0% improvement on R@1 (9k split) is also obtained compared to TS2-Net for video-to-text retrieval (44.3% *vs.* 46.3%). Furthermore, we reproduce TVMM [43] based on CLIP (ViT-B/32), which has 2.4% (45.8% *vs.* 48.2%) and 2.3% (44.0% *vs.* 46.3%) lower R@1 than ProST on the 9k split. The results demonstrate the effectiveness of ProST, which mines the spatial object details and temporal event diversity of videos, improving the model discrimination. Besides, ProST achieves a low MnR metric of 12.4 (text-to-video) on the 9k split, demonstrating the model robustness to erroneous results.

In Tab. 3, 4 and 5, we examine the performance of ProST on the DiDeMo, VATEX and LSMDC datasets respectively. Compared with TS2-Net [48], ProST achieves a substantial boost of 3.1%, 1.5% and 1.1% on the R@1 metric for DiDeMo, VATEX and LSMDC, respectively. ProST outperforms state-of-the-art methods by a large margin with the

Table 6. The ablation study on MSRVTT-9k to investigate the effectiveness of **Object-Phrase Prototype Matching (OPPM)** and **Event-Sentence Prototype Matching (ESPM)**. MdR results are not shown, as MdR=2.0 for all variants. **Definition of important symbols**: ✓ indicates default settings; ↻TPG or ↻SPG means replace with TPG or SPG to generate prototypes; P-P or O-W refers the alignment of patch-phrase or object-word; -F, -M or -R refers to removing frame decoder, attention mask or residual structure in Eq. 5; F-W or F̄-S means directly aligning frame global tokens and word tokens or aligning the mean of frame global tokens and text global token.

ID	OPPM		ESPM		Text → Video				Video → Text			
	SPG	Matching	TPG	Matching	R@1 ↑	R@5 ↑	R@10↑	MnR	R@1 ↑	R@5 ↑	R@10↑	MnR
1			✓	✓	45.6	72.2	81.1	14.3	44.9	71.6	82.3	10.0
2			↻SPG	✓	45.8	71.7	81.9	14.8	44.0	71.9	82.3	10.6
3	✓	✓			45.9	72.8	82.4	13.2	45.1	73.3	82.7	9.8
4	↻TPG	✓	✓	✓	46.3	73.3	81.8	13.5	45.0	73.1	82.4	9.8
5	✓	✓	↻SPG	✓	47.7	73.6	83.0	12.6	45.9	74.0	83.0	9.0
6	✓	P-P	✓	✓	46.0	72.5	81.6	14.9	43.1	72.4	82.1	10.8
7	✓	O-W	✓	✓	47.1	73.3	82.2	12.8	45.7	73.3	82.6	9.6
8	✓	✓	-F	✓	46.4	72.8	81.7	13.5	44.5	72.5	82.0	10.2
9	✓	✓	-M	✓	47.8	73.8	82.9	12.6	46.0	73.7	82.8	9.0
10	✓	✓	-R	✓	46.6	72.9	82.4	14.0	45.3	72.1	81.8	10.1
11		F-W	✓	✓	47.3	73.5	82.8	12.6	45.9	73.7	83.1	9.0
12	✓	✓		F̄-S	46.5	72.6	82.3	13.4	45.1	73.4	82.5	9.8
Ours	✓	✓	✓	✓	**48.2**	**74.6**	**83.4**	**12.4**	**46.3**	**74.2**	**83.2**	**8.7**

Table 7. The ablation study on MSRVTT-9k to investigate the configuration of the number of prototypes.

