PIDRo: Parallel Isomeric Attention with Dynamic Routing for Text-Video Retrieval

Peiyan Guan[1]*, Renjing Pei[2]†, Bin Shao[2], Jianzhuang Liu[2],
Weimian Li[2], Jiaxi Gu[2], Hang Xu[2], Songcen Xu[2], Youliang Yan[2], Edmund Y. Lam[1]†
The University of Hong Kong[1], Huawei Noah's Ark Lab[2]
{pyguan, elam}@eee.hku.hk, {peirenjing, shaobin3, liu.jianzhuang, liweimian,
xusongcen,yanyouliang}@huawei.com, {imjiaxi, chromexbjxh}@gmail.com

Abstract

Text-video retrieval is a fundamental task with high practical value in multi-modal research. Inspired by the great success of pre-trained image-text models with large-scale data, such as CLIP, many methods are proposed to transfer the strong representation learning capability of CLIP to text-video retrieval. However, due to the modality difference between videos and images, how to effectively adapt CLIP to the video domain is still underexplored. In this paper, we investigate this problem from two aspects. First, we enhance the transferred image encoder of CLIP for fine-grained video understanding in a seamless fashion. Second, we conduct fine-grained contrast between videos and texts from both model improvement and loss design. Particularly, we propose a fine-grained contrastive model equipped with parallel isomeric attention and dynamic routing, namely PIDRo, for text-video retrieval. The parallel isomeric attention module is used as the video encoder, which consists of two parallel branches modeling the spatial-temporal information of videos from both patch and frame levels. The dynamic routing module is constructed to enhance the text encoder of CLIP, generating informative word representations by distributing the fine-grained information to the related word tokens within a sentence. Such model design provides us with informative patch, frame and word representations. We then conduct token-wise interaction upon them. With the enhanced encoders and the token-wise loss, we are able to achieve finer-grained text-video alignment and more accurate retrieval. PIDRo obtains state-of-the-art performance over various text-video retrieval benchmarks, including MSR-VTT, MSVD, LSMDC, DiDeMo and ActivityNet.

1. Introduction

The amount of videos on the Internet has significantly increased recently. Efficiently finding target videos based on text description, referred to as text-video retrieval, is of high practical and research value. Over the past few years, various methods have been proposed for this task [8, 14, 25, 26, 11, 12].

Recently, large-scale contrastive text-image pre-training has achieved great success in many multi-modal text-vision understanding tasks [34, 18, 17, 35]. One representative method is CLIP, which trains a text encoder and an image encoder with over 400 million image-text pairs. Inspired by such success, some works directly adapt the pre-trained text and image encoders to the video domain and achieve great improvements [29, 15, 16, 13]. However, simply using the models without considering the differences between images and videos neglects the characteristics of videos. In this work, we study CLIP-based text-video retrieval. In spite of CLIP's remarkable performance for image classification, two major issues remain when using it in the video domain. The first is how to enhance the transferred image encoder for video understanding in a seamless fashion. The second lies in how to conduct finer-grained contrast between text and video.

For the first issue, the image encoder (a transformer) of CLIP conducts spatial attention within each frame and does not explore the cross-frame temporal relationship. Current methods mainly append a temporal transformer to the image encoder to learn the temporal information [15, 30]. However, it only conducts temporal attention in the frame-to-frame fashion and lacks fine-grained temporal modeling. Building a powerful video encoder is an important topic for video understanding. FROZEN employs divided space-time attention to learn the spatial-temporal information [4]. TS2-Net incorporates a token-shift module to enable patch-level cross-frame interaction [27]. However, these methods change the internal structure of the image encoder of text-

*This work was done during an internship at Huawei.
†Corresponding authors: Renjing Pei, Edmund Y. Lam

image pre-training models and may corrupt the transferred knowledge.

As for the second issue, fine-grained interaction is effective for better modality alignment. However, CLIP only conducts coarse-grained contrast between text and image with global features. It lacks the capability of capturing finer-level information. To solve this problem, some methods conduct fine-grained interaction over token representations [41, 20]. However, the performance improvement is limited when the CLIP-based models utilize such fine-grained contrastive loss for text-video retrieval [30]. This is because these models are dominated by the encoders of CLIP, which does not provide informative enough token representations, such as words and patches, for us to conduct effective fine-grained cross-modal interaction. Designing a good loss function alone is not enough, and correspondingly enhancing the encoders is also necessary. However, systematic fine-grained interaction from both model enhancement and loss design is rarely explored.

Based on the above analysis, we propose PIDRo, a CLIP-based model equipped with a novel parallel isomeric attention module and dynamic routing, for fine-grained text-video retrieval. Specifically, we design a new architecture with two parallel branches for comprehensive video modeling. One branch learns frame representations with spatial attention first and temporal attention second. The other one encodes the video in reverse order to acquire patch representations. Each of the two branches consists of a spatial transformer and a temporal transformer, which are arranged in different orders in the two branches to have different functions, working like isomers. Besides, we propose a dynamic routing module appended to the text encoder to enhance the word representations. Concretely, it is designed to dig out fine-grained information embedded in the global feature of the sentence and distribute it to the corresponding word tokens. These newly designed multi-modal encoders provide us with informative representations of words, patches and frames, which allows conducting effective fine-grained video-text contrast. Meanwhile, we do not change the internal architectures of CLIP's encoders, keeping CLIP's extendability during transfer. We then design a contrastive loss to conduct fine-grained video-text contrast on the learned representations (i.e., word-frame and word-patch), which calculates token-wise similarity scores between a text and a video.

