KDProR: A Knowledge-Decoupling Probabilistic Framework for Video-Text Retrieval

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

Abstract. Existing video-text retrieval methods predominantly focus on designing diverse cross-modal interaction mechanisms between captions and videos. However, those approaches diverge from human learning paradigms, where humans possess the capability to seek and associate knowledge from an open set, rather than rote memorizing all text-video instances. Motivated by this, we attempt to decouple knowledge from retrieval models through multi-grained knowledge stores and identify two significant benefits of our knowledge-decoupling strategy: (1) it ensures a harmonious balance between knowledge memorization and retrieval optimization, thereby improving retrieval performance; and (2) it can promote incorporating diverse open-world knowledge to augment video-text retrieval. To efficiently integrate information from knowledge stores, we further introduce a novel retrieval framework termed KDProR, which utilizes our proposed Expectation-Knowledge-Maximization (EKM) algorithm for optimization. Specifically, in E-step, KDProR obtains relevant contextual semantics from knowledge stores and achieves efficient knowledge injection through interpolation and alignment correction. During the K-step, KDProR calculates the knowledge KNN distribution by indexing the top-K acquired knowledge to refine the retrieval distribution, and in M-step, KDProR optimizes the retrieval model by maximizing the likelihood of the objective. Extensive experiments on various benchmarks prove that KDProR significantly outperforms previous state-of-the-art methods across all metrics. Remarkably, KDProR can uniformly and efficiently incorporate diverse open-world knowledge and is compatible with different interaction mechanisms and architectures.

Keywords: Video Retrieval · Knowledge Decoupling · Open World

1 Introduction

Video-text retrieval (VTR) is a significant and challenging task in cross-modal interaction [1, 12, 31, 42], aiming to align video (text) candidates with text (video) queries to identify the most relevant instances accurately [61]. The standard

* These authors contributed equally.
† Corresponding author.

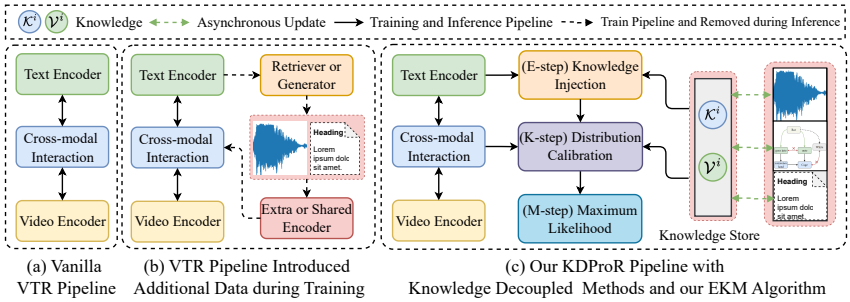

Fig. 1: The illustration of different VTR paradigms. (a) denotes vanilla VTR methods; (b) introduces additional data only during training. (c) achieves "open examination" with knowledge decoupling, thereby enhancing VTR in both closed and open sets.

paradigm of VTR for addressing cross-modal matching involves aligning video-text features acquired through visual and textual encoders via contrastive learning. With the advent of large-scale image-language pre-trained models [20, 44, 60, 62], recent VTR methods [3, 13, 35] have realized substantial advancements in retrieval performance by leveraging pre-trained image-language models, such as CLIP [44]. Benefiting from powerful CLIP, most of these methods focus on devising various cross-modal interaction strategies to enhance visual-text alignment, including cross-attention [13], multi-grained interaction [36], disentangled representation learning [48], etc, and have achieved considerable success.

However, if we conceptualize the training set as a "book" intuitively, this training-test paradigm for VTR can be likened to "video-text memorization" and "closed-book VTR examination", which diverges from the human learning process [37, 65]. Humans possess the capability to seek and associate knowledge from an open set, rather than relying on rote memorization of all text-video instances [47]. This "closed-book examination" approach may be further hindered by dense memorization overhead [37] and challenges in memorizing difficult samples [9, 11], echoing the proverb: "The palest ink is better than the best memory." Furthermore, previous open-set fine-tuning [18, 52] and pre-training [27, 59] methods have shown that incorporating auxiliary data from the open world can improve VTR performance. These insights encourage us to explore a novel VTR paradigm: "open-set examination", which involves constructing knowledge stores from the closed training set or open-world sources to decouple knowledge. As shown in Figure 1, this strategy introduces relevant knowledge from stores as reinforcement signals to help the model strike a balance between generalization and memory. However, this also brings two new challenges: (1) How to decouple knowledge from models and unify closed- and open-world knowledge; (2) How to efficiently incorporate decoupled knowledge to improve retrieval performance.

In this work, we propose a novel **k**nowledge-**d**ecoupling **pro**babilistic framework for VT**R** (KDProR), tackling the challenges with the following strategies:

(1) **KDProR utilizes multi-grained stores to decouple knowledge and standardize closed- and open-world knowledge sources.** We con-

struct multi-scale knowledge stores composed of local and global neural representations to decouple knowledge. This approach allows rare patterns to be memorized explicitly, rather than implicitly in model parameters. In addition, knowledge stores can be updated asynchronously and endowed with knowledge from closed- or open-world sources. In this work, we explore three different sources of additional knowledge, including spatio-temporal knowledge from pre-trained models, entity-relation structured knowledge and outputs of large language models (LLMs, e.g., GPT [40]) prompted by chain-of-thoughts (CoT) [50]. Moreover, we identify two core advantages of the knowledge decoupling strategy: it (a) ensures a harmonious balance between memorization and retrieval optimization; and (b) opens up a unified interface for injecting diverse open-world knowledge.

(2) **KDProR utilizes a novel Expectation-Knowledge-Maximization algorithm to achieve efficient knowledge injection and retrieval optimization.** We extend the traditional Expectation-Maximization (EM) algorithm [5] to the Expectation-Knowledge-Maximization (EKM) algorithm for optimizing KDProR. Specifically, in the E-step, KDProR obtains relevant contextual semantics and achieves efficient knowledge injection through interpolation and alignment correction. In the K-step, we calculate the k-nearest neighbor (KNN) distribution by indexing the top-K knowledge obtained to calibrate the original retrieval distribution. And in the M-step, we optimize the model by maximizing the likelihood of the target. Additionally, we provide a theoretical analysis of our KDProR framework from a probabilistic perspective in Section 3.7.

