KDProR: A Knowledge-Decoupling Probabilistic Framework for Video-Text Retrieval (Supplementary Materials)

Xianwei Zhuang*, Hongxiang Li*, Xuxin Cheng,
Zhihong Zhu, Yuxin Xie, Yuexian Zou†

ADSPLAB, School of ECE, Peking University, China
{xwzhuang, lihongxiang}@stu.pku.edu.cn, zouyx@pku.edu.cn

In this Appendix, we present the following items:

- **A Proof of Theorem 1**
- **B The Application Details of KDProR Framework**
- **C More Experimental Details**
- **D More Experimental Results and Analysis**
- **E More Discussions about KDProR**
- **F More Visualization Cases**

A Proof of Theorem 1

Theorem 1 *The retrieval model Θ optimized iteratively through the proposed EKM algorithm of step t satisfies the following properties:*

- ***Monotonic Increasing:*** $\forall t \geq 0, \quad \mathcal{L}(\Theta^{(t+1)}) \geq \mathcal{L}(\Theta^{(t)})$;
- ***Convergence:*** $\forall \epsilon > 0, \exists \delta > 0$, *if* $t > \delta, \quad \mid \Theta^{(t+1)} - \Theta^{(t)} \mid \leq \epsilon$.

Theorem 1 indicates that our EKM algorithm can ensure the model is positively optimized in each iteration, while also maintaining the convergence properties of vanilla EM algorithms. We provide the following proof:

Proof. In a closed world, the posterior distribution of z_i^f and z_i^c can be represented as:

$$Q(z_i^f) = p(z_i^f; x_i, \Theta), \quad Q(z_i^c) = p(z_i^c; z_i^f, x_i, \Theta). \tag{1}$$

The goal of optimization is to maximize the probability of video text matching pairs $x_i = (T_i, V_i)$ appearing. The optimization objective of the retrieval model

* These authors contributed equally.
† Corresponding author.

is to maximize the likelihood function as follows:

$$\begin{aligned}\Theta^* &= \arg\max \mathcal{L}(\Theta) \\ &= \arg\max \sum_i \log p\left(x_i;\Theta\right) \\ &= \arg\max \sum_i \log \sum_{z_i^f, z_i^c} p\left(x_i, z_i^f, z_i^c \mid \Theta\right) \\ &= \arg\max \sum_i \log \sum_{z_i^f, z_i^c} Q(z_i^f)Q(z_i^c)\frac{p\left(x_i, z_i^f, z_i^c;\Theta\right)}{Q(z_i^f)Q(z_i^c)}.\end{aligned} \tag{2}$$

Thus we can rewrite the objective function of the retrieval model Θ according to Jensen's inequality [9] as follows:

$$\begin{aligned}\Theta^* &= \arg\max \sum_i \log \sum_{z_i^f, z_i^c} Q(z_i^f)Q(z_i^c)\frac{p\left(x_i, z_i^f, z_i^c;\Theta\right)}{Q(z_i^f)Q(z_i^c)} \\ &\geq \arg\max \sum_i \sum_{z_i^f, z_i^c} Q(z_i^f)Q(z_i^c)\log\frac{p\left(x_i, z_i^f, z_i^c;\Theta\right)}{Q(z_i^f)Q(z_i^c)} \\ &= \arg\max \sum_i \sum_{z_i^f, z_i^c} Q(z_i^f)Q(z_i^c)\log p\left(x_i, z_i^f, z_i^c;\Theta\right).\end{aligned} \tag{3}$$

Therefore, we obtain a lower bound on the likelihood function, and we will maximize the likelihood function value by maximizing this lower bound. Therefore, the optimal parameter is represented as:

$$\Theta^* = \arg\max \sum_i \sum_{z_i^f, z_i^c} Q(z_i^f)Q(z_i^c)\log p\left(x_i, z_i^f, z_i^c;\Theta\right). \tag{4}$$