By addressing the above two issues, we are able to smoothly transfer the text-image pre-training model, CLIP, into text-video retrieval. We conduct comprehensive experiments on several text-video retrieval benchmarks. Our PIDRo achieves state-of-the-art performance on all benchmarks and sets new record of retrieval accuracy. The main contributions of PIDRo are summarized as follows:

1. We propose a parallel isomeric attention module for better video understanding. It models the temporal dependencies of videos in both frame and patch levels and does not undermine the original structure of the text-image pre-trained model.

2. We design a dynamic routing module to yield informative representations for word tokens. It distributes the fine-grained information related to different words but buried in the global feature to the corresponding words.

3. Our work leads to a new scheme of conducting effective fine-grained cross-modal interaction for CLIP-based methods via both model enhancement and loss design.

4. Extensive experiments on five widely used text-video retrieval benchmarks demonstrates the superiority of our method. Ablation studies also illustrate the effectiveness of our video and text encoders and fine-grained contrast.

2. Related Work

2.1. Text-Video Retrieval

Large-scale text-image pre-training models, such as CLIP [34], FILIP [41] and ALIGN [19], have demonstrated success across various downstream tasks. Recently, there are also some end-to-end trainable models like Frozen [4] and HiVLP [37] being proposed, which are designed to take advantage of both large-scale image and video captioning datasets. However, the collection and cleaning of the pre-training video data cost huge manpower for video pre-training. Researchers find that text-video retrieval models extended from the pre-trained model CLIP can also achieve state-of-the-art simply by transferring CLIP to video-related tasks. Therefore, a series of CLIP-based methods, such as CLIP4Clip [29], CLIP2Video [13], X-CLIP [30], TS2-Net [27], and CLIPPING [32] keep appearing. In order to maintain the pre-training knowledge of CLIP, those models are usually constructed on top of CLIP's encoders with the similarity calculation to obtain the results of video-text retrieval. For example, CLIP4Clip computes sequential similarity and X-CLIP calculates word-frame similarity.

2.2. Fine-Grained Understanding

The core of text-video retrieval models lies in modeling the semantic information of the two modalities and the interaction between them. Usually, more details and key information (e.g., description of a small object or insignificant movement) can improve the final retrieval results. There are mainly two ways for fine-grained improvement: model architecture [5, 3, 28] or loss design [16, 30, 38, 41].

For fine-grained model architectures, existing methods such as Timesformer [5] and VIVIT [3] forward all spatial-temporal patches extracted from videos through a transformer encoder. Timesformer finds that divided attention architecture is a good design. It separately applies temporal attention and spatial attention within each block of the net-