$\{N_o, N_p, N_e\}$	Text → Video				
	R@1 ↑	R@5 ↑	R@10↑	MdR ↓	MnR ↓
{6, 14, 3}	45.6	73.2	81.6	**2.0**	14.3
{6, 28, 3}	46.7	73.4	82.2	**2.0**	13.4
{12, 14, 3}	47.0	73.7	83.0	**2.0**	12.9
{24, 32, 3}	46.0	73.3	82.0	**2.0**	13.8
{12, 28, 1}	46.3	73.0	82.1	**2.0**	13.5
{12, 28, 2}	47.4	74.0	83.1	**2.0**	12.7
{12, 28, 4}	47.6	74.3	83.2	**2.0**	12.5
{12, 28, 3}	**48.2**	**74.6**	**83.4**	**2.0**	**12.4**

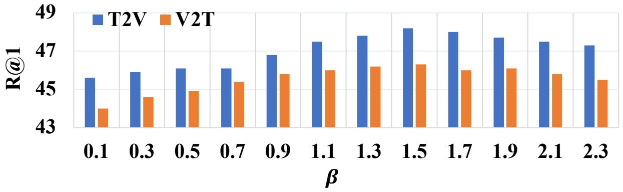

Figure 5. The R@1 results at different spatial matching factors β.

CLIP (ViT-B/16) backbone. These results suggest the versatility and generalization of ProST, which can effectively process videos from different domains (*e.g.*, movies). We owe the advantage of ProST over these works to the full use of dynamic spatio-temporal prototype matching.

4.3. Ablation Study

Analysis of progressive spatio-temporal prototype matching. In Tab. 6, we compare ProST with 12 variants, which proves that each module contributes to final results.

We elaborate on some important observations. First, we show the results relying only on **Object-Phrase Prototype Matching (OPPM)** or **Event-Sentence Prototype Matching (ESPM)**. To implement our model without OPPM, all tokens are fed directly into the frame decoder and event decoder (TPG) to get multiple event prototypes. When only ESPM is used, there is a 2.6% decrease in R@1. This confirms the complementary role of fine-grained spatial details for ESPM. Next, we replace TPG with SPG and find that there is a certain suppression effect on noise. However, both of the above variants perform poorly due to the lack of progressive prototype generation and matching. When we only train model with OPPM, the model still underperforms. This may be due to the lack of temporal understanding and the inability to resolve correspondence ambiguity.

Second, ↻TPG, P-P and O-W prove that there is indeed redundancy in the original video tokens. The sparse weights in SPG can alleviate this problem. However, FC layers in SPG cannot interact with intra-frame and inter-frame information, making it difficult to generate better event prototypes. This may be the reason for the drop in ↻SPG. Third, P-P and O-W also confirm the benefits of prototype matching. Raw patches or words without aggregation strategies may worsen the cross-modal alignment.

Fourth, the results for -F, -M and -R also decreased. Because intra-frame local object relations and global frame features jointly complement comprehensive frame-level spatial information, which is helpful for subsequent temporal learning. Fifth, from F-W and F̄-S, we can see the importance of prototype design mode. Directly using global frame features ([CLS] tokens) in ProST-S for frame-word alignments lacks object details. In F̄-S, the mean value of frame global features mixes all information equally, making it difficult to characterize dynamic and changeable videos.

Analysis of prototype configuration. We conduct experiments on various prototype configurations in Tab. 7. Im-