We further define the value of the latent variable at step t as:

$$Q(z_i^f;\Theta^{(t)}) = p(z_i^f; x_i, \Theta^{(t)}), \quad Q(z_i^c;\Theta^{(t)}) = p(z_i^c; z_i^f, x_i, \Theta^{(t)}). \tag{5}$$

Our EKM algorithm can be seen as a three-step iterative optimization process of Expectation, Knowledge, and Maximization. The E-step and K-step aim to estimate the posterior probability $Q(z_i^f)$ and $Q(z_i^c)$ on fine-grained and coarse-grained knowledge stores with given current $\Theta^{(t)}$. Then KDProR achieves efficient alignment of videos and texts and obtains $\Theta^{(t+1)}$ under knowledge guidance at M-step by maximizing the log-likelihood in Eq. 2 with known $Q(z_i^f)$ and

$Q(z_i^c)$. Therefore, through Eq. 2 and Eq. 3, we can further obtain:

$$\begin{aligned}\mathcal{L}(\Theta) &= \sum_i \log \sum_{z_i^f, z_i^c} Q(z_i^f;\Theta^{(t)})Q(z_i^c;\Theta^{(t)}) \frac{p\left(x_i, z_i^f, z_i^c;\Theta\right)}{Q(z_i^f;\Theta^{(t)})Q(z_i^c;\Theta^{(t)})} \\ &\geq \sum_i \sum_{z_i^f, z_i^c} Q(z_i^f;\Theta^{(t)})Q(z_i^c;\Theta^{(t)}) \log \frac{p\left(x_i, z_i^f, z_i^c;\Theta\right)}{Q(z_i^f;\Theta^{(t)})Q(z_i^c;\Theta^{(t)})} \\ &= F\left(\Theta, \Theta^{(t)}\right),\end{aligned} \tag{6}$$

where, $F\left(\Theta, \Theta^{(t)}\right)$ is a lower bound of the likelihood function $\mathcal{L}(\Theta^{(t+1)})$.

Therefore, we can further derive:

$$\mathcal{L}\left(\Theta^{(t+1)}\right) \geq F\left(\Theta^{(t+1)}, \Theta^{(t)}\right) \geq F\left(\Theta^{(t)}, \Theta^{(t)}\right) = \mathcal{L}\left(\Theta^{(t)}\right), \tag{7}$$

where, $F\left(\Theta^{(t+1)}, \Theta^{(t)}\right) \geq F\left(\Theta^{(t)}, \Theta^{(t)}\right)$ is guaranteed by the maximum likelihood process in M-step. In addition, due to the upper bound of the likelihood function and the monotonically increasing nature (*cf.* Eq. 7) of the EKM algorithm, it can be inferred from the monotonically bounded theorem that the iteration of EKM will eventually converge, which concludes the proof for Theorem 1.

In an open-world setting, the knowledge in the stores is independent of the training set, thus the distribution of latent variables can be represented as:

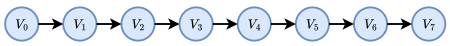
$$Q(z_i^f) = p(z_i^f, \Theta), \quad Q(z_i^c) = p(z_i^c; z_i^f, \Theta). \tag{8}$$

At this point, the latent variables z_i^f and z_i^c and training data x_i are independent of each other. When the introduction of additional knowledge does not conflict with maximum likelihood optimization, our EKM algorithm still conforms to the property of Theorem 1.

Table 1: Examples of structured knowledge of entity-relation information.

Captions	Relations			Entities		
	Subject	Relation	Object	Head	Span	Modifiers
a man in striped collared	man	discusses	jobs	man	a man	a
shirt discusses jobs in	man	in	news room	shirt	collared shirt	collared
news room of bloomberg	news room	of	bloomberg	bloomberg	bloomberg	-
a man with glasses	man	with	glasses	man	a man	a
and a goatee talking	goatee	with	glasses	goatee	a goatee	a
about his former job	goatee	about	jobs	job	former job	former

Table 2: Prompts and examples for extracting LLMs knowledge using CoT.

