Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video Retrieval

Jianfeng Dong[1,2], Minsong Zhang[1*], Zheng Zhang[1*], Xianke Chen[1]
Daizong Liu [3], Xiaoye Qu [4], Xun Wang [1,2], Baolong Liu [1,2 †]
[1]Zhejiang Gongshang University, [2]Zhejiang Key Lab of E-Commerce
[3]Peking University, [4]Huazhong University of Science and Technology
https://github.com/HuiGuanLab/DL-DKD

Abstract

Almost all previous text-to-video retrieval works assume that videos are pre-trimmed with short durations. However, in practice, videos are generally untrimmed containing much background content. In this work, we investigate the more practical but challenging Partially Relevant Video Retrieval (PRVR) task, which aims to retrieve partially relevant untrimmed videos with the query input. Particularly, we propose to address PRVR from a new perspective, i.e., distilling the generalization knowledge from the large-scale vision-language pre-trained model and transferring it to a task-specific PRVR network. To be specific, we introduce a Dual Learning framework with Dynamic Knowledge Distillation (DL-DKD), which exploits the knowledge of a large vision-language model as the teacher to guide a student model. During the knowledge distillation, an inheritance student branch is devised to absorb the knowledge from the teacher model. Considering that the large model may be of mediocre performance due to the domain gaps, we further develop an exploration student branch to take the benefits of task-specific information. In addition, a dynamical knowledge distillation strategy is further devised to adjust the effect of each student branch learning during the training. Experiment results demonstrate that our proposed model achieves state-of-the-art performance on ActivityNet and TVR datasets for PRVR.

1. Introduction

With the explosion of online videos, searching the videos of interest has been an indispensable activity in people's daily lives. Meanwhile, text-to-video retrieval (T2VR), retrieving videos *w.r.t.* a textual query from a large number of unlabeled videos, attracts growing attention recently [1, 6, 25, 35, 50, 59]. One basic assumption prerequisite for mainstream T2VR is that videos are pre-trimmed with short duration and supposed to be fully relevant to the query [29, 48, 61]. However, in practical applications, the majority of the existing videos are untrimmed. Besides, as queries are not known a priori, pre-trimmed video clips may not contain sufficient content to fully meet the query. Therefore, there is a huge gap between the literature and the real world for the mainstream T2VR task [8].

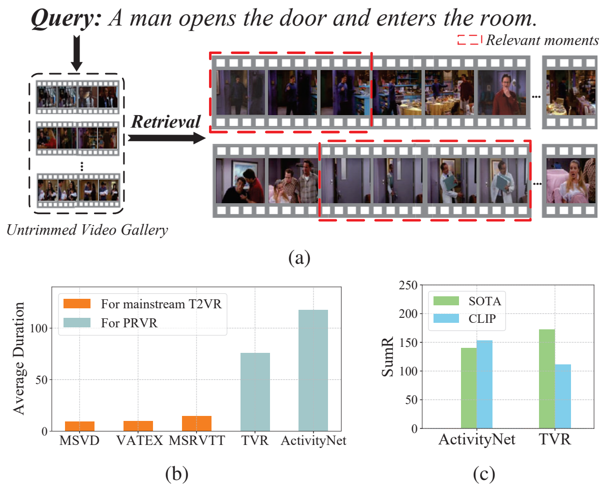


Figure 1. (a) An illustrative example of the PRVR task; (b) Mean values of video duration of different datasets; (c) Performance comparisons between the state-of-the-art (SOTA) [8] and the vanilla CLIP-based [2] methods, where CLIP shows huge performance divergences for different PRVR datasets.

To fill the gap, a new text-to-video retrieval subtask, *i.e.*, Partially Relevant Video Retrieval (PRVR), has been proposed recently in [8]. Different from previous T2VR, PRVR aims to retrieve the partially relevant untrimmed videos that contain at least one internal moment relevant to the given

*Both authors contributed equally to this work.
†Corresponding author: Baolong Liu (liubaolongx@gmail.com)

query (as exemplified in Fig. 1(a)). Besides, videos used for PRVR are much longer than that for T2VR (see Fig. 1(b)). In this work, we target the PRVR task, considering it is more consistent with practical video retrieval scenarios.

Recently, we notice an increasing use of large-scale pre-trained vision and language models, *e.g.*, Contrastive Language-Image Pre-training (CLIP) [46], for various cross-modal tasks, recognition [49, 53], semantic segmentation [55, 65] and person Re-Identification [20, 56], such as text-image retrieval [19, 47], visual question answer [4, 13], and achieving dominant performances. For text-to-video retrieval task, current works [2, 15, 36, 39] mainly focus on the learning of temporal aggregation layers on top of CLIP features, this is because the videos are mainly composed of image sequences while CLIP is trained only on image-text pairs. Different from short videos in these works, the PRVR task contains more complicated untrimmed videos with longer-duration moments of mixed query-relevant and query-irrelevant activities. Therefore, directly treating PRVR as mainstream text-to-video retrieval and aggregating the CLIP features across all the frames may lead to huge performance divergences on PRVR datasets. As shown in Fig. 1 (c), a vanilla CLIP performs superior on ActivityNet but depressing on TVR. Therefore, how to effectively transfer the knowledge of CLIP to PRVR models is still an open problem.