Extensive experiments on four benchmarks, i.e., MSR-VTT [55], DiDeMo [16], LSMDC [45] and ActivityNet [15], prove that KDProR significantly outperforms previous state-of-the-art methods in both closed and open-world settings. In summary, our KDProR has several appealing facets: (1) **Effectiveness**: It can significantly improve retrieval performance under both closed-set and open-set settings. (2) **Universality**: It can be used to introduce various open-world knowledge, while being compatible with various interaction modules and pre-trained foundation models (e.g., [27, 59]). (3) **Explainability**: The effectiveness and convergence of retrieval optimization via our EKM algorithm are theoretically guaranteed. The main contributions of this paper are presented as follows:

(1) We explore a novel knowledge-decoupling paradigm for VTR and construct multi-grained knowledge stores to uniformly introduce knowledge from closed sets and diverse open-world sources. (2) We propose a novel VTR framework termed KDProR, which utilizes a principled EKM algorithm to inject various knowledge into the model to improve retrieval performance. Theoretical analysis provides an underlying insight into the effectiveness and convergence of KDProR. (3) Experiments conducted on four benchmarks show that KDProR achieves new state-of-the-art results under both closed- and open-world settings.

2 Related work

Feature Representation for Video-Text Retrieval. Recent studies [1, 12, 31, 42] on VTR employ a bi-encoder architecture to extract text and video rep-

resentation. The selection of text and visual encoders changes with the development of research on feature extraction upon NLP and CV, e.g., Word2Vec [38] and BERT-based models [7, 32] for textual representation, and CNN-based [14] and ViT-based models [8, 63, 64] for visual tasks. These researches use independently pre-trained dual encoders for visual-textual semantic extraction [1, 12, 42, 57]. Recently, CLIP [43] pre-trained on a large-scale dataset with 400 million pairs has achieved excellent capability of cross-modal representation. And CLIP-based VTR methods [3, 13, 35, 53] have rapidly developed and achieved great success. Our KDProR also benefits from existing pre-training, while exploring a novel perspective for enhancing representation, i.e., knowledge decoupling.

Interaction Mechanism for Video-Text Retrieval. Many VTR methods explore different cross-modal interaction mechanisms for cosine similarity [35] to improve retrieval performance. Some studies [35] seek to enhance the spatiotemporal understanding ability of image encoders through LSTM [17] or temporal Transformers. MoE-based [3] and Cross-Transformers-based [13, 54] methods significantly improve the cross-modal interaction. [25, 26, 29, 36, 49, 56] explore hierarchical and multi-grained interactions, respectively. DRL [48] proposes disentangled representation learning for cross-modal interaction. Our KDProR is compatible with these cross-modal interaction strategies.

Open-World Knowledge for Video-Text Retrieval. The previous methods mainly introduce open-world knowledge in the pre-training and fine-tuning stages to enhance VTR performance. CLIP-ViP [59] conducts secondary pre-training of CLIP on large-scale video-text datasets, and UMT [27] achieves efficient pre-training to improve the spatio-temporal understanding and inject open-set knowledge into VTR tasks. TEFAL [18] utilizes additional audio features to enhance VTR retrieval performance and Cap4Video [52] utilizes auxiliary captions obtained by Web Search or generated by LLMs to enhance cross-modal interaction. KDProR is compatible with these pre-training methods, and we do not specify the type of open-world knowledge to be introduced, but instead achieve universal open-set knowledge injection through knowledge decoupling.

3 Methods

3.1 Settings and Feature Extraction.

Following previous work [35, 48], we utilize the dual-encoder architecture of CLIP [44] to achieve video and text feature extraction on training set $\mathcal{D} = \{x_i\}_i^N = \{(T_i, V_i)\}_i^N$, where T_i and V_i represent text and video instances respectively, and N is the size of the training set. Given a text query T_i, we initially add identifiers `[CLS]` and `[SEP]` to the sentence, and leverage the text encoder of CLIP to encode the text representation: $t_i = \text{TextEncoder}(T_i)$. For each video V_i, we uniformly select n_v frames $[F_1, \cdots, F_{n_v}]$ as keyframes, and employ the transformer-based encoder of CLIP to extract sequential features $v_i = \text{VideoEncoder}([F_1, \cdots, F_{n_v}])$.

Extensive work explores various cross-modal interaction strategies to improve retrieval performance [30,35,48]. Our work focuses on decoupling knowledge from

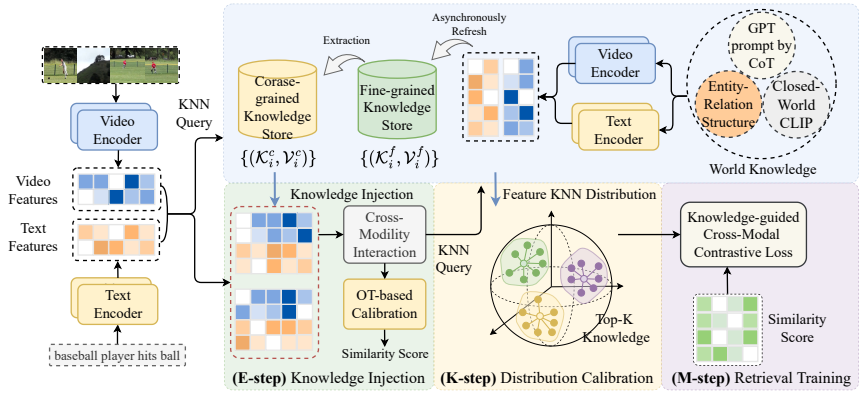

Fig. 2: The illustration of the proposed framework, consisting of (1) constructing a multi-grained knowledge store using open-world knowledge or closed-world knowledge (training dataset); (2) EKM algorithm that sequentially implements knowledge injection, distribution calibration, and target likelihood maximization.

models and exploring the impact of different knowledge sources on retrieval performance, rather than pre-training or interaction strategies. Thus we adopt the weighted token-wise interaction module of DRL [48] to achieve cross-modal interaction. Note that KDProR is also compatible with other interaction strategies.

3.2 Decoupling Knowledge Store

Our KDProR involves augmenting CLIP-based retrieval models with semantic information from decoupled knowledge stores. Thus, the first step of our methods is to build a multi-grained knowledge store. As illustrated in Figure 2, our stores contain fine-grained and coarse-grained caches, where each key and value (text-video) are representations of knowledge.