Original Captions	a scene from the ice age video game is shown
Prompts	This is a subtitle for a video{}. I will take 8 pictures from the video. The previous frame is {}. Please provide possible descriptive sentences for the next pictures:
CoTs	
V_0	A majestic glacier looms in the distance, hinting at the icy challenges ahead.
V_1	Our fearless protagonist, a prehistoric mammal, surveys the frozen landscape with determination.
V_2	Crystalline icicles hang perilously from rocky outcrops, adding to the chilling atmosphere.
V_3	A herd of ancient creatures traverse the snowy expanse, their footprints marking their journey.
V_4	The frigid winds whip through the tundra, carrying whispers of distant adventures.
V_5	A towering ice wall presents a formidable obstacle, testing the resilience of our heroes.
V_6	In the heart of the frozen wilderness, a hidden cave offers refuge from the elements.
V_7	As day fades into night, the aurora dances across the sky, casting an ethereal glow upon the icy terrain.

B The Application Details of KDProR Framework

Video-Text Pretraining. Vanilla CLIP has difficulty understanding spatial-temporal semantics of the video [13]. Aiming at improving temporal understanding of videos, we utilize CLIP-ViP [26] which is further pre-trained on a large-scale video-text dataset HD-VILA-100M [25] to extract pretraining video-text knowledge. Specifically, we adopt video and text encoders of the pre-trained CLIP-ViP to extract video and text features from the training dataset for constructing fine-grained and coarse-grained knowledge stores.

Entity-Relation Structure. Objects, attributes of objects, and relationships between objects are essential to the understanding of textual structure and visual scenes. Previous studies [7] have found vanilla CLIP has difficulty understanding these structures, which are critical to the joint representation of vision and language. In this work, we attempt to parse captions to scene graphs $\mathcal{G}$ through the Scene Graph Parser [23], then employ $\mathcal{G}$ to obtain subject-relation-object triplets. Subsequently, we use these triplets as substitutes for the original captions and then employ CLIP to extract text-video features to construct stores.

Specifically, we use a pre-trained scene graph parsing model to extract structured relationships. We present two parsed entity relationship structures and their corresponding structured triplets in Table 1,

CoT Generation of LLMs. LLMs have demonstrated remarkable success across a wide range of textual tasks. In this work, we attempt to use a strategy similar to CoT [22] to generate auxiliary captions for each video frame by prompting LLMs. Specifically, we treat original captions as "answers" and then prompt LLMs to generate video frame descriptions with contextual relevance step by step as auxiliary captions. Subsequently, we utilize CLIP to extract neural representations of auxiliary captions and corresponding video frames to construct fine- and coarse-grained knowledge stores. We adopt the `gpt-3.5-turbo-1106` [17] to obtain the CoT knowledge.

We concatenate prompts and captions and use LLMs to obtain the description of the first possible frame. Then we concatenate prompts, captions, and descriptions of the previous frame and use LLMs to obtain the next possible frame description. We iteratively obtain descriptions of all possible video frames in sequence. The auxiliary captions obtained through this chain prompt technique are highly consistent with the original captions, while also having the feature of temporal continuity. In Table 2 , we present the auxiliary captions obtained using the CoT technique.

C More Experimental Details

C.1 More Details of Experimental Settings

In K-step, the threshold α of the indicator function $\phi(\cdot, \cdot)$ used for coarse-grained distribution calibration is set to 0.8 on MSR-VTT and DiDeMo datasets, and 0.7 on ActivityNet and LSMDC datasets. We set the dimension of visual and textual representations to 512. We set the number of iterations as 5 across all datasets in the Sinkhorn-Knopp algorithm within E-step. Following [15, 21], the batch size is set to 128 for MSR-VTT, LSMDC, and 64 for ActivityNet and DiDeMo. We use the FAISS library [4] to achieve top-K retrieval of the knowledge stores.