To this end, we propose a Dual Learning framework with Dynamic Knowledge Distillation (DL-DKD) to purify the knowledge of CLIP into the PRVR. Specifically, we develop an effective teacher-student network where the CLIP model is adopted as the teacher and a dual-branch student model is devised to acquire the knowledge. The reason why we introduce two student branches is that CLIP may suffer from domain gap issues due to complicated datasets. Therefore, one inheritance student branch is introduced to directly absorb the beneficial knowledge of the teacher model on a specific domain, while another exploration student branch is utilized to only explore the task-specific property of the training data. In addition, motivated by the fact that human beings first learn from teachers and slowly carry out self-living evolutionary learning once they have formed their own preliminary cognition. Thus, a dynamic knowledge distillation strategy is devised, namely, the inheritance branch takes the prime position at the beginning and the exploration branch gradually becomes more prominent during the training process. In this manner, our DL-DKD is able to take the advantage of both the powerful generalization-ability of CLIP and the benefits of task-specific model convergence on the PRVR data while alleviating their limitations, achieving more robust and effective retrieval. To sum up, the contributions of this work are threefold:

- We propose a knowledge distillation framework that contains a dual-branch student network to acquire appropriate knowledge selectively for partially relevant video retrieval. Meanwhile, our framework supports single-teacher and multiple-teacher distillation.
- We explore how to take the advantage of the powerful generalization-ability of the large model and the benefits of the task-specific model simultaneously while alleviating their limitations, and propose a dynamical knowledge distillation strategy.
- Extensive experiments demonstrate the effectiveness of the above contributions, and our proposed model achieves state-of-the-art performance on the challenging PRVR task.

2. Related Work

2.1. Text-to-video Retrieval

Given a textual query, the task of T2VR [1, 25, 32, 34, 35, 58, 61] aims to retrieve relevant videos with the query from a set of pre-trimmed video clips. The dominant methods typically project videos and queries into a common space for measuring the cross-modal similarity [17, 18, 35, 52]. They usually learn the cross-modal similarity using a large amount of video-text pairs, based on the initial video features extracted by pre-trained vision models and the text features obtained by pre-trained language models. Additionally, we observe an increasing use of large-scale pre-training vision-language models, such as CLIP [46], for text-to-video retrieval [2, 15, 23, 36, 39]. For instance, Hu *et al*. [23] utilize CLIP to extract both video and text features as extra features. Other works adapt CLIP for text-to-video retrieval by introducing the similarity calculation module between the representation of text and video frames [39], frame-wise attentions [2], and a temporal difference block for capturing motions between frames [15].

In practice, videos are generally untrimmed containing much background content [24, 45, 54, 60, 64]. However, in the traditional T2VR, videos are typically pre-trimmed with short duration and are supposed to be fully relevant to the query [29–31, 33, 48, 61], which leads to a huge gap between the literature and the real world. To overcome this limitation, a new text-to-video retrieval subtask, *i.e.*, PRVR, has been proposed [8]. By contrast, videos in PRVR are typically untrimmed, and it aims to retrieve partially relevant videos with the query. An untrimmed video is considered to be partially relevant to a given textual query if it contains a moment relevant to the query. Although it is more consistent with real applications, PRVR had been neglected for a long time.

2.2. Knowledge Distillation

Knowledge distillation is the process of transferring knowledge from a large model (teacher) to a smaller one

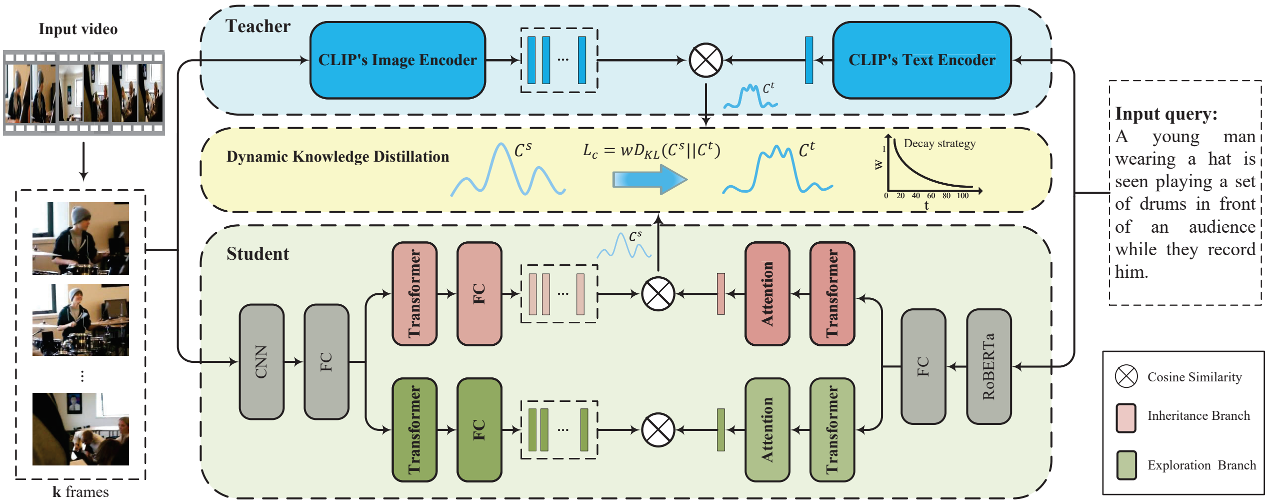


Figure 2. An overview of our proposed DL-DKD framework. Given an untrimmed video and query input, we design a teacher-student network to transfer the knowledge of a large-scale pre-training model CLIP into the PRVR. In detail, we first take the CLIP model as the teacher to provide the generalization knowledge. Then, a dual-branch student model is devised to control the learning effect on knowledge transfer with a dynamic knowledge distillation strategy. In particular, an inheritance student branch is introduced to absorb the beneficial knowledge of the teacher model, while an exploration student branch is utilized to explore the task-specific property of the training data.