Fine-grained Knowledge Store. The fine-grained knowledge store can be represented as $\mathcal{S}^f = \{\mathcal{S}_i^f\}_{i=1}^{K^f} = \{(\mathcal{K}_i^f, \mathcal{V}_i^f)\}_{i=1}^{K^f}$, where $\mathcal{S}_i^f = (\mathcal{K}_i^f, \mathcal{V}_i^f)$ denotes the i-th key-value pair in the fine-grained knowledge store, and $\mathcal{K}_i^f$ is the size of our fine-grained knowledge store. Under the closed world setting, we utilize text and visual encoders of CLIP to obtain the feature key-value pairs on the training dataset as $\mathcal{K}_i^f = t_i$ and $\mathcal{V}_i^f = v_i$.

Coarse-grained Knowledge Store. The coarse-grained knowledge store is designed to encapsulate global semantics. It equips retrieval models with global, coarse-grained insights from the knowledge store, enhancing the comprehensive knowledge understanding. Similarly, the coarse-grained storage can be represented as: $\mathcal{S}^c = \{\mathcal{S}_i^c\}_{i=1}^{K^c} = \{(\mathcal{K}_i^c, \mathcal{V}_i^c)\}_{i=1}^{K^c}$, where $\mathcal{S}_i^c = (\mathcal{K}_i^c, \mathcal{V}_i^c)$ denotes the i-th key-value pair in the coarse-grained knowledge store, and $\mathcal{K}_i^c$ is the size of our coarse-grained knowledge store. We adopt the K-mean strategy to extract knowledge from fine-grained storage into more advanced representations:

$$\mathcal{K}_i^c = \phi\left(c_i^t\right), \forall c_i^t \in \mathcal{C}^t, \qquad \mathcal{V}_i^c = \phi\left(c_i^v\right), \forall c_i^v \in \mathcal{C}^v, \tag{1}$$

where, $\mathcal{C}^t = \{c_1^t, \cdots, c_{K^c}^t\}$ denotes the cluster set obtained by K-Mean over fine-grained knowledge $\{\mathcal{K}_i^f\}$, $\mathcal{C}^v = \{c_1^v, \cdots, c_{K^c}^v\}$ represents the cluster set obtained by K-Mean over $\{\mathcal{V}_i^f\}$, and $\phi(\cdot)$ represents the Max-Pooling aggregate function.

Note that our framework is universal, and we will introduce more applications for constructing diverse knowledge stores in Section 3.6.

3.3 E-step for Multi-Grained Knowledge Injection

Knowledge Interpolation. In E-step, KDProR initially attempts to inject fine-grained and coarse-grained knowledge into the model by knowledge interpolation of features. We treat with t_i and v_i as anchors to retrieve the fine-grained knowledge store, and obtain two sets of top-K similar vector pairs $\mathcal{N}_{t_i}^f$ and $\mathcal{N}_{v_i}^f$, respectively. $\mathcal{N}_{t_i}^f = \{(\mathcal{K}_{t_i}^f, \mathcal{V}_{t_i}^f)\}$ and $\mathcal{N}_{v_i}^f = \{(\mathcal{K}_{v_i}^f, \mathcal{V}_{v_i}^f)\}$ are two sets of key-value knowledge pairs, where $\mathcal{K}_{t_i}^f$ and $\mathcal{V}_{v_i}^f$ are the nearest neighbors of t_i and v_i, respectively. After the model retrieves the top-K candidates for each t_i and v_i, their corresponding representation $\mathcal{K}_{t_i}^f$ and $\mathcal{V}_{v_i}^f$ obtained from knowledge-store will be incorporated into original features t_i and v_i to act as demonstration learning. Specifically, we adopt the nearest neighbor sets to inject decoupled knowledge into text and video features as:

$$t_i^f = \rho t_i + (1-\rho)\frac{\sum_{(\mathcal{K},\mathcal{V})\in\mathcal{N}_{t_i}^f}\mathcal{K}}{|\mathcal{N}_{t_i}^f|}, \quad v_i^f = \rho v_i + (1-\rho)\frac{\sum_{(\mathcal{K},\mathcal{V})\in\mathcal{N}_{v_i}^f}\mathcal{V}}{|\mathcal{N}_{v_i}^f|}, \tag{2}$$

where, ρ is a hyperparameter used to express the interpolation ratio. Subsequently, we utilize the weighted token-wise interaction module [48] to enhance textual and visual representations injected with additional knowledge.

However, the textual and visual knowledge within the top-K sets, obtained by t_i and v_i respectively, may not correspond on a one-to-one basis due to their derivation from two independent top-K operations. To mitigate the issue of mismatched video-text features during knowledge injection, we utilize the optimal transport (OT) algorithm to facilitate alignment correction within a batch.

Alignment Correction via Optimal Transport. We first define $\langle \mathbf{Q}, \mathbf{S}^f \rangle = \mathrm{tr}(\mathbf{Q}^\top \mathbf{S}^f) \in \mathbb{R}^{B\times B}$ as the similarity matrix calculated from the updated text and video pairs $\{(t_i^f, v_i^f)\}_{i=1}^B$ in a batch. $\mathbf{Q} \in \mathbb{R}^{B\times B}$ represents the corresponding transport assignment matrix, where $\mathbf{Q}[i,j]$ denotes the probabilities of aligning t_i^f with v_i^f. The optimal transport aims to establish flexible alignment between videos and captions by maximizing global text-video similarity $\mathrm{tr}(\mathbf{Q}^\top \mathbf{S}^f)$. The objective of optimal transport can be formulated as:

$$\max_{\mathbf{Q}\in\mathcal{Q}} \langle \mathbf{Q}, \mathbf{S}^f \rangle + \varepsilon H(\mathbf{Q}) \quad \text{s.t. } \mathcal{Q} = \left\{ \mathbf{Q} \in \mathbb{R}^{B\times B} \mid \mathbf{Q}\mathbf{1}_B = \boldsymbol{\mu}, \mathbf{Q}^\top \mathbf{1}_B = \boldsymbol{\nu} \right\} \tag{3}$$

where, $\boldsymbol{\mu} \in \mathbb{R}^B$ and $\boldsymbol{\nu} \in \mathbb{R}^B$ indicate the relative importance of each video and caption, $\mathbf{1}_B$ represents the vector of ones in dimension B, $H(\mathbf{Q})$ denotes a entropy regular term [4] and ε controls its smoothness. Following [46], we initialize $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ to a uniform distribution $\mathbf{1}_B$. The optimal transport problem

in Eq. 3 can be solved using the Sinkhorn-Knopp algorithm [46] to obtain the optimal assignment scheme $\mathbf{Q}^*$. Based on this, we can obtain the similarity matrix after adding alignment correction $\mathbf{Q}^*$ as follows:

$$\mathbf{S}^{f^*} = ((1-\beta)\mathbf{I} + \beta\mathbf{Q}^*)\,\mathbf{S}^f, \tag{4}$$

where, $\mathbf{I}$ is the identity matrix and β is a hyperparameter.