We further explain the different settings of KDProR in all experiments: **KDProR**$^{+}$ denotes we use vanilla CLIP to extract video and text features on the training set to construct knowledge stores. At this point, the stores contain the original knowledge of the closed world. **KDProR**$^{\top}$ represents our utilization of fine-tuning CLIP to extract video and text features and construct knowledge stores. We use base models to fine-tune the CLIP on the training set through five epochs to obtain the fine-tuned CLIP. Here, the knowledge in the stores comes from the original knowledge of CLIP and additional information provided by the training set, which is still a closed-world setting. **KDProR***, **KDProR**† and **KDProR**‡ denote the open-world settings introduced in Section B, respectively.

`https://github.com/vacancy/SceneGraphParser`
`https://platform.openai.com/`
`https://github.com/facebookresearch/faiss`

Table 3: Video-to-Text retrieval results on DiDeMo, ActivityNet and LSMDC datasets without any post-processing (e.g., [2] and [1]). $^{+}$ and $^{\top}$ are closed-set settings and denote constructing knowledge stores on the training set using the original CLIP and the fine-tuned CLIP upon the training set, respectively. *, † and ‡ denote pre-trained, structural and CoT knowledge, respectively.

Method	DiDeMo				ActivityNet				LSMDC			
	R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓
Video-to-Text Retrieval using CLIP (ViT-B/32)												
CLIP4clip [15]	42.5	70.6	80.2	2.0	42.5	74.1	85.8	2.0	20.8	39.0	48.6	12.0
X-CLIP [16]	43.1	72.2	-	-	43.9	73.9	-	-	22.5	42.2	-	-
DRL [21]	45.4	72.6	82.1	2.0	42.2	74.0	86.2	2.0	24.9	44.1	53.8	9.0
PromptSwitch [3]	-	-	-	-	-	-	-	-	22.0	40.8	50.3	-
DiffusionRet [11]	46.2	74.3	82.2	2.0	43.8	75.3	86.7	2.0	23.0	43.5	51.5	9.0
KDProR^{+}	46.4	76.1	84.2	2.0	44.8	74.2	86.2	2.0	24.9	46.4	54.7	7.0
KDProR$^{\top}$	47.1	75.0	83.3	2.0	45.9	75.8	86.9	2.0	25.5	47.6	55.9	7.0
KDProR*	48.4	**77.0**	83.9	**1.0**	46.5	75.9	86.9	**1.0**	25.8	48.3	57.1	**6.0**
KDProR†	47.5	76.7	83.9	2.0	45.8	75.4	87.2	2.0	25.4	46.5	**58.2**	7.0
KDProR‡	**49.1**	76.8	**84.9**	**1.0**	**47.0**	**76.7**	**88.0**	**1.0**	**26.4**	**48.6**	57.1	7.0

D More Experimental Results and Analysis

D.1 Video-to-Text Retrieval on DiDeMo, ActivityNet and LSMDC

As shown in Figure 3, we present a performance comparison of video-to-text retrieval on three datasets, i.e., LSMDC, DiDeMo and ActivityNet. It can be observed that our KDProR achieves better performance than state-of-the-art results in the closed setting. More noteworthy is that our method further improves retrieval performance by introducing open-world knowledge, significantly outperforming state-of-the-art methods in all metrics.

D.2 Video-to-Text Retrieval on Different Base Models.

We further equip KDProR with two powerful base models in addition to DRL, i.e., CLIP4clip [15] and EMCL [10], and evaluate on MSR-VTT. Table 4 shows that KDProR can be applied to successfully boost all baselines as a plug-and-play module under both closed-set or open-world settings. Our method can significantly improve the performance of the base model in both text-to-video and video-to-text retrieval. The significant improvements prove the generalization ability of KDProR.