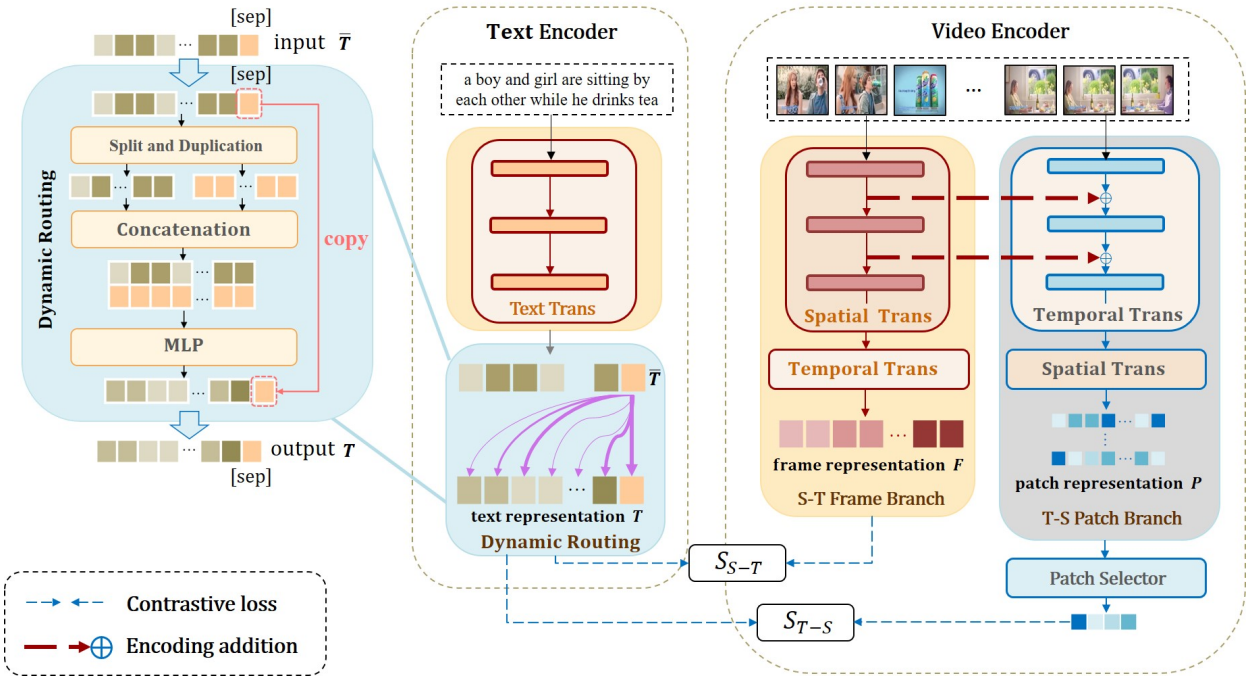

Figure 1. Diagram of PIDRo. It consists of a text encoder and a video encoder. The text encoder first employs CLIP's text transformer to learn a representation for each word token, and then has a dynamic routing module to enhance them by conducting information redistribution within the sentence. The video encoder contains two branches. The S-T frame branch learns frame representations by conducting spatial attention first and temporal attention across the frames second, while the T-S patch branch performs the two attentions in reverse order and learns patch cube representations. Fine-grained interaction are performed upon these representations, including a token-wise interaction between text and frame representations S_{S-T} and another interaction between text and patch representations S_{T-S}.

work. However, when the divided attention architecture is adapted to a CLIP-based model, we find that it either provides no performance gain with a small learning rate or increases the loss during training with a large learning rate.

For fine-grained loss design, FILIP [41] uses a token-wise similarity between visual and textual tokens to guide the contrastive objective in the pre-training. It successfully leverages the fine-grained expressiveness between image patches and words by word-patch alignment. Motivated by FILIP, X-CLIP [30] introduces token-wise interaction to CLIP-based text-video retrieval with word-frame alignment. However, since CLIP is a text-image pre-training framework, most textual information converges to the last token ([sep] token [34]). Therefore, only modifying the contrastive loss may not be enough for CLIP-based text-video retrieval. Besides, the word-frame alignment still lacks fine-grained information compared with the word-patch alignment. In this paper, we focus on improving both the model architecture and loss design to leverage fine-grained effectiveness, and simultaneously transfer the pre-training knowledge of CLIP to the video domain.

3. Methodology

3.1. Overview

In text-video retrieval, we aim to learn a function which calculates the similarity between text descriptions and video clips. This similarity function is expected to give higher scores to those related video-text pairs and lower scores to those unrelated. Our PIDRo is a fine-grained contrast model equipped with parallel isomeric attention and dynamic routing. The overall framework is depicted in Fig. 1, which is built upon the pre-trained CLIP model. The parallel isomeric attention module serves as the video encoder with two branches learning frame and patch representations, respectively. Our text encoder is constructed by appending the dynamic routing module to CLIP's text encoder (a transformer) to acquire informative representations of word tokens. Such model design allows us to fully take advantage of the text-image pre-trained model without changing its internal structure when transferring it to the video domain. The encoders provide informative text and video representations in sequence, from which we can conduct cross-modal token-wise interaction to achieve fine-grained contrast.

3.2. Video Encoder

Given a video clip $V \in \mathbb{R}^{N_f \times H \times W \times 3}$ of N_f sampled frames with spatial size of $H \times W$, each frame of it is divided into N_p non-overlapping patches. Those patches are fed into the video encoder, i.e., the parallel isomeric attention module with two parallel branches. The first one consists of two cascaded transformers: a spatial and a temporal. It takes raw frames as the input and generates spatial frame representations. We call it the S-T frame branch. The second one has a temporal transformer and a spatial transformer. Its input is the patch cubes, each of which contains patches at the same location of all frames, and it acquires temporal patch representations. We refer to it as the T-S patch branch. These two branches both consist of spatial and temporal transformers, which, however, are in different orders for different purposes, just like isomers.

3.2.1 Parallel Isomeric Attention

The S-T frame branch first employs a spatial transformer to encode each frame, which is CLIP's image encoder. A [cls] token is prepended to the sequence of the patch tokens of each frame. The output corresponding to this token is used as the representation of this frame. The spatial transformer does not consider temporal dependencies of the video. Thus, a temporal transformer is appended to the image encoder for cross-frame temporal correlation. It consists of 4 layers, each of which has 8 heads. By feeding the representation of each frame to it, we are able to acquire the final sequential frame representations $F = [f_1, f_2, \cdots, f_{N_f}]^\top \in \mathbb{R}^{N_f \times d}$, where d is the dimensionality of the token features.

The S-T frame branch is a generic architecture for CLIP-based models. However, its exploration of temporal correlation is not enough. To further enhance the understanding of the video, we build another T-S patch branch, which perceives the video from another view, patch-level temporal dynamics. Specifically, this branch first employs a temporal transformer to perform the attention within each patch cube. It has the same structure as the spatial transformer in the first branch. To better leverage the pre-trained knowledge, we also conduct interaction between these two transformers. Except for the last layer, we add the encoding of each layer in the spatial transformer to the encoding of that same layer in the temporal transformer (see the supplementary materials for the details). Such interaction transfers CLIP's spatial attention knowledge to the temporal transformer. We also prepend a [cls] token to the token sequence of each cube, whose corresponding output serves as the patch cube representation. The representations capture the patch-level cross-frame relationship. To further enhance the spatial correlation among the cubes, similar to the S-T frame branch, we append a spatial transformer to the the temporal one. It shares the same structure as the temporal transformer in the S-T frame branch. In this way, we obtain the patch representations $P = [p_1, p_2, \cdots, p_{N_p}]^\top \in \mathbb{R}^{N_p \times d}$. Compared to the S-T frame branch, T-S has more fine-grained temporal modeling but coarser spatial modeling. These two branches complement each other and provide a comprehensive understanding of videos.