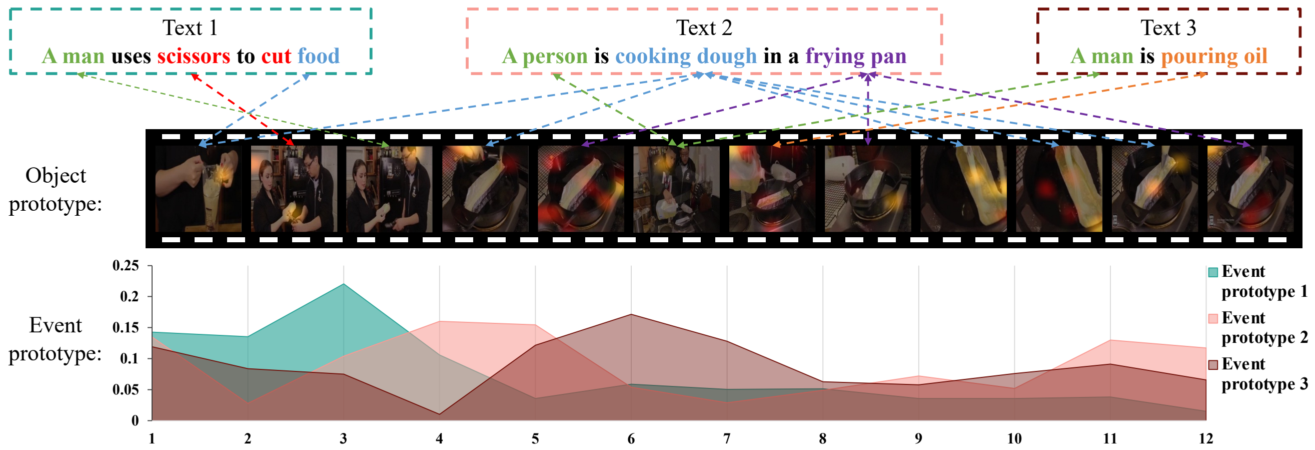

Figure 6. Visualization of the object and event prototypes. We sample 12 frames in the video and object prototypes are shown as highlighted response regions in the frame. Then, we show cross-attention event weights of Eq. 6 in a line graph. Best viewed in color.

proper configuration of spatial prototypes leads to performance degradation, possibly because too many prototypes may introduce local noise, while few prototypes are not enough to express semantics. Furthermore, the best result is obtained when the number of event prototypes is 3, which may be related to the degree of semantic divergence of texts.

Hyperparameter analysis. As shown in Fig. 5, we investigate the spatial matching factor β and present results for text-video (T2V) and video-text retrieval (V2T). As β grows, the performance increases at first and reaches the best results, then decreases as a whole. A small β may not fully exploit the underlying fine-grained spatial information, while a large β may overemphasize local matching and destroy the overall dynamic semantic understanding.

4.4. Qualitative Results

Visualization of prototypes. In Fig. 6, we first present object prototypes in video frames. Many redundant patches are removed by prototypes, and the spatially highlighted object part and the phrase can be well locally corresponded. Then, we compute cross-attention for 3 event queries and 12 video frames according to Eq. 6, whose weights are shown in a line graph. The weights of different event prototypes on different frames are quite different, which proves that our model can learn temporal relations and event prototypes can express diverse semantics. For example, event prototype 1 has larger weights on the first three frames, which indicates that it may correspond to the semantics contained in text 1. Similarly, event prototype 2 or 3 corresponds to text 2 or 3.

Retrieval results. Top-3 TVR results are illustrated in Fig. 7. Specifically, for query 97, although the top-2 recalled videos are women in the same dress, the difference is that the correct video discusses the cushion seat, while the incorrect video does not. This proves that object-phrase prototype matching provides fine-grained spatial knowledge (*i.e.*, objects). For query 353, ProST pays attention to both the brief human speech at the beginning of the video and the subsequent earth display. However, using only event prototypes will result in a matching error. This shows that ProST can make a trade-off between the spatio-temporal matching relations to achieve better discrimination.

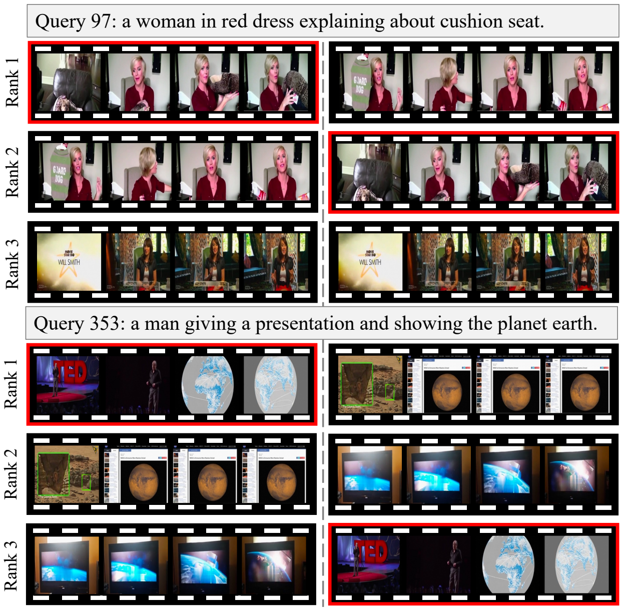
Query 97: a woman in red dress explaining about cushion seat.
Rank 1
Rank 2
Rank 3
Query 353: a man giving a presentation and showing the planet earth.
Rank 1
Rank 2
Rank 3