D.3 Analysis of Out-Domain Settings

To further evaluate the performance of our method in out-domain retrieval settings, we train our KDProR on the source domain dataset (MSR-VTT) and then test it on the target domain (DiDeMo and LSMDC) to evaluate the generalization, as results shown in Table 5. It can be observed that our method

Table 4: Video-to-Text retrieval results on MSR-VTT using our re-implemented CLIP4clip [15] and EMCL [10] as base models. $^{+}$ and $^{\top}$ are closed-set settings, *, † and ‡ are open-world settings, whose meanings are consistent with those in Table 3.

	CLIP4clip [15] as Base Model					EMCL [10] as Base Model				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Base Models	43.7	71.0	81.9	**2.0**	12.0	45.2	73.2	83.1	**2.0**	9.9
+ KDProR^{+}	45.8	73.9	82.6	**2.0**	11.0	46.8	73.8	84.5	**2.0**	8.7
+ KDProR$^{\top}$	**46.9**	74.0	82.9	**2.0**	10.7	46.9	73.7	84.2	**2.0**	8.5
+ KDProR*	46.6	74.1	82.9	**2.0**	10.7	47.1	73.7	**84.7**	**2.0**	8.7
+ KDProR†	46.2	73.9	82.7	**2.0**	10.9	46.9	73.9	84.5	**2.0**	8.3
+ KDProR‡	46.8	**74.6**	**83.2**	**2.0**	**10.5**	**47.3**	**74.2**	84.5	**2.0**	**8.2**

Table 5: Text-to-video retrieval performance in out-domain retrieval settings. "MSR-VTT->DiDeMo" and "MSR-VTT->LSMDC" denote the generalization results on unseen target-domain test datasets (DiDeMo and MSR-VTT) using pre-trained models on the MSRVTT (source domain). $^{\top}$ is closed-set settings and denotes constructing knowledge stores on the training set using the fine-tuned CLIP upon the training set. ‡ denotes CoT knowledge.

Methods	MSR-VTT -> DiDeMo				MSR-VTT -> LSMDC			
	R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓
Text-to-Video Retrieval using CLIP (ViT-B/32)								
CLIP4clip [15]	31.8	57.0	66.1	4.0	15.3	31.3	40.5	21.0
EMCL [10]	30.0	56.1	65.8	**3.0**	16.6	29.3	36.5	24.0
DiffusionRet [11]	33.2	59.3	68.4	**3.0**	17.1	32.4	41.0	21.0
KDProR$^{\top}$	34.7	**61.3**	70.1	**3.0**	18.4	**35.2**	43.1	**19.0**
KDProR‡	**36.8**	60.8	**72.3**	**3.0**	**18.9**	34.5	**43.2**	**19.0**

has significantly better out-domain retrieval performance than other methods in both closed and open-world settings, e.g., +3.6% R@1 text-to-video retrieval gain under domain transfer from MSR-VTT to DiDeMo with our KDProR using CoT knowledge, and +1.8% R@1 text-to-video retrieval gain under domain transfer from MSR-VTT to LSMDC with our KDProR using CoT knowledge. These results further confirm the effectiveness and generalization of our KDProR under out-domain retrieval settings.

D.4 More Results on CLIP (ViT-B/16) and Comparison with More Baselines

We conduct experiments using CLIP(ViT-B/16) as the backbone network and compare our KDProR with more baseline methods, as results shown in Table 6. It can be observed that our KDProR is significantly superior to other retrieval methods in the closed-world set. In addition, our KDProR significantly outperforms the fine-tuning and pre-training methods with open-set settings utilizing additional data to assist retrieval. Impressively, our KDProR is superior to Om-

Table 6: Compared with more baselines using CLIP(ViT-B/16) [15] as the backbone. $^{+}$ and $^{\top}$ are closed-set settings, *, † and ‡ are open-world settings, whose meanings are consistent with those in Table 3. OmniVL [20] and UMT [12] denote the utilization of additional video for open-set pre-training methods.