We can implement the coarse-grained knowledge injection and obtain the realigned similarity matrix $\mathbf{S}^{c*}$ in a similar way.

3.4 K-step for Retrieval Distribution Calibration

Existing studies [23, 41] have shown that non-parametric KNN classification or nearest-neighbor interpolation methods have extraordinary potential in improving robustness and generalization. It is intuitively to leverage the KNN's retrieval results as the prior knowledge to guide the parameters training for video-text pairs that are difficult to establish deep correlations. In practice, we aggregate the prediction of KNN into the original probability distribution of video-text retrieval to achieve the multi-grained calibration of retrieval distribution.

Fine-grained Distribution Calibration. We utilize the set of k-nearest neighbors $\mathcal{N}_{t_i}^f$ and $\mathcal{N}_{v_i}^f$ retrieved by querying the open-book knowledge store in E-step as fine-grained top-K knowledge. Subsequently, we calculate the KNN distribution according to the softmax of the similarity and aggregation probability mass of each video-text pair in the retrieved target as:

$$p_{knn}^f(v_i \mid t_i) = \frac{\sum_{(\mathcal{K},\mathcal{V})\in N_{t_i}^f} \varphi(v_i,\mathcal{V})\, e^{d(v_i,\mathcal{V})}}{\sum_{(\mathcal{K},\mathcal{V})\in N_{t_i}^f} \varphi(v_i,\mathcal{V})}, \quad \varphi(v_i,\mathcal{V}) = \begin{cases} \mathbf{1}, & \mathcal{V} \in N_{v_i}^f \\ 0, & \mathcal{V} \notin N_{v_i}^f \end{cases}, \tag{5}$$

where, $d(\cdot,\cdot) \in [-1,1]$ denotes the cosine similarity function, and $\varphi(v_i,\mathcal{V})$ is an indicator function used to determine whether $\mathcal{V}$ is within the top-K nearest neighbor set $N_{v_i}^f$ of the v_i.

We can further obtain the KNN probability distribution $p_{knn}^f(t_i \mid v_i)$ for video-to-text retrieval in a completely symmetrical manner.

Coarse-grained Distribution Calibration. Consistent with Eq. 5, we employ the KNN algorithm to obtain top-K knowledge, utilizing closed-world or open-world knowledge from the coarse-grained knowledge store to refine the probability distribution function pertinent to the text-to-video retrieval. Nevertheless, since the coarse-grained knowledge store is derived from the distillation of fine-grained knowledge, it is impracticable to apply indicator functions (*cf.* Eq. 5) to ascertain the presence of an inclusion relation between textual features and their nearest neighbor sets. To tackle this challenge, we introduce a novel indicator function, denoted as $\phi(\cdot,\cdot)$, parameterized by a threshold α. $\phi(\cdot,\cdot)$ is designed to evaluate whether $\mathcal{V}$ resides within the top-K nearest neighbor set $N_{v_i}^c$ of the v_i via the cosine similarity function $d(\cdot,\cdot)$, and is formulated as:

$$\phi(v_i,\mathcal{V}) = \begin{cases} \mathbf{1}, & d(\mathcal{V},\mathcal{V}^c) < \alpha, \exists(\mathcal{K}^c,\mathcal{V}^c) \in N_{v_i}^c \\ 0, & otherwises \end{cases}. \tag{6}$$

Based on this, we can obtain a closed-form expression for the KNN distribution used for coarse-grained distribution calibration as follows:

$$p_{knn}^{c}\left(v_i \mid t_i\right)=\frac{\sum_{(\mathcal{K},\mathcal{V})\in N_{t_i}^{c}} \phi\left(v_i,\mathcal{V}\right) e^{d(v_i,\mathcal{V})}}{\sum_{(\mathcal{K},\mathcal{V})\in N_{t_i}^{c}} \phi\left(v_i,\mathcal{V}\right)}. \tag{7}$$

We can further obtain the KNN probability distribution $p_{knn}^{c}\left(t_i \mid v_i\right)$ for coarse-grained video-to-text retrieval in a completely symmetrical manner.

3.5 M-step for Unified Retrieval Optimization

In the M-step, we optimize the whole model by maximizing the retrieval likelihood objective over the realigned similarity matrix. Specifically, we utilize the re-aligned multi-grained similarity matrix in Eq. 4 obtained from E-step to calculate the InfoNCE loss:

$$\begin{aligned}
\mathcal{L}_{t2v} &= -\frac{1}{B}\sum_{i}^{B}\left(\log\frac{\exp\left(\mathbf{S}^{f*}_{ii}\right)}{\sum_{j=1}^{B}\exp\left(\mathbf{S}^{f*}_{ij}\right)}+\log\frac{\exp\left(\mathbf{S}^{c*}_{ii}\right)}{\sum_{j=1}^{B}\exp\left(\mathbf{S}^{c*}_{ij}\right)}\right),\\
\mathcal{L}_{v2t} &= -\frac{1}{B}\sum_{i}^{B}\left(\log\frac{\exp\left(\mathbf{S}^{f*}_{ii}\right)}{\sum_{j=1}^{B}\exp\left(\mathbf{S}^{f*}_{ji}\right)}+\log\frac{\exp\left(\mathbf{S}^{c*}_{ii}\right)}{\sum_{j=1}^{B}\exp\left(\mathbf{S}^{c*}_{ji}\right)}\right),
\end{aligned} \tag{8}$$

where B is the batch size. Subsequently, we propose utilizing the KNN distribution ($cf.$ Eq. 5 and 7) to guide the training process. The KNN calibrator reweights the InfoNCE loss by adjusting the relative loss for the correctly-matched or mismatched video-text instances identified by KNN, respectively. We focus on exploiting the results of KNN distribution for calibrating the training of retrieval models. Specifically, we apply the negative log-likelihood of p_{knn}^{f} in Eq 5 and p_{knn}^{c} in Eq. 7 as the calibration factor:

$$\begin{aligned}
F_{t2v}^{f} &= -\log p_{knn}^{f}\left(v_i \mid t_i\right), \quad F_{v2t}^{f} = -\log p_{knn}^{f}\left(t_i \mid v_i\right);\\
F_{t2v}^{c} &= -\log p_{knn}^{c}\left(v_i \mid t_i\right), \quad F_{v2t}^{c} = -\log p_{knn}^{c}\left(t_i \mid v_i\right).
\end{aligned} \tag{9}$$

Finally, we can optimize the retrieval model by maximizing the following **knowledge-guided cross-modal contrastive loss** as:

$$\mathcal{L}_{v2t}^{*}=(1+\lambda_1 F_{v2t}^{f}+\lambda_2 F_{v2t}^{c})\mathcal{L}_{v2t};\quad \mathcal{L}_{t2v}^{*}=(1+\lambda_1 F_{t2v}^{f}+\lambda_2 F_{t2v}^{c})\mathcal{L}_{t2v}, \tag{10}$$

where λ_1 and λ_2 are hyper-parameters to determine the proportion of each loss term. Hence, in the M-step, the complete training objective of KDProR can be formulated as $\mathcal{L}=\mathcal{L}_{v2t}^{*}+\mathcal{L}_{t2v}^{*}$.