(student) [22], which has been widely employed in various tasks, such as image classification [44], and object detection [5]. Recently, a number of works demonstrate that knowledge distillation is also beneficial to learning text-to-video retrieval models [7, 16, 41]. For instance, Croitoru *et al*. [7] distill multiple strong text encoders into one text encoder. Miech *et al*. [41] aims to learn an efficient text-to-visual retrieval model via distilling from a cross-attention model of high performance to a dual encoder model. These works usually have a prerequisite that teacher models have pretty good performance. In this work, we relax this prerequisite, allowing teacher models of mediocrity to be used for knowledge distillation.

3. Method

We propose a dual learning framework with dynamic knowledge distillation for the PRVR task, which exploits the knowledge of CLIP as the teacher guidance to dynamically distill and balance the learning of dual student branches. As shown in Fig. 2, the whole architecture mainly consists of two parts: 1) *Teacher model*: To leverage the powerful generalization ability of the large model trained on the large-scale data, a vision-language pre-training model like CLIP [46] is taken to serve as the teacher model to guide the task-specific model converge on the PRVR data. 2) *Student model*: Since the vanilla pre-trained large models may have huge performance differences on different datasets due to their task-specific domain gaps, a single-branch student model may simply get stuck into the underfitting problem by the above teacher model. To learn to selectively acquire appropriate knowledge from the teacher model when it performs differently on various data, we propose a two-branch, *i.e.*, *Inheritance-Exploration* student network. This student model is split into two branches, in which one branch is utilized to inherit the beneficial knowledge of the teacher model on a specific data domain, while the other is to explore and fit the domain-specific property on the training data. A dynamical knowledge distillation strategy is further applied on the teacher-student framework to adjust the effect of each student branch learning during the training. In the following, we will illustrate the details of each component.

3.1. CLIP Teacher Model

The large-scale vision-language model CLIP [46] was pre-trained on a great amount of image-text data, and is now commonly employed as a strong vision-language backbone enabling zero-shot knowledge transfer to various downstream tasks [36, 51]. Therefore, we also resort to CLIP, utilizing it as the teacher model to appropriately guide our student model training. Note that other vision-language pre-training models, such as TCL [39], can also be employed here. We conduct experiments with different teacher models in Section 4.4, showing the remarkable generalizability of our proposed framework.

As depicted in the top of Fig. 2, given a video-text pair (V, Q) consisting of a video $V = \{I_i\}_{i=1}^{k}$ of k frames and a textual query Q as input, we feed them into the CLIP's image and text encoders to obtain the corresponding video feature $F^t = \{f_i^t\}_{i=1}^{k} \in \mathbb{R}^{d \times k}$ and query feature $q^t \in \mathbb{R}^d$, respectively. The video feature is comprised of a sequence of k frame features, and the dimensions of the frame and the query features are both d. Considering that semantic-aligned similarity matters a lot in our retrieval task, we aim to transfer the collected knowledge of video-query semantic-aware similarity distribution from the teacher model to the student model. Formally, for a pair of video V and query Q, their semantic similarity distribution $C^t \in \mathbb{R}^k$ is formulated as:

$$C^t = [cos(f_1^t, q^t), cos(f_2^t, q^t), ..., cos(f_k^t, q^t)], \quad (1)$$

where cos denotes the cosine similarity.

Different from many works [21,40] devoting to distilling image features from a teacher model to a student model, our teacher model is committed to guide the student model by constraining the consistent semantic similarity distributions between the teacher-student models.

3.2. Dual-branch Student Model

To inherit the beneficial knowledge of the teacher model on a specific data domain while learning to explore and fit the domain-specific property on the training data, we develop a dual-branch student model. As illustrated in Fig. 2, the student model contains two branches, *i.e.*, an inheritance branch and an exploration branch. Specifically, the inheritance student branch is devised to absorb the large-scale knowledge from the teacher branch. Besides, the exploration student branch is introduced to learn the data-specific property by fitting the training data to alleviate the teacher's performance-drop problem due to the domain gaps. By jointly training the two student branches, we can obtain prime performance not only on datasets with a similar distribution to the training data of CLIP but also on datasets with distinct domain gaps.

3.2.1 Inheritance Student Branch

As for the inheritance branch, it is expected to learn the collected knowledge from the semantic similarity distribution C^t of the teacher model.

Multi-Modal Encoding. Given an input video $V = \{I_i\}_{i=1}^k$, a pre-trained 2D CNN with an FC layer is employed to extract the higher-level CNN features of the video as $F^{s'} \in \mathbb{R}^{z \times k}$, where each video frame is represented as a z-dimensional feature vector. Then, after an operation of a standard Transformer with positional embedding and another FC layer, $F^{s'}$ is projected into the joint latent space for the latter multi-modal similarity measurement. This encoded visual feature $F^s \in \mathbb{R}^{z \times k}$ is denoted as:

$$F^s = \{f_1^s, f_2^s, ..., f_k^s\} = FC(Trans(FC(F^{s'}) + PE)), \quad (2)$$

where PE stands for positional embedding, and $Trans$ is a standard Transformer.

For an input query Q, following [8, 27], we utilize the pre-trained RoBERTa [38] with an FC layer to generate the word-level features $Q^{s'} = \{w_i^{s'}\}_{i=1}^{n_s} \in \mathbb{R}^{z \times n_s}$. To further obtain the contextual features of the query text, we first feed $Q^{s'}$ into a standard Transformer to obtain $Q^s = \{w_i^s\}_{i=1}^{n_s} \in \mathbb{R}^{z \times n_s}$, and then employ an attention layer to generate the sentence-level feature $q^s \in \mathbb{R}^z$ via attentive aggregation as:

$$q^s = \sum_{i=1}^{n_s} \alpha_i \times w_i^s, \quad \alpha = Softmax(WQ^s), \quad (3)$$

where $Softmax$ denotes the softmax layer, $W \in \mathbb{R}^{1 \times z}$ is a trainable variable, and $\alpha \in \mathbb{R}^{1 \times n_s}$ indicates the attention vector. q^s is in the joint latent space with F^s.