	Method	Text-to-Video Retrieval					Video-to-Text Retrieval				
		R@1	R@5	R@10	MdR↓	MnR↓	R@1	R@5	R@10	MdR↓	MnR↓
Using CLIP (ViT-B/16) as backbone											
Closed-Set	CLIP2TV [6]	48.3	74.6	82.8	2.0	14.9	46.5	75.4	84.9	**2.0**	10.2
	CenterCLIP [27]	48.4	73.8	82.0	2.0	13.8	47.7	75.0	83.3	**2.0**	10.2
	TS2-Net [14]	49.4	75.6	85.3	2.0	13.5	46.6	75.9	84.9	**2.0**	8.9
	DRL [21]	50.2	**76.5**	84.7	**1.0**	-	48.9	76.3	85.4	**2.0**	-
	UATVR [5]	50.8	76.3	85.5	**1.0**	12.4	48.1	76.3	85.4	**2.0**	8.0
	KDProR^{+}	51.4	76.4	85.7	**1.0**	11.8	49.2	**76.8**	**86.0**	**2.0**	8.1
	KDProR$^{\top}$	**51.7**	76.1	**86.1**	**1.0**	**11.7**	**49.7**	76.6	85.8	**2.0**	**7.9**
Open-Set	OmniVL [20]	47.8	74.2	83.8	-	-	-	-	-	-	-
	UMT-B/5M [12]	46.3	72.7	82.0	-	-	-	-	-	-	-
	TEFAL [8]	49.9	76.2	84.4	2.0	11.4	-	-	-	-	-
	Cap4Video [24]	51.4	75.7	83.9	**1.0**	12.4	49.0	75.2	85.0	**2.0**	8.0
	KDProR*	51.7	76.1	85.5	**1.0**	11.8	49.6	**76.1**	85.7	**2.0**	8.1
	KDProR†	51.5	75.9	84.8	**1.0**	11.6	49.1	75.9	85.4	**2.0**	8.3
	KDProR‡	**52.1**	**76.3**	**85.7**	**1.0**	**11.4**	**49.8**	75.5	**86.0**	**2.0**	**7.8**

niVL [20] and UMT-B/5M [12], which utilize an additional 14M training pairs and 5M training pairs for pre-training.

D.5 Running Analysis

We evaluate the runtime parameters of our method on the server with Intel(R) Gold6330 CPU@2.00GHz and 8 RTX4090, as results shown in Table 7. We evaluate the results of different experiments under the same environmental settings. Our method adopts DRL [21] as the base model and evaluates with a 128 batch size on MSR-VTT dataset.

It can be observed that our KDProR only incurs negligible overhead in training time and memory compared to the baseline models of EMCL and DRL, which achieves significant performance gains. The additional time overhead generated by our method mainly comes from the retrieval of additional knowledge from multi-grained knowledge stores. We can observe that compared to the base model, our method only increases the training time for each epoch by 80 seconds. In addition, we maintain multi-grained knowledge stores during training and inference, which incurs additional memory overhead since it is necessary to load the knowledge base into memory.

By comparing the results of closed and open-world settings, it can be observed that the time and memory costs of our KDProR are independent of the source of knowledge. This is mainly because the time and space complexity mainly depend on the computational complexity and parameter of the model itself.

Compared to the computational complexity of the model, the cost introduced by our knowledge-decoupling strategy can be almost negligible.

Table 7: Runtime analysis of different methods. We re-implement the EMCL [10] and DRL-WTI [21] methods in our experiment. $^{\top}$ is closed-set settings and denotes constructing knowledge stores on the training set using the fine-tuned CLIP upon the training set. ‡ denotes CoT knowledge.