3.2.2 Patch Token Selector

Although the patch representations give more fine details, the number N_p of patches is much larger than the number N_f of frames. To reduce the computation cost of subsequent similarity calculation, we use a patch selector to select informative patches that contain salient semantics. This is based on a key observation that, among all the patches, only a few are related to the text. Specifically, we apply an MLP followed by a Softmax function to compute the importance score for each patch, which is formulated as:

$$S = \text{Softmax}(\text{MLP}(P)) \in \mathbb{R}^{N_p}. \tag{1}$$

Based on the K most importance scores, we generate $I \in \{0, 1\}^{K \times N_p}$, where each row of I has only one "1" that corresponds to one of the K importance score, and each column of I has at most one "1". Using I, we can select K patch representations from P:

$$P_s = IP \in \mathbb{R}^{K \times d}. \tag{2}$$

To make such top-K selection differentiable, we adopt the perturbed maximum method proposed in [6]. These K patch representations are used for the subsequent similarity calculation.

3.3. Text Encoder with Dynamic Routing

The text encoder of CLIP is employed for the textual representation learning. It is a transformer with 12 layers and 8 heads with feature dimensionality of 512. Given a text input with N_t word tokens, a [cls] token and a [sep] token are added to the beginning and end of the word tokens, respectively. This text encoder generates a representation for each token. Thus, for each sentence, we have a sequence of text representations $\overline{T} = [\bar{t}_0, \bar{t}_1, \cdots, \bar{t}_{N_t+1}]^\top \in \mathbb{R}^{(N_t+2) \times d}$. In CLIP's training, since only $\bar{t}_{N_t+1}$ is used to represent the whole sentence, this global representation cannot capture the fine-grained details of the sentence. To deal with this problem, we expand it with a dynamic routing module as shown in Fig. 1. This module can boost the word representations by achieving the knowledge redistribution within $\overline{T}$. It concatenates the representation of the [sep] token, which contains rich global information, with the representation of each word token, which has limited local information. The concatenated representations are then fed into an MLP to

dig out the deeply buried information related to each word from the global feature and integrate it to the corresponding token, by which the word representations are enhanced. This process for the i-th text token is formulated as:

$$t_i = \text{MLP}([\bar{t}_i, \bar{t}_{sep}]) \in \mathbb{R}^d, \quad 0 \le i \le N_t, \tag{3}$$

where $[\cdot]$ denotes the concatenation operation. Besides, the representation of the last token stays the same, i.e., $t_{N_t+1} = \bar{t}_{N_t+1}$. Finally, we acquire the textual representations $T = [t_0, t_1, \cdots, t_{N_t+1}]^\top$.

3.4. Similarity Calculation

Our text and video encoders provide sequential representations of fine granularity, which allows conducting fine-grained contrastive learning for text-video retrieval. We perform token-wise similarity calculation between the text and video modalities. The video encoder has two kinds of representations from the two branches, which are interacted with the text representations separately.

Take the contrast between text and frames as an example. Given two sequences of the textual representations T and the frame representations F, we calculate the similarity between them as:

$$\begin{aligned} S_{S-T} = \frac{1}{2}\Bigg(&\sum_{i=0}^{N_t+1} \alpha_i \max_{1\le j\le N_f} \{t_i^\top f_j\} \\ &+ \sum_{j=1}^{N_f} \beta_j \max_{0\le i\le N_t+1} \{t_i^\top f_j\} \Bigg), \end{aligned} \tag{4}$$

where we use SoftMax to obtain the normalized weights $\alpha_i = \frac{\exp(\eta \cdot \max_{1\le j\le N_f}\{t_i^\top f_j\})}{\sum_{k=0}^{N_t+1}\exp(\eta \cdot \max_{1\le j\le N_f}\{t_k^\top f_j\})}$ for the i-th text representation and $\beta_j = \frac{\exp(\eta \cdot \max_{0\le i\le N_t+1}\{t_i^\top f_j\})}{\sum_{k=1}^{N_f}\exp(\eta \cdot \max_{0\le i\le N_t+1}\{t_i^\top f_k\})}$ for the j-th frame representation, and η is the temperature parameter, which is set to 100 empirically in our experiments. We compute the similarity between the text and patch cube representations S_{T-S} in a similar way. The final similarity score between the text and the video is defined as:

$$S = S_{S-T} + \lambda S_{T-S}, \tag{5}$$

where λ is a balancing weight.

Given a training batch of B text-video pairs, we use symmetric InfoNCE as the training objective, which is represented:

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\tau \cdot S(T_i, V_i))}{\sum_{j=1}^{B}\exp(\tau \cdot S(T_i, V_j))}, \tag{6}$$

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\tau \cdot S(T_i, V_i))}{\sum_{j=1}^{B}\exp(\tau \cdot S(T_j, V_i))}, \tag{7}$$

$$\mathcal{L} = (\mathcal{L}_{t2v} + \mathcal{L}_{v2t})/2, \tag{8}$$

where $S(T_i, V_j)$ denotes the similarity between the i-th sentence and j-th video in the batch, and τ is another temperature parameter empirically set to 100 in all the experiments.