Transferring Knowledge from Teacher to Student. To absorb knowledge from the teacher by learning the consistency of video-query semantic similarity distribution between the teacher and student branches further, we first calculate the similarity distribution $C^s \in \mathbb{R}^k$ of current student branch between F^s and q^s:

$$C^s = [cos(f_1^s, q^s), cos(f_2^s, q^s), ..., cos(f_k^s, q^s)]. \quad (4)$$

Then, we design a distribution distillation to transfer knowledge from the pre-trained teacher model to the inheritance student branch. Specifically, our distribution distillation strategy is to capture the consistency of the similarity distributions of the teacher and the student. A similarity consistency constraint is configured to guide the learning of the inheritance branch with the teacher model [51]. In detail, given the teacher-similarity distribution C^t and the student-similarity distribution C^s of a video-text pair (V, Q), the semantic consistency loss $\mathcal{L}_c$ is formulated by exploiting the KL divergence as:

$$\mathcal{L}_c = D_{KL}(C^s || C^t) = \sum_{i=1}^{k} C_i^s log \frac{C_i^s}{C_i^t}, \quad (5)$$

where the subscript i indicates the i-th elements in the corresponding similarity distribution.

Besides, we also employ self-similarity learning to make the partially relevant video-text pairs near and irrelevant pairs far away in the learned space. Following the previous work [8], by constructing the positive and negative video-text samples, both triplet ranking loss [11, 14] and InfoNCE loss [42, 63] are jointly utilized to self-train the inheritance branch, which can be noted as $\mathcal{L}_s$. Overall, to train our inheritance student branch, we simultaneously optimize both $\mathcal{L}_s$ and $\mathcal{L}_c$ to learn the self-similarity and the similarity consistency. The final loss $\mathcal{L}_I$ of this branch can be termed as a weighted sum of $\mathcal{L}_s$ and $\mathcal{L}_c$ as:

$$\mathcal{L}_I = \mathcal{L}_s + w\mathcal{L}_c, \quad (6)$$

where w is a hyper-parameter to balance the contribution of $\mathcal{L}_s$ and $\mathcal{L}_c$.

3.2.2 Exploration Student Branch

Since the teacher model can not always perform well on various data due to the domain gap, the inheritance student branch may be prone to its mistake when the teacher model is of mediocre performance. Therefore, we design another student branch, called the exploration branch, to only learn the data-specific property of the training set without any guidance from the teacher branch. By jointly optimizing the two branches in a dual learning manner, we can effectively take advantage of the teacher models on well-performing data while mitigating the negative impact of the teacher model's performance degradation on certain data. In contrast to the inheritance branch that updates its similarity distribution by referring to the teacher knowledge, the exploration branch is devised to learn the data-specific knowledge directly from the on-site training data. As shown in Fig. 2, both two branches share the same network architecture. We also utilize the triplet ranking loss [11, 14] and InfoNCE loss [42,63] to jointly train the exploration branch, and note the overall loss as $\mathcal{L}_E$.

3.3. Dynamic Knowledge Distillation

Although we can directly joint learn the two student branches, this training process remains two-aspect concerns: (1) Firstly, as we mentioned, the CLIP teacher model may have huge performance differences on different datasets due to their task-specific domain gaps. Therefore, when the teacher model is of mediocre performance, how to reduce the impact of the inheritance branch while strengthen the exploration branch learning is important. (2) Secondly, it is worth noticing that continuously pushing the student model to mimic the similarity distribution of the teacher model during the whole training period may limit the student model in acquiring the data-specific knowledge.

Based on the above two observations, we propose a dynamical learning paradigm to adjust and distill the knowledge learned from the dual branches. The main idea is: learning more knowledge from the teacher at the beginning of the training when the knowledge of the teacher is beneficial, while learning more from the on-site data gradually otherwise when the student model getting stronger. Specifically, to obtain a more balanced and better distillation result from the dual-branch learning, a dynamic distillation strategy is introduced. It is devised to tune the hyper-parameter w in Eq.(7) online during the model training instead of setting it to a fixed constant one like most previous works. At the beginning of the training, we set a larger initial value to w to learn more knowledge from the teacher, then we decay w smoothly according to the training epochs. Formally, w is computed as:

$$w = w_0 g(t), \tag{7}$$

where w_0 is an initial weight, t indicates t-th epoch during the training, and $g(\cdot)$ is the decay strategy function.

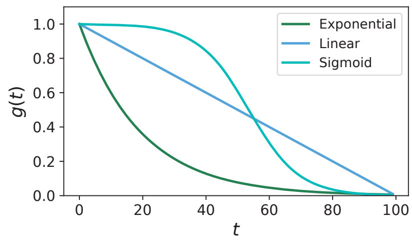


Figure 3. Different decay strategies during the training.

Particularly, inspired by [3, 37], we implement the dynamic distillation strategy with three different types of decay strategy functions, and the experiments demonstrate the results are all promising. As illustrated in Fig. 3, the three decay strategy functions are:

- Exponential decay: $g(t) = k^t$, where $k < 1$ is the factor that control the decay trend.
- Linear decay: $g(t) = kt + b$, where $k < 0$ and b is for controlling the downward trend of the slash.
- Sigmoid decay: $g(t) = \frac{k}{k+e^{\frac{t}{k}}}$, where k is a hyper-parameter to control the decay.