Figure 7. Top-3 TVR results on MSRVTT-9k. Left: the ranked videos by our ProST. Right: the ranked videos with only event-sentence prototype matching. Red: correct videos.

5. Conclusion

We have proposed a novel text-video retrieval framework, ProST, to decompose the matching process into complementary object-phrase and event-sentence prototype alignments. In the object-phrase prototype matching stage, we design the spatial prototype generation mechanism to focus on important video content and strengthen the fine-grained spatial alignment. In the event-sentence prototype matching phase, we use the temporal prototype generation mechanism to progressively generate diverse event prototypes and learn dynamic one-to-many relations. In this paper, we hope not only to provide insights into the importance of complementary spatio-temporal matching but also to facilitate future work that advances text-video retrieval by solving design flaws rather than mostly trial and error.

Acknowledgment. This work is supported by the National Key Research and Development Program of China (2022YFB3104700), the National Nature Science Foundation of China (62121002, 62022076, U1936210, 62102384).

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2, 3, 6

[2] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *CVPR*, 2022. 6

[3] Soravit Changpinyo, Jordi Pont-Tuset, Vittorio Ferrari, and Radu Soricut. Telling the what while pointing to the where: Multimodal queries for image retrieval. In *ICCV*, 2021. 1

[4] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. VALOR: vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 1

[5] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, 2020. 3, 4, 5, 6

[6] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. Vista: vision and scene text aggregation for cross-modal retrieval. In *CVPR*, 2022. 1

[7] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 6

[8] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019. 2

[9] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021. 3, 5

[10] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *ICCV*, 2021. 2, 6

[11] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *TPAMI*, 2022. 2, 4

[12] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. SVTR: scene text recognition with a single visual model. In *IJCAI*, pages 884–890, 2022. 1

[13] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Cross-modal deep variational hashing. In *ICCV*, 2017. 1

[14] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improved visual-semantic embeddings. In *BMVC*, 2018. 1

[15] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 3, 6

[16] Sheng Fang, Shuhui Wang, Junbao Zhuo, Xinzhe Han, and Qingming Huang. Learning linguistic association towards efficient text-video retrieval. In *ECCV*, 2022. 1

[17] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 2, 5, 6

[18] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *CVPR*, 2022. 3, 6

[19] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. In *NeurIPS*, 2020. 3

[20] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 2022. 1, 3, 5, 6

[21] Ning Han, Jingjing Chen, Chuhao Shi, Yawen Zeng, Guangyi Xiao, and Hao Chen. Bic-net: Learning efficient spatio-temporal relation for text-video retrieval. *arXiv preprint arXiv:2110.15609*, 2021. 1

[22] Laura Hanu, James Thewlis, Yuki M Asano, and Christian Rupprecht. Vtc: Improving video-text retrieval with user comments. In *ECCV*, 2022. 1

[23] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 5

[24] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. Lightweight attentional feature fusion: A new baseline for text-to-video retrieval. In *ECCV*, 2022. 1

[25] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *ICCV*, 2021. 1

[26] Kaixiang Ji, Jiajia Liu, Weixiang Hong, Liheng Zhong, Jian Wang, Jingdong Chen, and Wei Chu. Cret: Cross-modal retrieval transformer for efficient text-video retrieval. In *SIGIR*, 2022. 2

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 1

[28] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. In *NeurIPS*, 2022. 2, 3