3.6 Applications of Our KDProR Framework

In this section, as illustrated in Figure 3, we present three distinct applications for constructing open-world knowledge stores, including pre-trained knowledge, entity-relation structured knowledge, and CoT knowledge within LLMs.

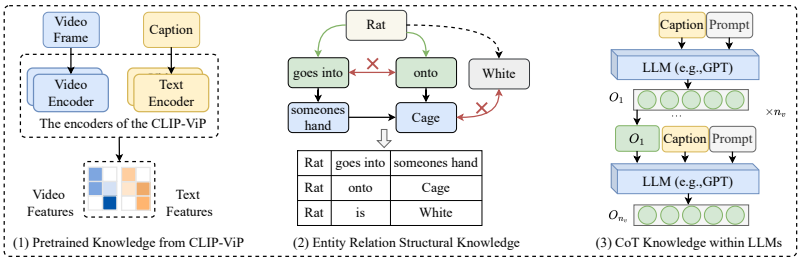

Fig. 3: The illustration of proposed open-world knowledge stores, including (1) video-text pretraining knowledge, (2) entity-relation structure knowledge, and (3) knowledge within LLMs prompted by the CoT technique.

Video-Text Pretraining. As shown in Figure 3, we utilize CLIP-ViP [59], which is further pre-trained on HD-VILA-100M [58] dataset, to extract pre-training video-text knowledge for constructing our knowledge stores.
Entity-Relation Structure. We attempt to parse captions to scene graphs $\mathcal{G}$ through the Scene Graph Parser [51], 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.
CoT Generation of LLMs. In this work, we attempt to use a strategy similar to CoT [50] to generate auxiliary captions for each video frame by prompting LLMs. Subsequently, we utilize CLIP to extract neural representations of auxiliary captions and corresponding video frames to construct knowledge stores.

3.7 Theoretical Analysis

From a probabilistic perspective, the optimization goal of the VTR model Θ trained on dataset $\mathcal{D} = \{x_i\}_i^N = \{(T_i, V_i)\}_i^N$ is to maximize the log-likelihood $\mathcal{L}(\Theta)$ for text-video pairs. We consider the knowledge introduced from fine-grained and coarse-grained stores as two related latent variables z_i^f and z_i^c. In a closed world, the posterior distribution of z_i^f and z_i^c can be represented as:

$$Q(z_i^f) = p(z_i^f; x_i, \Theta), \quad Q(z_i^c) = p(z_i^c; z_i^f, x_i, \Theta). \tag{11}$$

Thus we can rewrite the objective function of the retrieval model Θ according to Jensen's inequality [19] as follows:

$$\Theta^* = \arg\max \sum_i \sum_{z_i^f, z_i^c} Q(z_i^f) Q(z_i^c) \log p\left(x_i, z_i^f, z_i^c; \Theta\right). \tag{12}$$

To sum up, 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. 12 with known $Q(z_i^f)$ and $Q(z_i^c)$, and satisfies the Theorem 1:

Table 1: Results on MSR-VTT without any post-processing (e.g., [3] and [2]). $^{+}$ 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 in Section 3.6, respectively.

Backbone	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↓
Non-CLIP	CE [31]	20.9	48.8	62.4	6.0	28.2	20.6	50.3	64.0	5.3	25.1
	MMT [12]	26.6	57.1	69.6	4.0	24.0	27.0	57.5	69.7	3.7	21.3
	Support [42]	27.4	56.3	67.7	3.0	-	26.6	55.1	67.5	3.0	-
	Frozen [1]	31.0	59.6	70.5	3.0	-	-	-	-	-	-
CLIP ViT-B/32	CLIP4clip [35]	44.5	71.4	81.6	**2.0**	15.3	42.7	70.9	80.6	**2.0**	11.6
	X-pool [13]	46.9	72.8	82.2	**2.0**	14.3	-	-	-	-	-
	EMCL [21]	46.8	73.1	83.1	**2.0**	-	46.5	73.5	83.5	**2.0**	-
	TS2-Net [33]	47.0	74.5	83.8	**2.0**	13.0	45.3	74.1	83.7	**2.0**	9.2
	DRL [48]	47.4	74.6	83.8	**2.0**	-	45.3	73.9	83.3	**2.0**	-
	STAN [28]	46.9	72.8	82.8	**2.0**	-	-	-	-	**-**	-
	UATVR [10]	47.5	73.9	83.5	**2.0**	12.9	46.9	73.8	83.8	**2.0**	8.6
	PromptSwitch [6]	47.8	73.9	82.2	-	14.1	46.0	74.3	**84.8**	-	8.5
	KDProR^{+}	48.4	74.4	84.2	**2.0**	12.2	47.1	**74.6**	84.4	**2.0**	8.9
	KDProR$^{\top}$	48.7	74.4	84.1	**2.0**	12.0	47.2	**74.6**	84.4	**2.0**	8.7
	KDProR*	49.2	74.2	**84.4**	**2.0**	12.0	47.1	74.0	84.1	**2.0**	8.9
	KDProR†	49.0	74.6	84.3	**2.0**	12.0	47.3	74.3	83.9	**2.0**	8.7
	KDProR‡	**49.6**	**75.1**	**84.4**	**2.0**	**11.6**	**48.2**	74.2	84.2	**2.0**	**8.1**

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.

4 Experiments

4.1 Experiments Settings

Datasets. We conduct experiments on four benchmarks for video-text retrieval tasks including: MSR-VTT [55], DiDeMo [16], LSMDC [45] and ActivityNet [15]. MSR-VTT [55] comprises 10,000 videos with 20 captions each. We train on 9,000 videos and their captions, testing on 1k-A with 1,000 video-text pairs. DiDeMo [16] provides 10,000 videos annotated with 40,000 sentences. We concatenate all descriptions of a video into one paragraph for evaluation. LSMDC [45] includes 118,081 movie-extracted videos, with our dataset partitioned into 109,673 for training, 7,408 for validation, and 1,000 for testing. ActivityNet [15] includes 20,000 YouTube videos. Following [35], we concatenate all captions similar to DiDeMo and evaluate on the 'val1' split.