3.4. Training and Inference

During the whole framework training, we only optimize the student model by jointly minimizing the losses of the inheritance branch and the exploration branch. The overall training loss $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_E. \tag{8}$$

During the inference, only the student model is utilized to obtain the retrieval results. Given a video-text pair, we first calculate their similarities from both inheritance and exploration branches, resulting in $S_I(Q, V)$ and $S_E(Q, V)$. The final similarity is computed as:

$$S(Q, V) = (1 - \beta)S_I(Q, V) + \beta S_E(Q, V), \tag{9}$$

where β is a hyper-parameter to balance the two similarities. Given a textual query, all candidate videos are sorted in terms of their final similarities with the query.

4. Experiments

4.1. Experimental Setup

4.1.1 Datasets

In order to validate the effectiveness of our model, we adopt the long untrimmed video datasets ActivityNet Captions [26] and TVR [27]. Note that the pre-trained CLIP

Table 1. The effectiveness of dual learning with both inheritance and exploration branches. Our proposed model not only outperforms the single-branch counterparts but also performs better than simple two-branch baselines.

Branch		**ActivityNet**					**TVR**				
Inheritance	**Exploration**	R@1	R@5	R@10	R@100	SumR	R@1	R@5	R@10	R@100	SumR
✓	✗	7.6	23.4	35.4	76.2	142.6	11.1	28.8	39.3	80.5	159.7
✗	✓	5.9	20.0	32.3	73.7	131.9	11.0	28.5	39.1	81.9	160.6
✓	✓	**8.0**	**25.0**	**37.5**	**77.1**	**147.6**	**14.4**	**34.9**	**45.8**	**84.9**	**179.9**
Double-Inheritance		6.9	22.4	34.8	75.9	140.0	12.5	31.9	43.0	83.1	170.5
Double-Exploration		6.5	22.4	34.6	75.8	139.3	13.1	32.9	43.4	83.4	173.1

performs well on Activitynet Captions, but mediocrely on TVR. Here we briefly introduce these two datasets.

ActivityNet Captions [26] is originally developed for dense video captioning task. As captions are partially relevant with the corresponding videos (a caption is typically associated with a specific moment in a video), it has been re-purposed for partially relevant video retrieval. It contains around 20K videos from YouTube, and the average length of videos is around 118 seconds. On average, each video has around 3.7 moments with a corresponding sentence description. For a fair comparison, we adopt the same data partition used in [8]. For ease of reference, we refer to the dataset as ActivityNet.

TV show Retrieval (TVR) [27] is originally developed for video corpus moment retrieval, and now can be also used for partially relevant video retrieval. It contains 21.8K videos collected from 6 TV shows, and the average length of videos is around 76 seconds. Each video is associated with 5 natural language sentences that describe a specific moment in the video. As a moment is typically a part of a video, sentences are partially relevant to videos. We utilize the same data partition in [8, 62, 63].

4.1.2 Evaluation Metrics

Following the previous work [8], we utilize the rank-based metrics, namely $R@K$ ($K = 1, 5, 10, 100$). $R@K$ stands for the fraction of queries that correctly retrieve desired items in the top K of the ranking list. The performance is reported in percentage (%). The SumR is also utilized as the overall performance, which is defined as the sum of all recall scores. Higher scores indicate better performance.

4.1.3 Implementation Details

For the CLIP teacher model, we adopt a Vision Transformer based ViT-B/32 provided by OpenAI[1], and encode video frames and query sentences to 512-D features. For the student model, we directly utilize the video and sentence features provided by [8] as the input. For the model training, we set the initial learning rate to 0.00025 and use the same learning schedule as [27]. We use the early stop schedule that the model will stop when the evaluated SumR exceeds 10 epochs without promotion. The maximum number of epochs is set to 100. We choose exponential decay as the default one unless otherwise stated, where the initial weight w_0 is 0.1 and the hyper-parameter k in exponential decay is 0.95. During the inference, we empirically set the weights of the inheritance branch and the exploration branch to 0.3 and 0.7 for similarity fusion. Additionally, we use PyTorch to build the model framework and train models on NVIDIA RTX 3090 GPU with a batch size of 128.

[1] https://github.com/openai/CLIP

4.2. Ablation Studies

4.2.1 Effectiveness of Dual Learning

In order to verify the effectiveness of our proposed dual learning with both inheritance and exploration branches, we compare it to the counterparts with the inheritance branch or exploration branch only. The results on both ActivityNet and TVR are summarized in Table 1. Note that the teacher model CLIP we used performs pretty well on ActivityNet, while it performs mediocrely on TVR. On both datasets, the model with both branches consistently performs the best, which demonstrates the effectiveness of our proposed dual learning structure with both inheritance and exploration branches. Especially on TVR where the pre-trained CLIP only achieves SumR score of 110, the model using dual learning obtains a relative SumR gain of around 12% when compared to single-branch counterparts, which is more significant than that on ActivityNet. The result demonstrates that dual learning is more important when the teacher model is of mediocre performance.

Additionally, we also try to verify whether the improvements come from the combination of two branches. We compare our model to the baselines of simply combining two exploration branches (Dual-exploration) or two inheritance branches (Dual-inheritance) without our dynamic distillation strategy. Their worse performance compared to ours demonstrates that the architecture of our dual-branch exploration and inheritance with the dynamic distillation contributes a lot to the final performance.

Table 2. The effectiveness of dynamic knowledge distillation. Note that *Fixed* indicates the model using knowledge distillation with a fixed weight during the training.