4. Experiments

4.1. Datasets

MSR-VTT [40] consists of 10,000 video clips with 20 captions for each clip. We follow the 'Training-9K' split, where 9,000 videos with the corresponding captions are for training and the left 1,000 pairs for testing.

MSVD [7] contains 1,970 videos, each of which is annotated with 40 captions. We use the split of 1200, 100, and 670 videos for training, validation and testing, respectively.

LSMDC [36] consists of 118,081 short videos and captions. We adopt 109673, 7408 and 1000 for training, validation and testing, respectively.

DiDeMo [2] dataset is split into training, validation and testing sets containing 8,395, 1,065 and 1,004 videos, respectively. Following previous works [26, 23, 5], the multiple descriptions of each video are concatenated.

ActivityNet [22] contains 20,000 videos collected from YouTube. We use the same split as in [14, 29, 39], where all captions of each video are concatenated.

4.2. Experimental settings

Implementation Details. We train the whole model in three steps. First, the text transformer and the S-T frame branch are trained for 5 epochs without the dynamic routing module and the T-S patch branch. Second, we freeze the two parts trained in the first step and only optimize the dynamic routing module for 3 epochs without the T-S patch branch. In the last step, we fix the parameters of the modules optimized in the first two steps and only train the T-S patch branch and the patch selector for another 2 epochs. The learning rate for the modules of CLIP is set to $1e{-}7$, while the learning rate for all the other modules is set to $1e{-}4$. We adopt the cosine learning rate schedule with a linear warm-up in each training step. The spatial transformer in the S-T frame branch is ViT-B/32 if not specified. The model is trained by Adam optimizer [21] with a batch size of 128. The max text and frame lengths are set to 32 and 12 for MSR-VTT, MSVD and LSMDC, and to 64 and 64 for DiDeMo and ActivityNet, respectively.

Evaluation Metrics. The metrics Recall at rank K (R@K), mean rank (MnR) and median rank (MdR) are used for evaluation. Similar to previous works, we set $K = 1, 5$ and 10 in the experiments.

4.3. Ablation Study

Effect of Dynamic Routing (DR) and T-S Patch Branch (T-S). In Table 1, we provide the ablation study of the dynamic routing and T-S patch branch on MSR-VTT.

Method	Text-to-Video (t2v)					Video-To-Text (v2t)				
	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
Base_model	46.9	73.8	82.8	**2.0**	13.7	45.9	74.2	83.6	**2.0**	9.4
Base_model + DR	47.5	74.4	82.9	**2.0**	13.3	46.6	74.4	83.7	**2.0**	9.1
Base_model + T-S	47.9	74.5	83.1	**2.0**	12.9	46.8	74.5	84.0	**2.0**	8.8
Base_model + DR + T-S	**48.2**	**74.9**	**83.3**	**2.0**	**12.6**	**47.4**	**74.8**	**84.1**	**2.0**	**8.7**

Table 1. Ablation study of the key components on MSR-VTT. DR: dynamic routing. T-S: T-S patch branch.

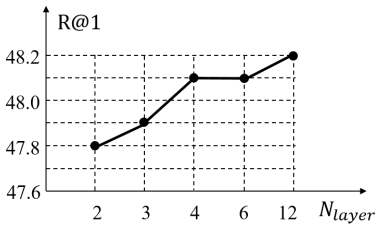

Figure 2. R@1 accuracies for text-to-video retrieval with different numbers of the layers in the temporal transformer of the T-S frame branch on MSR-VTT.

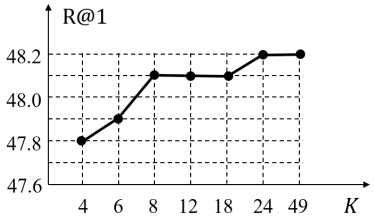

Figure 3. R@1 accuracies for text-to-video retrieval with different numbers of selected patch tokens K on MSR-VTT. The number of the layers of the temporal transformer in the T-S patch branch is 12.

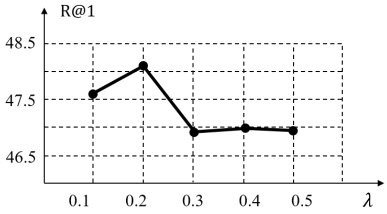

Figure 4. R@1 accuracies for text-to-video retrieval with different weights λ on MSR-VTT.

Base_model has only the text transformer and the S-T frame branch. As can be seen, either DR or T-S is helpful, and the full model equipped with both components gives the best result. This study shows the effectiveness of the DR and T-S modules on generating informative word and patch representations.

Temporal Transformer in T-S. We analyze the effect of the layer number N_l of the temporal transformer in the T-S patch branch. The R@1 accuracies of different layer numbers for text-to-video (t2v) retrieval are presented in Fig. 2. As can be seen, the accuracy usually drops when fewer layers are employed. However, the drop is slight and the accuracy is reduced only by 0.4%, even in the worst case, while is still better than state-of-the-art methods compared later. This demonstrates that conducting patch-level temporal attention is effective for learning informative patch representations, while the size of this transformer can be small.