[29] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *CVPR*, 2023. 2

[30] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video retrieval with disentangled conceptualization and set-to-set alignment. *arXiv preprint arXiv:2305.12218*, 2023. 2

[31] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. *arXiv preprint arXiv:2303.09867*, 2023. 2

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[33] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 3, 6

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1

[35] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*, 2023. 1

[36] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*, 2020. 6

[37] Pandeng Li, Yan Li, Hongtao Xie, and Lei Zhang. Neighborhood-adaptive structure augmented metric learning. In *AAAI*, 2022. 1

[38] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. *arXiv preprint arXiv:2307.02869*, 2023. 1

[39] Pandeng Li, Hongtao Xie, Jiannan Ge, Lei Zhang, Shaobo Min, and Yongdong Zhang. Dual-stream knowledge-preserving hashing for unsupervised video retrieval. In *ECCV*, 2022. 2

[40] Pandeng Li, Hongtao Xie, Shaobo Min, Jiannan Ge, Xun Chen, and Yongdong Zhang. Deep fourier ranking quantization for semi-supervised image retrieval. *TIP*, 2022. 1

[41] Pandeng Li, Hongtao Xie, Shaobo Min, Zheng-Jun Zha, and Yongdong Zhang. Online residual quantization via streaming data correlation preserving. *TMM*, 2022. 1

[42] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. 1

[43] Chengzhi Lin, Ancong Wu, Junwei Liang, Jun Zhang, Wenhang Ge, Wei-Shi Zheng, and Chunhua Shen. Text-adaptive multiple visual prototype matching for video-text retrieval. In *NeurIPS*, 2022. 2, 3, 4, 5, 6

[44] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *ECCV*, 2022. 1

[45] Hongying Liu, Ruyi Luo, Fanhua Shang, Mantang Niu, and Yuanyuan Liu. Progressive semantic matching for video-text retrieval. In *ACM MM*, 2021. 1

[46] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *ICCV*, 2021. 2, 3, 6

[47] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 1, 2, 6

[48] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, 2022. 2, 3, 4, 5, 6

[49] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[50] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1

[51] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 2022. 1, 2, 3, 4, 6

[52] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 6

[53] Shaobo Min, Hongtao Xie, Hantao Yao, Xuran Deng, Zheng-Jun Zha, and Yongdong Zhang. Hierarchical granularity transfer learning. In *NeurIPS*, 2020. 1

[54] Shaobo Min, Hantao Yao, Hongtao Xie, Zheng-Jun Zha, and Yongdong Zhang. Multi-objective matrix normalization for fine-grained visual recognition. *TIP*, 2020. 1

[55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[56] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 6

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 5, 6

[58] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Joseph Pal, Hugo Larochelle, Aaron C. Courville, and Bernt Schiele. Movie description. *IJCV*, 2017. 5

[59] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, pages 1979–1988, 2019. 3

[60] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *ICCV*, 2019. 1

[61] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 1

[62] Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. In *ICCV*, 2021. 1

[63] Alex Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *CVPR*, 2022. 2

[64] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 5

[65] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *CVPR*, 2021. 1, 3, 4, 6

[66] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1

[67] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, 2019. 1

[68] Zeyu Wang, Yu Wu, Karthik Narasimhan, and Olga Russakovsky. Multi-query video retrieval. In *ECCV*, 2022. 1

[69] Michael Wray, Gabriela Csurka, Diane Larlus, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*, 2019. 2

[70] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *CVPR*, 2021. 1

[71] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016. 5

[72] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. In *ICLR*, 2023. 1

[73] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *ICCV*, 2021. 1

[74] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *ICLR*, 2022. 4

[75] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. In *BMVC*, 2019. 5

[76] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*, 2023. 1

[77] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, 2018. 2

[78] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *SIGIR*, 2022. 2, 3, 4, 6

[79] Tianlun Zheng, Zhineng Chen, Jinfeng Bai, Hongtao Xie, and Yu-Gang Jiang. Tps++: Attention-enhanced thin-plate spline for scene text recognition. *IJCAI*, 2023. 1