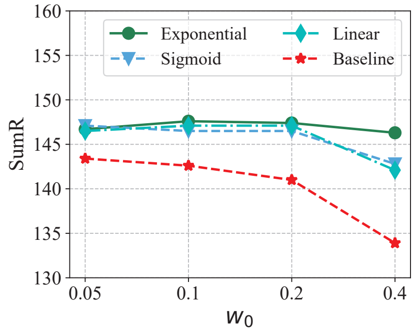
Branch	Distillation	ActivityNet					TVR				
		R@1	R@5	R@10	R@100	SumR	R@1	R@5	R@10	R@100	SumR
Single	✗	5.9	20.0	32.3	73.7	131.9	11.0	28.5	39.1	81.9	160.6
	✓ (Fixed)	**7.6**	23.4	35.4	**76.2**	142.6	11.1	28.8	39.3	80.5	159.7
	✓ (Dynamic)	7.5	**24.1**	**36.2**	76.0	**143.8**	**12.3**	**30.3**	**41.1**	**82.5**	**166.1**
Dual	✗	6.8	22.3	34.5	75.6	139.1	13.4	32.9	43.4	83.4	173.1
	✓ (Fixed)	7.7	24.9	36.6	77.1	146.4	14.1	33.1	44.1	84.6	175.9
	✓ (Dynamic)	**8.0**	**25.0**	**37.5**	**77.1**	**147.6**	**14.4**	**34.9**	**45.8**	**84.9**	**179.9**

4.2.2 Effectiveness of Dynamic Knowledge Distillation

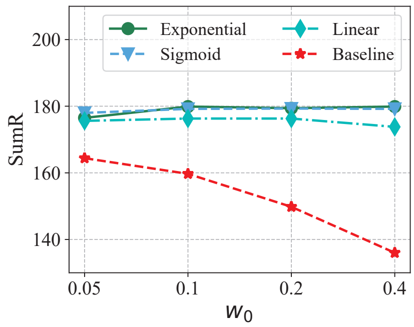
Table 2 shows the ablation study results of dynamic knowledge distillation for both single-branch and dual-branch networks. We compare it to the counterparts without any knowledge distillation or using knowledge distillation with a fixed weight. The fixed weight is set to 0.1 which is the same as the initial weight used in our full model. To ease of reference, we refer to the latter as fixed knowledge distillation. For the single-branch, we observe a phenomenon that the fixed knowledge distillation is beneficial on ActivityNet, but it hurts the performance on TVR. It allows us to conclude that the common knowledge distillation with a fixed weight is not suitable when the performance of the teacher model is mediocre. By contrast, on both datasets our proposed dynamic knowledge distillation consistently achieves performance gain over the ones without the distillation. For the dual-branch, the fixed knowledge distillation also consistently improves the performance on both datasets, but still worse than our dynamic knowledge distillation. The results not only again confirm the effectiveness of our dual learning with two branches, but also demonstrate the advantage of the dynamic knowledge distillation.

4.2.3 Influence of Decay Strategies

In this section, we explore three decay strategies with various initial weights, and also include the model using the distillation with the fixed weight as the baseline. The results on ActivityNet and TVR are demonstrated in Fig. 4. On the whole, the three decay strategies give similar results. Considering the relatively more stable performance of exponential decay, we choose it as the default decay strategy. Additionally, all three variants with dynamic knowledge distillation consistently outperform the baseline, which again confirms the effectiveness of our proposed model. What is more, we find that dynamic knowledge distillation is much less sensitive to the initial weight than the baseline. It makes the dynamic knowledge distillation more appealing, as it alleviates the cumbersome efforts of hyper-parameter tuning.



(a) ActivityNet



(b) TVR

Figure 4. The influence of decay strategies in our dynamic knowledge distillation. The three decay strategies give comparable performances. Besides, they are not very sensitive to the initial weight, making them appealing for hyper-parameter tuning. 7

4.3. Comparison with the State-of-the-Art

Table 3 summarizes the comparison results with other methods on ActivityNet. Our proposed model outperforms all the competitor models with clear margins. Among all methods, only our model utilizes the knowledge distillation, the results justify the viability of using the knowledge distillation for partially relevant video retrieval. In addition, although the previous best-performing model MS-SL also utilizes two branches, their two branches solely learn from the training data without extra knowledge. By contrast, in our model with the dual learning paradigm where one branch

Table 3. Performance comparison on ActivityNet. Models are sorted in ascending order in terms of their overall performance.

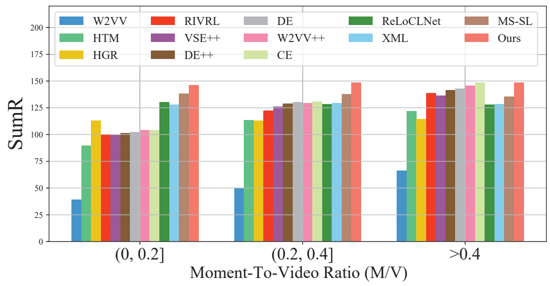
Model	R@1	R@5	R@10	R@100	SumR
W2VV [9]	2.2	9.5	16.6	45.5	73.8
HTM [43]	3.7	13.7	22.3	66.2	105.9
HGR [6]	4.0	15.0	24.8	63.2	107.0
RIVRL [12]	5.2	18.0	28.2	66.4	117.8
VSE++ [14]	4.9	17.7	28.2	67.1	117.9
DE++ [11]	5.3	18.4	29.2	68.0	121.0
DE [10]	5.6	18.8	29.4	67.8	121.7
W2VV++ [28]	5.4	18.7	29.7	68.8	122.6
CE [35]	5.5	19.1	29.9	71.1	125.6
ReLoCLNet [63]	5.7	18.9	30.0	72.0	126.6
XML [27]	5.3	19.4	30.6	73.1	128.4
MS-SL [8]	7.1	22.5	34.7	75.8	140.1
DL-DKD (Ours)	**8.0**	**25.0**	**37.5**	**77.1**	**147.6**

Table 4. Performance comparison on the TVR dataset.