Table 2: Text-to-Video retrieval results on DiDeMo, ActivityNet and LSMDC datasets without any post-processing (e.g., [3] and [2]). $^{+}$ and $^{\top}$ are closed-set settings and *, † and ‡ are open-world settings, whose meanings are all consistent with those in Table 1.

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↓
Text-to-Video Retrieval using CLIP (ViT-B/32)												
CLIP4clip [35]	43.4	70.2	80.6	2.0	40.5	72.4	83.6	**2.0**	22.6	41.0	49.1	11.0
X-CLIP [36]	45.2	74.0	-	-	44.3	74.1	-	-	23.3	43.0	-	-
TS2-Net [33]	41.8	71.6	82.0	2.0	41.0	73.6	84.5	**2.0**	23.4	42.3	50.9	9.0
DRL [48]	47.9	73.8	82.7	2.0	44.2	74.5	86.1	**2.0**	24.9	45.7	55.3	7.0
PromptSwitch [10]	-	-	-	-	-	-	-	-	23.1	41.7	50.5	-
UATVR [10]	43.1	71.8	82.3	2.0	-	-	-	-	-	-	-	-
DiffusionRet [22]	46.7	74.7	82.7	2.0	45.8	75.6	86.3	**2.0**	24.4	43.1	54.3	8.0
KDProR^{+}	48.8	76.1	84.1	2.0	45.8	75.4	85.0	**2.0**	26.1	45.9	57.3	7.0
KDProR$^{\top}$	49.2	75.3	83.6	2.0	46.4	76.2	85.7	**2.0**	26.5	47.8	57.8	6.0
KDProR*	51.7	78.3	84.9	**1.0**	48.3	76.8	86.1	**2.0**	26.9	50.2	58.3	**5.0**
KDProR†	50.5	76.3	**85.1**	**1.0**	47.5	76.1	85.7	**2.0**	27.6	**50.6**	58.1	**5.0**
KDProR‡	**53.2**	**79.2**	85.0	**1.0**	**49.1**	**77.9**	**86.8**	**2.0**	**28.2**	**50.6**	**59.2**	**5.0**

Metrics. Following [35, 48], we utilize recall at rank K (R@K), median rank (MdR) and mean rank (MnR) as metrics to validate the effectiveness of KDProR.

Implementation Details. Following [48], we adopt CLIP's ViT-B/32 as the visual encoder and fine-tune the model with 5 epochs on all training sets. We adopt Adam [24] as the optimizer with a cosine warm-up method [34]. Following [35, 48], the frame and caption length are 12 and 32 for MSR-VTT and LSMDC, 64 and 64 for ActivityNet and DiDeMo. The initial learning rate for vision and text encoder is set to 10^{-7} and other modules are 10^{-4}. The hyper-parameter ρ in Eq. 2 is set to 0.2 on MSR-VTT and LSMDC, 0.1 on ActivityNet and DiDeMo for optimal performance. β in Eq. 4 are set to 0.2, and λ_1 and λ_2 in Eq. 9 are set to 0.2 and 0.3, respectively. We use grid search to obtain the optimal parameters. We uniformly set the top-5 relevant knowledge for KNN retrieval, and the size of the coarse-grained knowledge store K^c is set to 256. We adopt the `gpt-3.5-turbo-1106` [39] to obtain the CoT knowledge in Section 3.6. All experiments are conducted on 2 Tesla A100 GPUs (80G) and 8 RTX 4090 GPUs.

4.2 Main Results

Comparisons to State-of-the-arts. In this section, we compare KDProR with state-of-the-art methods on four datasets, as results shown in Table 1 and 2. Analysis of Table 1 and 2 yields several insights:

(1) Our KDProR, employing multi-grained knowledge stores constructed from vanilla and fine-tuned CLIP, can achieve performance gains of 1% and 1.3% R@1 on MSR-VTT, 1.6% and 2.2% R@1 on ActivityNet compared to DRL [48], respectively. This indicates that KDProR can achieve a harmonious balance between memorization and optimization, thereby improving performance.

Table 3: Comparisons to Other Baseline Methods in Text-to-Video Retrieval. ‡ denotes the utilization of open-world CoT knowledge in Section 3.6.

Baseline Methods	Knowledge Source	MSR-VTT			LSMDC			DiDeMo		
		R@1↑	R@5↑	MnR↓	R@1↑	R@5↑	MnR↓	R@1↑	R@5↑	MnR↓
Data Augmentation	Captions from GPT	47.3	74.7	12.1	26.4	46.9	43.9	49.7	76.3	14.2
Feature Concatenate	Pretrained CLIP-ViP	47.8	74.5	11.9	26.1	46.1	44.5	50.3	76.3	13.4
Cap4Video [52]	Captions from GPT	49.3	74.3	12.0	-	-	-	52.0	**79.4**	10.5
TEFAL [18]	Audio Signal	49.4	**75.9**	12.0	26.8	46.1	44.4	-	-	-
Our KDProR‡	Captions from GPT	**49.6**	75.1	**11.6**	**28.2**	**50.6**	**34.9**	**53.2**	79.2	**10.2**

Table 4: Text-to-Video retrieval results on MSR-VTT using our re-implemented CLIP4clip [35] and EMCL [21] as base models. $^{+}$ and $^{\top}$ are closed-set settings, *, † and ‡ are open-world settings, whose meanings are consistent with those in Table 1.

Method	CLIP4clip [35] as Base Model					EMCL [21] as Base Model				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Base Model	43.6	70.6	81.0	**2.0**	16.1	46.6	73.5	82.7	**2.0**	13.6
+ KDProR^{+}	45.5	73.4	82.7	**2.0**	14.1	47.8	73.5	83.1	**2.0**	13.8
+ KDProR$^{\top}$	46.4	74.0	82.9	**2.0**	13.8	48.1	74.0	83.5	**2.0**	13.7
+ KDProR*	47.0	74.4	82.9	**2.0**	13.6	48.7	73.6	83.6	**2.0**	13.7
+ KDProR†	46.9	74.0	82.8	**2.0**	13.6	48.4	73.7	83.8	**2.0**	13.3
+ KDProR‡	**47.3**	**74.6**	**83.3**	**2.0**	**13.4**	**49.0**	**74.8**	**83.8**	**2.0**	**12.9**

(2) The integration of additional open-world knowledge significantly bolsters the predictive capabilities of retrieval models. KDProR of employing three open-world knowledge (i.e., pre-trained, structural, and CoT knowledge) in Section 3.6 achieves superior performance, e.g., +1.8%, +1.6% and + 2.2% R@1 on MSR-VTT, 3.8%, 2.6% and 5.3% R@1 on DiDeMo compared to DRL [48], respectively. We attribute the performance gain to incorporating open-set knowledge.