Method	K	N_l	Inference time (s)	Memory (MB)	Params (M)	FLOPs (G)
B	-	-	22.07	8521	138	53
B + DR	-	-	22.11	8539	139	54
B + T-S	49	12	25.67	10675	178	66
B + DR + T-S	49	12	25.87	10687	179	67
B + DR + T-S	49	4	23.65	10582	152	58
B + DR + T-S	8	12	25.71	9092	179	67

Table 2. Efficiency study of PIDRo on MSR-VTT. B represents Base_model.

Patch Selector. The patch selector is used to select the K most informative patch tokens for similarity calculation. We study the effect of different K. The accuracies of R@1 for text-to-video retrieval are shown in Fig. 3. It is observed that our model gives the best performance when all patch tokens are selected. The accuracy decreases only a little when $K = 8$. This validates the observation that only a few tokens are valuable among all the patches and our patch selector is able to pick them out.

Branch Balancing Weight. We control the contributions to the final similarity calculation from the S-T frame branch and T-S patch branch with the branch balancing weight λ in Eq. 5. We vary λ in the range of $[0.1, 0.5]$ with a step size of 0.1. As can be seen in Fig. 4, the R@1 accuracy increases as λ goes from 0.1 to 0.2 and decreases while λ continues to increase. Too large λ makes the model focus more on patch representations, and vice versa makes it rely more on the pre-trained knowledge of CLIP. As a result, we choose $\lambda = 0.2$ for all the following experiments.

Efficiency Study. In Table 2, we give the efficiency study of our PIDRo. We conduct model inference on MSR-VTT with a single V100 GPU and a batch size of 128 and record the inference time and memory consumption. The parameter numbers and FLOPs are also calculated. It can be seen that compared with Base_model, even with both the DR and T-S modules, the computational cost and complexity of the full model (fourth row) increase slightly. Besides, selecting fewer most informative patch tokens K and reducing the layer number N_l improve the model efficiency, with the accuracies dropping a little as shown in Figs. 2 and 3.

Method	Text-to-Video					Video-To-Text				
	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
HERO [24]	16.8	43.4	57.7	-	-	-	-	-	-	-
MDMMT [12]	38.9	69.0	79.7	**2.0**	16.5	-	-	-	-	-
Support Set [31]	30.1	58.5	69.3	3.0	-	30.1	58.5	69.3	3.0	-
CLIP4Clip [29]	44.5	71.4	81.6	**2.0**	15.3	42.7	70.9	80.6	**2.0**	11.6
CLIP2Video [13]	45.6	72.6	81.7	**2.0**	14.6	43.3	72.3	82.1	**2.0**	10.2
X-CLIP [30]	46.1	73.0	83.1	**2.0**	13.2	46.8	73.3	84.0	**2.0**	9.1
CLIP2TV [15]	46.1	72.5	82.9	**2.0**	15.2	43.9	73	82.8	**2.0**	11.1
TS2-Net [27]	47.0	74.5	**83.8**	-	13.0	45.3	74.1	83.7	-	9.2
PIDRo (ours)	**48.2**	**74.9**	83.3	**2.0**	**12.6**	**47.4**	**74.8**	**84.1**	**2.0**	**8.7**
CLIP2TV* [15]	49.3	74.7	83.6	2.0	13.5	46.9	75	**85.1**	2.0	10
TS2-Net* [27]	49.4	75.6	85.3	-	13.5	46.6	75.9	84.9	-	8.9
PIDRo* (ours)	**50.2**	**77.0**	**85.4**	**1.0**	**12.5**	**49.4**	**76.3**	84.6	**1.0**	**8.4**
CLIP2TV* + DSL [9]	52.9	78.5	86.5	1.0	12.8	54.1	77.4	85.7	1.0	9.0
TS2-Net* + DSL [9]	54.0	79.3	87.4	-	-	-	-	-	-	-
PIDRo* (ours) + DSL [9]	**55.9**	**79.8**	**87.6**	**1.0**	**10.7**	**54.5**	**78.3**	**87.3**	**1.0**	**7.5**

Table 3. Retrieval results on MSR-VTT-1kA. The methods with and without * use patch sizes of 16×16 (ViT-B/16) and 32×32 (ViT-B/32), respectively.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CE [26]	19.8	49.0	63.8	6.0	23.1
Support Set [31]	28.4	60.0	72.9	4	-
Straight-CLIP [33]	37.0	64.1	73.8	3.0	-
Frozen [4]	33.7	64.7	76.3	3.0	-
TeachText-CE+ [10]	25.4	56.9	71.3	4.0	-
CLIP4Clip-meanP [29]	46.2	76.1	84.6	**2.0**	10.0
CLIP4Clip-seqTransf [29]	45.2	75.5	84.3	**2.0**	10.3
PIDRo (ours)	**47.5**	**77.5**	**86.0**	**2.0**	**9.2**

Table 4. t2v results on the MSVD dataset.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
MMT [14]	12.9	29.9	40.1	19.3	75.0
Straight-CLIP [33]	11.3	22.7	29.2	56.5	-
MDMMT [12]	18.8	38.5	47.9	12.3	58.0
CLIP4Clip-meanP [29]	20.7	38.9	47.2	13.0	65.3
CLIP4Clip-seqTransf [29]	22.6	41.0	49.1	11.0	61.0
X-CLIP [30]	23.3	43.0	-	-	56.0
TS2-Net [27]	23.4	42.3	50.9	9.0	56.9
PIDRo (ours)	**25.4**	**43.9**	**54.0**	**8.0**	**50.3**

Table 5. t2v results on the LSMDC dataset.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CE [26]	20.5	47.7	63.9	6.0	23.1
ClipBERT+ [23]	21.3	49.0	63.5	6.0	-
MMT [14]	28.7	61.4	-	3.3	16.0
Support Set [31]	29.2	61.6	-	3.0	-
HiT [25]	29.6	60.7	-	3.0	-
CLIP4Clip-seqTransf [29]	40.5	72.4	-	2.0	7.5
X-CLIP [30]	44.3	74.1	-	-	7.9
TS2-Net [27]	41.0	73.6	84.5	2.0	8.4
PIDRo (ours)	**44.9**	**74.5**	**86.1**	**2.0**	**6.4**

Table 6. t2v results on the ActivityNet dataset.