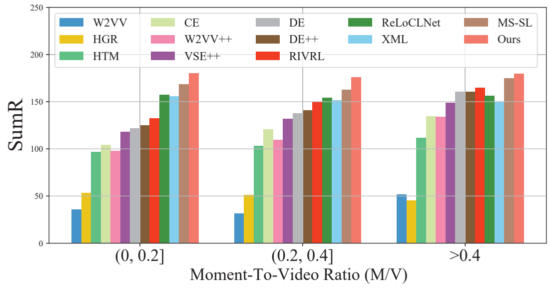
Model	R@1	R@5	R@10	R@100	SumR
W2VV [9]	2.6	5.6	7.5	20.6	36.3
HGR [6]	1.7	4.9	8.3	35.2	50.1
HTM [43]	3.8	12.0	19.1	63.2	98.2
CE [35]	3.7	12.8	20.1	64.5	101.1
W2VV++ [28]	5.0	14.7	21.7	61.8	103.2
VSE++ [14]	7.5	19.9	27.7	66.0	121.1
DE [10]	7.6	20.1	28.1	67.6	123.4
DE++ [11]	8.8	21.9	30.2	67.4	128.3
RIVRL [12]	9.4	23.4	32.2	70.6	135.6
XML [27]	10.0	26.5	37.3	81.3	155.1
ReLoCLNet [63]	10.7	28.1	38.1	80.3	157.1
MS-SL [8]	13.5	32.1	43.4	83.4	172.4
DL-DKD (Ours)	**14.4**	**34.9**	**45.8**	**84.9**	**179.9**

mainly learns from the teacher model and the other learns from the training data. The better performance of our model demonstrates the effectiveness of our proposed dual learning paradigm. Table 4 demonstrates the results on TVR, where our proposed model still performs the best in terms of all metrics.

Thus far all the comparisons are holistic. To gain a more fine-grained comparison, we group the test queries according to their *moment-to-video ratio* (M/V) [8]. M/V of the query is defined as its relevant moment's length ratio in the entire video. The smaller M/V indicates less relevant content while more irrelevant content in the target video with respect to the query, showing more challenging of the corresponding queries. Fig. 5 demonstrates the results on ActivieyNet and TVR. Our proposed model consistently performs the best, which again verifies its effectiveness. Additionally, for each competitor model, they usually perform worse in the group of lower M/V than that of higher M/V.



(a) ActivityNet



(b) TVR

Figure 5. Performance of different models on different types of queries. Queries are grouped according to their M/V. The smaller M/V indicates more challenging of queries.

Table 5. Performance with different teacher models. Our proposed framework supports various teacher models, and also allows for distilling for multiple teachers jointly.

Dataset	Teacher	R@1	R@5	R@10	R@100	SumR
ActivityNet	CLIP	8.0	25.0	37.5	77.1	147.6
	TCL	7.3	24.1	36.2	76.4	144.0
	CLIP+TCL	8.1	25.3	37.7	77.6	148.6
TVR	CLIP	14.4	34.9	45.8	84.9	179.9
	TCL	13.5	33.1	44.6	84.1	175.3
	CLIP+TCL	15.1	35.4	46.5	84.5	181.6

By contrast, our proposed model achieves more balanced performance in all groups, which shows that our model is less sensitive to irrelevant content in videos.

4.4. Extension to Multi-Teacher Distillation

While the focus of our work in the Method Section is only the single-teacher distillation, it is natural to consider whether the framework can be extended to multi-teacher distillation. Therefore, we adopt another vision-language pre-training model TCL [57] as an extra teacher model. As shown in Table 5, utilizing TCL as the teacher model still gives better performance than the previous state-of-the-art works. Additionally, with the joint use of CLIP and TCL as teacher models (their output distributions are fused by simple summation), it brings a further performance boost over the single-teacher distillation. We believe this extension may be useful in scenarios where various vision-language pre-training models are available during training.

4.5. Complementarity between the two branches

Recall that our proposed model consists of an inheritance branch and an exploration branch, here we explore their complementarity. We measure the complementarity via Pearson correlation coefficient between the similarity distributions of two branches, *i.e.*, C^t and C^s. Besides our model, we also compute the correlation coefficient of the common two-branch baseline without distillation (*i.e*. Double-Exploration). On the ActivityNet dataset, our model achieves a coefficient of 0.622, while the baseline obtains a coefficient of 0.749. Note that the lower coefficient indicates the less correlated between two branches and more complementary. The result demonstrates that the two branches of our model are more complementary, which to some extent illustrates why our model is better than the two-branch baseline.

5. Conclusions

In this paper, we have investigated the meaningful but challenging text-to-video subtask of PRVR from a new perspective of knowledge distillation. A novel framework, *i.e.*, KL-DKD, has been proposed to distill the generalization knowledge from the large-scale vision-language pre-trained model to a task-specific network. Extensive experiments on both ActivityNet and TVR datasets support the following conclusions: (1) Dual learning of an inheritance branch and an exploration branch is necessary for knowledge distillation. (2) Besides, the dynamic knowledge distillation further improves performance, especially when the teacher model is of mediocre performance. (3) For state-of-the-art performance, we recommend dual learning with dynamic knowledge distillation for PRVR.