(3) KDProR demonstrates superior performance compared to all baselines across all benchmarks. This demonstrates that our KDProR offers a twofold advantage: it can unify and incorporate diverse additional knowledge and balance memorization and retrieval to significantly improve performance.

Comparisons to Other Baseline Methods. We further compare our method with other baselines in Table 3, including Cap4Video [52] and TEFAL [18]. In addition, we compare it with two baseline methods that introduce open-world knowledge, namely using auxiliary captions generated by LLM as data augmentation and directly concatenating the features generated by CLIP-VIP with the original text-video features. The results indicate that introducing open-world knowledge can enhance performance, and our KDProR achieves better performance gains than `Data Augmentation` and `Feature Concatenate`. Impressively, KDProR significantly outperforms state-of-the-art open-set baselines, e.g., +1.4% R@1 by Cap4Video on LSMDC and +1.2% R@1 by TEFAL on DiDeMo.

Generalization Analysis. We further equip KDProR with two powerful base models in addition to DRL, i.e., CLIP4clip [35] and EMCL [21], and evaluate

Table 5: Ablation experiments on MSR-VTT 9k dataset. ‡ denotes the utilization of open-world CoT knowledge in Section 3.6.

G.	Method	Text-to-Video Retrieval				Video-to-Text Retrieval			
		R@1↑	R@5↑	R@10↑	MnR↓	R@1↑	R@5↑	R@10↑	MnR↓
1	w/o all Knowledge Stores (baseline)	46.9	74.0	82.4	12.5	45.8	74.0	84.3	8.8
	w/o Fine-grained Knowledge Store	48.1	74.4	83.3	12.1	46.8	73.5	84.0	8.5
	w/o Coarse-grained Knowledge Store	48.4	74.5	83.7	11.9	47.2	73.8	84.0	8.4
2	w/o the whole E-step	47.7	73.9	83.4	12.0	46.2	73.5	84.1	8.9
	w/o Knowledge Interpolation E-step	48.2	73.6	83.7	12.0	47.4	73.9	83.8	8.7
	w/o OT Alignment Correction in E-step	48.9	74.6	84.2	11.8	46.6	73.4	83.6	8.9
3	w/o the whole K-step	47.9	73.8	83.1	12.3	47.0	73.6	84.0	8.7
	w/o Fine-grained Calibration in K-step	48.7	74.5	83.8	11.8	47.6	73.8	84.0	8.5
	w/o Coarse-grained Calibration in K-step	48.5	74.3	83.3	11.8	47.3	73.9	83.6	8.5
4	Our Full KDProR‡	**49.6**	**75.1**	**84.4**	**11.6**	**48.2**	**74.2**	**84.2**	**8.1**

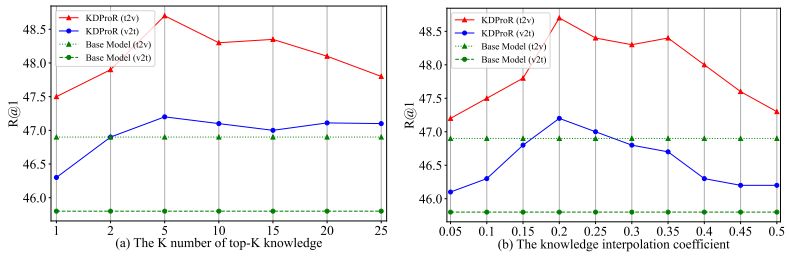

Fig. 4: Ablation experiment on MSR-VTT under closed-world settings. Effect of (a) the number K of top-K knowledge; (b) the knowledge interpolation coefficient ρ.

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. The significant improvements prove the generalization ability of KDProR.

4.3 Ablation Study and Analysis

Effect of the Multi-Grained Knowledge Store. As shown in Groups 1 and 4 in Table 5, removing either fine-grained or coarse-grained knowledge stores will result in a significant decrease in retrieval performance, e.g., -1.5% and -1.2% R@1 text-to-video retrieval on MSR-VTT, respectively. When all stores are removed, the model will lose the ability to decouple knowledge and introduce additional knowledge, resulting in a significant drop in performance. This verifies the effectiveness of our multi-grained knowledge decoupling strategy.

Effect of the Knowledge Injection in E-step. As shown in Groups 2 and 4 in Table 5, we evaluate the effect of the knowledge injection in E-step (i.e., Eq. 2) by removing different components. We can observe that when the knowledge interpolation strategy is removed, the performance drops sharply, indicating that our strategy can effectively inject knowledge from stores. We further remove the

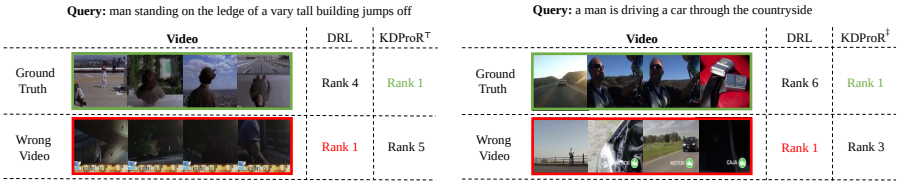

Fig. 5: The text-video results on the MSR-VTT 1K-A test set. $^{\top}$ and ‡ are close- and open-world settings respectively, whose meanings are consistent with those in Table 1.

OT alignment correction strategy (i.e., Eq. 4) and observe a significant decrease in video-to-text retrieval performance, proving our strategy's effectiveness.

Effect of the Distribution Calibration in K-step. We conduct experiments to study the effect of the KNN distribution calibration strategy as illustrated in Table 5 with Groups 3 and 4. It can be seen that the performance of KDProR significantly decreases without KNN distribution to calibrate the retrieval distribution in Eq. 9. Moreover, the lack of distribution calibration (fine-grained or coarse-grained) in the K-step can also affect retrieval performance. This verifies that the multi-grained distribution calibration strategy can calibrate the retrieval distribution to inject additional knowledge sufficiently and effectively.