	EMCL [10]	DRL [21]	KDProR$^{\top}$	KDProR‡
Training Time for a Epoch ($\times 10^3$s)	**1.85**	1.98	2.06	2.05
Memory (GB)	**18.8**	22.4	23.1	23.1
R@1	46.6	46.9	48.7	**49.6**

E More Discussions about KDProR

E.1 Further Clarification of Relevant Concepts

Open-world or open-set settings mean introducing data other than the training set and pre-trained CLIP for augmenting VTR. Our work aims to explore knowledge decoupling and knowledge enhancement strategies under "open-book examination". "open-book examination" means that we can still obtain data or features in the training set during the testing phase. KDProR can still obtain knowledge from knowledge stores during the testing phase to improve inference performance. Thus, the learning paradigm of KDProR belongs to this new paradigm of "open-book examination". In addition, although studies Cap4Video [24] and TEFAL [8] explore using auxiliary captions and additional audio to enhance VTR (open-world settings), they still rely on memorizing all knowledge during the training phase, which still belongs to the learning paradigm of "closed-book examination".

E.2 Comparison with Pre-Training Methods

Due to the fact that the image-language model CLIP [19] is pre-trained on a large number of image-text pairs, it may have a problem of insufficient understanding of spatio-temporal features [26]. To alleviate this issue, the latest studies [12, 26] explore pre-training techniques on video-text datasets to construct foundation models for the understanding of videos. CLIP-ViP [26] conducts secondary pre-training of CLIP on large-scale video-text datasets to improve the spatio-temporal understanding and inject open-set knowledge into VTR tasks. UMT [12] achieves efficient pre-training on a large number of video-language pairs. Compared to these methods, our KDProR has several appealing facets: (1) Our method avoids the high cost of video-text pre-training and supports the introduction of temporal knowledge that is beneficial for improving VTR

in a low-cost manner. (2) Our KDProR is compatible with these pre-training methods and can improve VTR by borrowing knowledge from pre-training foundation models, such as using pre-training knowledge from CLIP-ViP. (3) Our approach also supports the introduction of various types of open-world knowledge, such as entity-relation structured knowledge and auxiliary knowledge from ChatGPT [18].

E.3 Limitations and Further Work

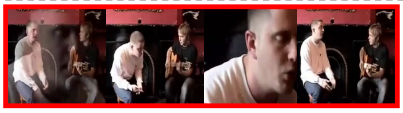
We explore a novel knowledge-decoupling paradigm for VTR. Our approach, KDProR, is capable of accommodating various interaction mechanisms and integrating diverse types of open-world knowledge. However, it necessitates the maintenance of knowledge decoupling stores throughout the training and testing phases, resulting in a slight increase in memory overhead.

Furthermore, our technique holds potential applicability across a range of visual-language tasks, such as visual question answering, by incorporating diverse open-world knowledge to enhance cross-modal comprehension. This area also constitutes our direction for future research.

F More Visualization Cases

We provide more cases of videos retrieved by CLIP4clip [15], DRL [21] and KDProR. As illustrated in Figure 1, KDProR successfully retrieves ground-truth videos in both closed- and open-set settings. This result further intuitively demonstrates the effectiveness of our method.

Query: a man with a guitar sings on a farm

Video	CLIP4clip	DRL	KDProR$^{\top}$	KDProR‡
	Rank 6	Rank 5	Rank 1	Rank 1
	Rank 1	Rank 3	Rank 4	Rank 3
	Rank 2	Rank 8	Rank 3	Rank 8
	Rank 3	Rank 2	Rank 5	Rank 6

Query: a football player with a football

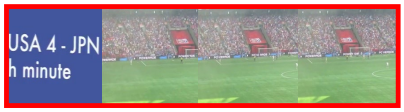
Video	CLIP4clip	DRL	KDProR$^{\top}$	KDProR‡
	Rank 5	Rank 3	Rank 1	Rank 1
	Rank 1	Rank 2	Rank 3	Rank 8
	Rank 2	Rank 8	Rank 7	Rank 4
	Rank 3	Rank 1	Rank 2	Rank 3

Fig. 1: The text-video results on the MSR-VTT 1K-A test set. $^{\top}$ and ‡ are close- and open-world settings respectively, whose meanings are consistent with those in Table 3. Ground-truth videos are surrounded by green borders, while incorrect videos are surrounded by red borders.