Acknowledgements. This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No.2023C01212), National Natural Science Foundation of China (No. 61976188), Young Elite Scientists Sponsorship Program by CAST (No. 2022QNRC001), the open research fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems, and the Fundamental Research Funds for the Provincial Universities of Zhejiang.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 1, 2

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A clip-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022. 1, 2

[3] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015. 5

[4] Paola Cascante-Bonilla, Hui Wu, Letao Wang, Rogerio S Feris, and Vicente Ordonez. Simvqa: Exploring simulated environments for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5056–5066, 2022. 2

[5] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 3

[6] 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, 8

[7] Ioana Croitoru, Simion-Vlad Bogolin, Yang Liu, Samuel Albanie, Marius Leordeanu, Hailin Jin, and Andrew Zisserman. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 3

[8] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 246–257, 2022. 1, 2, 4, 6, 8

[9] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 2018. 8

[10] 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. 8

[11] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2022. 4, 5, 8

[12] Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5680–5694, 2022. 8

[13] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021. 2

[14] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, pages 935–943, 2018. 4, 5, 8

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

[16] Sheng Fang, Shuhui Wang, Junbao Zhuo, Xinzhe Han, and Qingming Huang. Learning linguistic association towards

efficient text-video retrieval. In *European Conference on Computer Vision*, pages 254–270, 2022. 3

[17] Zerun Feng, Zhimin Zeng, Caili Guo, and Zheng Li. Exploiting visual semantic reasoning for video-text retrieval. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1005–1011, 2021. 2

[18] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229, 2020. 2

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5006–5015, 2022. 2

[20] Konrad Habel, Fabian Deuser, and Norbert Oswald. Clip-reident: Contrastive training for player re-identification. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, pages 129–135, 2022. 2

[21] Shitian He, Huanxin Zou, Yingqian Wang, Runlin Li, Fei Cheng, Xu Cao, and Meilin Li. Enhancing mid–low-resolution ship detection with high-resolution feature distillation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 4

[22] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3

[23] 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 *European Conference on Computer Vision*, pages 444–461. Springer, 2022. 2

[24] Wei Ji, Renjie Liang, Zhedong Zheng, Wenqiao Zhang, Shengyu Zhang, Juncheng Li, Mengze Li, and Tat-seng Chua. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23013–23022, 2023. 2

[25] Weike Jin, Zhou Zhao, Pengcheng Zhang, Jieming Zhu, Xiuqiang He, and Yueting Zhuang. Hierarchical cross-modal graph consistency learning for video-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1114–1124, 2021. 1, 2

[26] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017. 5, 6

[27] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463, 2020. 4, 5, 6, 8

[28] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. W2VV++: Fully deep learning for ad-hoc video search. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1786–1794, 2019. 8

[29] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. Sea: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 23:4351–4362, 2020. 1, 2

[30] Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *IEEE Transactions on Multimedia*, 2023. 2

[31] Daizong Liu and Wei Hu. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4536–4545, 2022. 2

[32] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021. 2

[33] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4092–4101, 2022. 2

[34] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020. 2

[35] 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, 2, 8

[36] Yu Liu, Huai Chen, Lianghua Huang, Di Chen, Bin Wang, Pan Pan, and Lisheng Wang. Animating images to transfer clip for video-text retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1906–1911, 2022. 2, 3

[37] Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Scheduled sampling based on decoding steps for neural machine translation. *arXiv preprint arXiv:2108.12963*, 2021. 5

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[39] 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*, 508:293–304, 2022. 2, 3

[40] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022. 4

[41] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Effi-

cient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021. 3

[42] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 4, 5

[43] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 8

[44] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020. 3

[45] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4280–4288, 2020. 2

[46] 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. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3

[47] Siyu Ren and Kenny Q Zhu. Leaner and faster: Two-stage model compression for lightweight text-image retrieval. *arXiv preprint arXiv:2204.13913*, 2022. 2

[48] Xue Song, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia*, 24:2914–2923, 2021. 1, 2

[49] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[50] Yabing Wang, Jianfeng Dong, Tianxiang Liang, Minsong Zhang, Rui Cai, and Xun Wang. Cross-lingual cross-modal retrieval with noise-robust learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 422–433, 2022. 1

[51] Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, and Qi Tian. Contextual similarity distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9489–9498, 2022. 3, 4

[52] Peng Wu, Xiangteng He, Mingqian Tang, Yiliang Lv, and Jing Liu. Hanet: Hierarchical alignment networks for video-text retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3518–3527, 2021. 2

[53] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Language supervised training for skeleton-based action recognition. *arXiv preprint arXiv:2208.05318*, 2022. 2

[54] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. Visual relation grounding in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 447–464, 2020. 2

[55] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021. 2

[56] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *arXiv preprint arXiv:2210.10276*, 2022. 2

[57] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 8

[58] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1339–1348, 2020. 2

[59] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, 2021. 1

[60] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing*, 31:1204–1216, 2022. 2

[61] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 471–487, 2018. 1, 2

[62] Bowen Zhang, Hexiang Hu, Joonseok Lee, Ming Zhao, Sheide Chammas, Vihan Jain, Eugene Ie, and Fei Sha. A hierarchical multi-modal encoder for moment localization in video corpus. *arXiv preprint arXiv:2011.09046*, 2020. 6

[63] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–695, 2021. 4, 5, 6, 8

[64] Qi Zheng, Jianfeng Dong, Xiaoye Qu, Xun Yang, Yabing Wang, Pan Zhou, Baolong Liu, and Xun Wang. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–21, 2023. 2

[65] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021. 2