4.4. Comparison with State-of-the-Arts

MSR-VTT-1kA. We compare the proposed PIDRo with other state-of-the-art methods on the five benchmarks. Table 3 presents the results on the MSR-VTT dataset. With the dynamic routing and T-S patch branch, our model is able to capture fine-grained information of texts and videos with rich semantics, resulting in more accurate retrieval. Our method achieves t2v R@1 48.2% and v2t R@1 47.4% and outperforms previous methods significantly as shown in Table 3. In addition, by employing ViT-B/16 as the base encoder and DSL [9] in inference, our PIDRo yields remarkable R@1 accuracies of 55.9% and 54.5% for t2v and v2t, respectively.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CE [26]	16.1	41.1	-	8.3	43.7
ClipBERT [23]	21.1	47.3	61.1	6.3	-
TeachText-CE+ [10]	21.6	48.6	62.9	6.0	-
Frozen [4]	31.0	59.8	72.4	3.0	-
CLIP4Clip-seqLSTM [29]	43.4	69.9	80.2	**2.0**	17.5
CLIP4Clip-meanP [29]	43.4	70.2	80.6	**2.0**	17.5
X-CLIP [30]	45.2	74.0	-	-	14.6
TS2-Net [27]	41.8	71.6	82.0	**2.0**	14.8
PIDRo (ours)	**48.6**	**75.9**	**84.4**	**2.0**	**11.8**

Table 7. t2v results on the DiDeMo dataset.

Other Benchmarks. Tables 4-7 present the t2v retrieval results on the MSVD, LSMDC, ActivityNet and DiDeMo datasets, respectively, where no post-processing such as DSL [9] is used. It can be observed that our method achieves consistent and significant improvements across different datasets, demonstrating the generalization and robustness of PIDRo. Besides, the v2t results on these four datasets are given in the supplemental materials.

4.5. Visualization of Fine-Grained Alignment

In this section, we analyze PIDRo's capability of capturing fine-grained cross-modal correspondence with the dynamic routing and T-S patch branch. We visualize the word-frame alignment and some retrieval examples.

Word-Frame Alignment with Dynamic Routing. The word-frame alignment is conducted based on the similarity

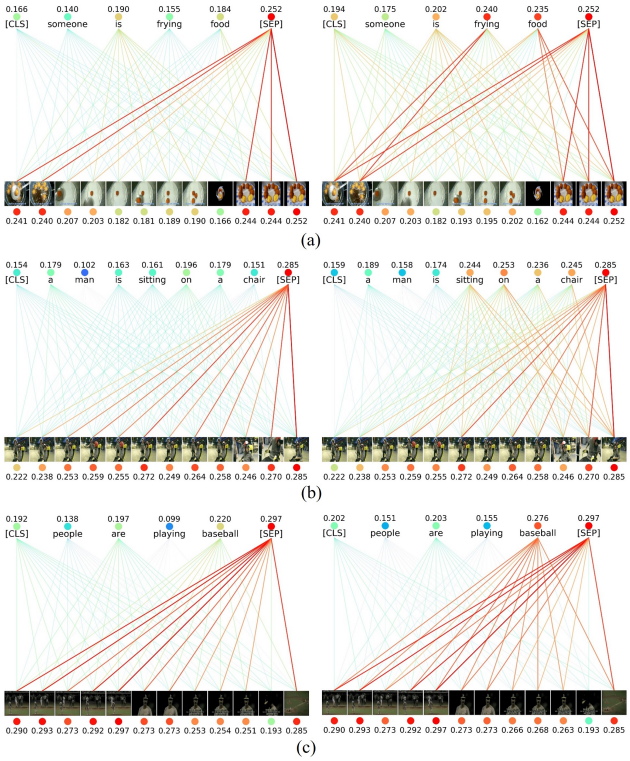

Figure 5. Visualizations of word-frame alignment. The left column shows the alignment without the dynamic routing and the right column is with the dynamic routing. (Best viewed on screen.)

between the word and frame tokens. Specifically, for a word token, we calculate its similarity score to each frame and then connect it with the frame. The visualization result of some examples are presented in Fig. 5, where the red lines indicate large scores and the blue lines small scores, while other colors denote the scores in between. We also present the maximum similarity score for each token. As can be seen, the similarity scores between the key words and related frames are greatly increased when the dynamic routing module is used. For example, in Figs. 5 (a) and (b), our model with the dynamic routing is able to connect the "frying food" and "sitting on a chair" to the related frames of the corresponding videos. In addition, in the last example, the crucial word for retrieving the correct video, 'baseball', is also successfully matched to the related frames when the model is equipped with the dynamic routing. These visualization results further validate the effectiveness of our dynamic routing module on enhancing word representations.