References

1. Bogolin, S.V., Croitoru, I., Jin, H., Liu, Y., Albanie, S.: Cross modal retrieval with querybank normalisation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5194–5205 (2022)
2. Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D.: Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. arXiv preprint arXiv:2109.04290 (2021)
3. Deng, C., Chen, Q., Qin, P., Chen, D., Wu, Q.: Prompt switch: Efficient clip adaptation for text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15648–15658 (2023)
4. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. arXiv preprint arXiv:2401.08281 (2024)
5. Fang, B., Liu, C., Zhou, Y., Yang, M., Song, Y., Li, F., Wang, W., Ji, X., Ouyang, W., et al.: Uatvr: Uncertainty-adaptive text-video retrieval. arXiv preprint arXiv:2301.06309 (2023)
6. Gao, Z., Liu, J., Chen, S., Chang, D., Zhang, H., Yuan, J.: Clip2tv: An empirical study on transformer-based methods for video-text retrieval. ArXiv **abs/2111.05610** (2021)
7. Huang, Y., Tang, J., Chen, Z., Zhang, R., Zhang, X., Chen, W., Zhao, Z., Lv, T., Hu, Z., Zhang, W.: Structure-clip: Enhance multi-modal language representations with structure knowledge. arXiv preprint arXiv:2305.06152 (2023)
8. Ibrahimi, S., Sun, X., Wang, P., Garg, A., Sanan, A., Omar, M.: Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12020–12030 (2023)
9. Jensen, J.L.W.V.: Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Mathematica **30**, 175–193 (1906)
10. Jin, P., Huang, J., Liu, F., Wu, X., Ge, S., Song, G., Clifton, D., Chen, J.: Expectation-maximization contrastive learning for compact video-and-language representations. Advances in Neural Information Processing Systems **35**, 30291–30306 (2022)
11. Jin, P., Li, H., Cheng, Z.L., Li, K., Ji, X., Liu, C., ming Yuan, L., Chen, J.: Diffusionret: Generative text-video retrieval with diffusion model. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 2470–2481 (2023)
12. Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., Qiao, Y.: Unmasked teacher: Towards training-efficient video foundation models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 19891–19903 (2023)
13. Liu, R., Huang, J., Li, G., Feng, J., Wu, X., Li, T.H.: Revisiting temporal modeling for clip-based image-to-video knowledge transferring. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6555–6564 (2023)
14. Liu, Y., Xiong, P., Xu, L., Cao, S., Jin, Q.: Ts2-net: Token shift and selection transformer for text-video retrieval. ArXiv **abs/2207.07852** (2022)
15. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
16. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. Proceedings of the 30th ACM International Conference on Multimedia (2022)

17. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.J.: Training language models to follow instructions with human feedback. ArXiv **abs/2203.02155** (2022)
18. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
20. Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y.G., Yuan, L.: Omnivl: One foundation model for image-language and video-language tasks. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 5696–5710. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/259a5df46308d60f8454bd4adcc3b462-Paper-Conference.pdf
21. Wang, Q., Zhang, Y., Zheng, Y., Pan, P., Hua, X.: Disentangled representation learning for text-video retrieval. ArXiv **abs/2203.07111** (2022)
22. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
23. Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., Ma, W.Y.: Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6602–6611 (2019)
24. Wu, W., Luo, H., Fang, B., Wang, J., Ouyang, W.: Cap4video: What can auxiliary captions do for text-video retrieval? 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10704–10713 (2022)
25. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5026–5035 (2021)
26. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language alignment. In: International Conference on Learning Representations (2023)
27. Zhao, S., Zhu, L., Wang, X., Yang, Y.: Centerclip: Token clustering for efficient text-video retrieval. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (2022)