Effect of the number K of top-K Knowledge. In Figure 4 (left), we show the effect of the number K of top-K knowledge. On the one hand, we find that fewer K ($K < 5$) means less knowledge is selected, which limits our KDProR's generalization ability for different knowledge. On the other hand, a larger K ($K > 5$) requires more retrieval time, which increases the cost of learning.

Effect of the Knowledge Interpolation Coefficient. In Figure 4 (right), we show the effect of the knowledge interpolation coefficient ρ on KDProR. We can observe that KDProR achieves the optimal balance between knowledge injection and optimization when $\rho = 0.2$. Deviations from this optimal ρ, either towards a smaller or larger number, result in performance degradation, which is attributed to inadequate knowledge acquisition and insufficient optimization, respectively.

Cases Study. We provide two cases of videos retrieved by KDProR and DRL [48]. As illustrated in Figure 5, KDProR successfully retrieves ground-truth videos in both closed- and open-set settings, whereas DRL fails to do so.

5 Conclusion

In this work, we explore a novel knowledge-decoupling VTR paradigm that utilizes multi-grained knowledge stores to unify closed- and open-world knowledge. This strategy allows VTR models to achieve an efficient balance between memorization and retrieval optimization, while also opening up a unified interface for injecting various open-world knowledge. We further propose a principled EKM algorithm to achieve an efficient injection of knowledge. Our method could be applied to other visual-language tasks, e.g., visual question answering, and injecting various open-world knowledge for augmenting cross-modal understanding.

Acknowledgements

We would like to thank all reviewers for their insightful comments. This paper was partially supported by NSFC (No: 62176008).

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: 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. pp. 1728–1738 (2021)
2. 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)
3. 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)
4. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em - algorithm plus discussions on the paper. In: "" (1977)
6. 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)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2020)
9. Elangovan, A., He, J., Verspoor, K.M.: Memorization vs. generalization : Quantifying data leakage in nlp performance evaluation. In: Conference of the European Chapter of the Association for Computational Linguistics (2021)
10. 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)
11. Feldman, V.: Does learning require memorization? a short tale about a long tail. Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (2019)
12. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 214–229. Springer (2020)
13. Gorti, S.K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., Yu, G.: X-pool: Cross-modal language-video attention for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5006–5015 (2022)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015)

15. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 961–970 (2015)
16. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.C.: Localizing moments in video with natural language. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 5804–5813 (2017)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**, 1735–1780 (1997)
18. 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)
19. Jensen, J.L.W.V.: Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Mathematica **30**, 175–193 (1906)
20. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. ArXiv **abs/2102.05918** (2021)
21. 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)
22. 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)
23. Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., Lewis, M.: Generalization through memorization: Nearest neighbor language models. arXiv preprint arXiv:1911.00172 (2019)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)
25. Li, H., Cao, M., Cheng, X., Li, Y., Zhu, Z., Zou, Y.: G2l: Semantically aligned and uniform video grounding via geodesic and game theory. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 11998–12008 (2023), `https://api.semanticscholar.org/CorpusID:260164585`
26. Li, H., Cao, M., Cheng, X., Zhu, Z., Li, Y., Zou, Y.: Exploiting auxiliary caption for video grounding. In: AAAI Conference on Artificial Intelligence (2023), `https://api.semanticscholar.org/CorpusID:257772084`
27. 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)
28. 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)
29. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 11895–11905 (2021)
30. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11915–11925 (2021)
31. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487 (2019)

32. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv **abs/1907.11692** (2019)
33. Liu, Y., Xiong, P., Xu, L., Cao, S., Jin, Q.: Ts2-net: Token shift and selection transformer for text-video retrieval. In: European Conference on Computer Vision. pp. 319–335. Springer (2022)
34. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv: Learning (2016)
35. 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)
36. 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)
37. Meng, Y., Zong, S., Li, X., Sun, X., Zhang, T., Wu, F., Li, J.: Gnn-lm: Language modeling based on global contexts via gnn. ArXiv **abs/2110.08743** (2021)
38. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (2013)
39. 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)
40. 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)
41. Papernot, N., McDaniel, P.: Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. arXiv preprint arXiv:1803.04765 (2018)
42. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824 (2020)
43. 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)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
45. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3202–3212 (2015)
46. Su, B., Hua, G.: Order-preserving wasserstein distance for sequence matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1049–1057 (2017)
47. Tänzer, M., Ruder, S., Rei, M.: Memorisation versus generalisation in pre-trained language models. In: Annual Meeting of the Association for Computational Linguistics (2021)

48. Wang, Q., Zhang, Y., Zheng, Y., Pan, P., Hua, X.: Disentangled representation learning for text-video retrieval. ArXiv **abs/2203.07111** (2022)
49. Wang, Z., Sung, Y.L., Cheng, F., Bertasius, G., Bansal, M.: Unified coarse-to-fine alignment for video-text retrieval. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 2804–2815 (2023)
50. 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)
51. 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)
52. 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)
53. Wu, X., Li, H., Luo, Y., Cheng, X., Zhuang, X., Cao, M., Fu, K.: Uncertainty-aware sign language video retrieval with probability distribution modeling. ArXiv **abs/2405.19689** (2024), `https://api.semanticscholar.org/CorpusID:270123137`
54. Xie, Y., Zhu, Z., Zhuang, X., Liang, L., Wang, Z., Zou, Y.: Gpa: Global and prototype alignment for audio-text retrieval. In: Interspeech (2024)
55. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5288–5296 (2016)
56. Xu, Z., Chen, Z., Zhang, Y., Song, Y., Wan, X., Li, G.: Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17503–17512 (2023)
57. Xu, Z., Huang, J., Liu, T., Liu, Y., Han, H., Yuan, K., Li, X.: Enhancing fine-grained multi-modal alignment via adapters: A parameter-efficient training framework for referring image segmentation. In: 2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ ICML 2024) (2024)
58. 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)
59. 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)
60. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Trans. Mach. Learn. Res. **2022** (2022)
61. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: European Conference on Computer Vision (2018)
62. Yuan, L., Chen, D., Chen, Y.L., Codella, N.C.F., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., Zhang, P.: Florence: A new foundation model for computer vision. ArXiv **abs/2111.11432** (2021)

63. Zhao, Y., Li, K., Cheng, Z., Qiao, P., Zheng, X., Ji, R., Liu, C., Yuan, L., Chen, J.: Graco: Granularity-controllable interactive segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3501–3510 (June 2024)
64. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16965–16974 (June 2024)
65. Zheng, X., Jiang, J.: An empirical study of memorization in nlp. In: Annual Meeting of the Association for Computational Linguistics (2022)