Retrieval with T-S Patch Branch. To show the effect of the T-S patch branch, we visualize some text-video retrieval results in Fig. 6. For better observation, we uniformly select 3 frames for each video in the visualization. In the first two examples, our model with the T-S patch branch is able to find the correct videos by distinguishing the small objects, such as 'monkey' and 'open box'. In the last example, without the T-S patch branch, the model does not capture the

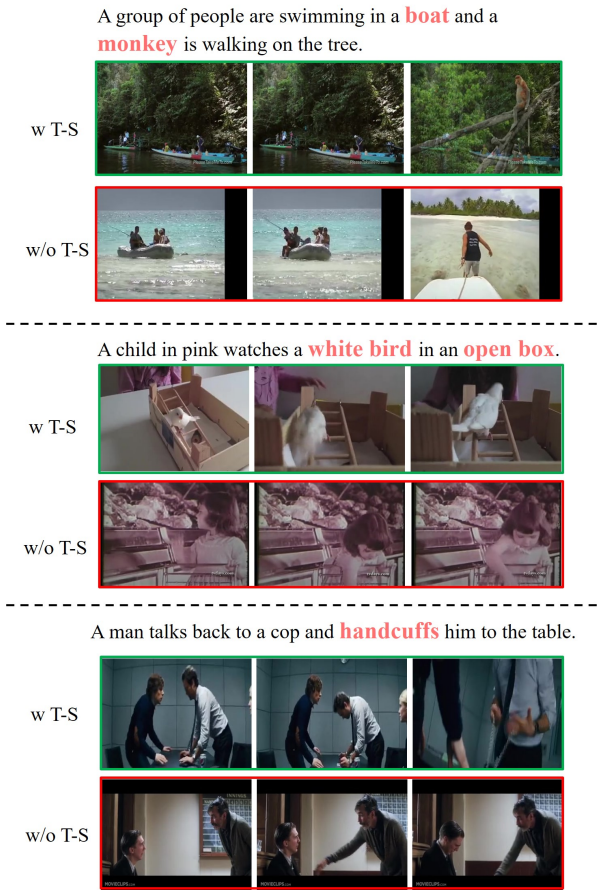

Figure 6. Visualizations of text-video retrieval examples with and without the T-S patch branch. The key words important for video retrieval are in red. The correct results are in green boxes while those incorrect are in red boxes.

subtle movement 'handcuffs'. In comparison, it retrieves the correct video when the T-S patch branch is used. These results again verify the effectiveness of this branch on capturing small objects and insignificant movements in videos.

5. Conclusion

In this paper, we present PIDRo, a fine-grained contrast model that effectively transfers CLIP to the video domain. We build a parallel isomeric attention module tailored for video encoding, which uses another branch to learn fine temporal dynamics of videos. Besides, a dynamic routing module is designed to enhance the word representations through fine-grained information redistribution. The two modules provide informative representations and allows fine-grained cross-domain contrast with token-wise interaction. The experimental results demonstrate the superiority of PIDRo.

Acknowledgements

We gratefully acknowledge the support of MindSpore [1], CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research. This work is also supported in part by the Research Grants Council of Hong Kong (GRF 17201822) and the University of Hong Kong (104006536).

References

[1] https://www.mindspore.cn/. 9

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

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 2

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021. 1, 2, 7

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 2, 5

[6] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable pertubed optimizers. In *NeurIPS*, 2020. 4

[7] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011. 5

[8] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 1

[9] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. In *AAAI*, 2022. 7

[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. 7

[11] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9346–9355, 2019. 1

[12] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *CVPR*, 2021. 1, 7

[13] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 1, 2, 7

[14] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 1, 5, 7

[15] Zijian Gao, Jingyu Liu, Weiqi Sun, Sheng Chen, Dedan Chang, and Lili Zhao. Clip2tv: Align, match and distill for video-text retrieval. *arXiv preprint arXiv:2111.05610*, 2022. 1, 7

[16] Satya Krishna Gorti, Noel 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, 2

[17] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1

[18] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[19] 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*, 2021. 2

[20] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020. 2

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

[22] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 5

[23] 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. 5, 7

[24] 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. 7

[25] 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. 1, 7

[26] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 1, 5, 7

[27] 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. 1, 2, 7

[28] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 2

[29] 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. *arXiv preprint arXiv:2104.08860*, 2021. 1, 2, 5, 7

[30] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP:: End-to-end multi-grained contrastive learning for video-text retrieval. *arXiv preprint arXiv:2207.07285*, 2022. 1, 2, 3, 7

[31] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 7

[32] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *CVPR*, 2023. 2

[33] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *MCPR*, 2021. 7

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021. 1, 2, 3

[35] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 1

[36] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 5

[37] Bin Shao, Jianzhuang Liu, Renjing Pei, Weimian Li, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Hivlp: Hierarchical interactive video-language pre-training. In *ICCV*, 2023. 2

[38] Jie Shao, Xin Wen, Bingchen Zhao, and Xiangyang Xue. Temporal context aggregation for video retrieval with contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3268–3278, January 2021. 2

[39] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 5

[40] 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

[41] 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. *arXiv:2111.07783*, 2021. 2, 3