Using Multimodal Contrastive Knowledge Distillation for Video-Text Retrieval

Wentao Ma, Qingchao Chen, Tongqing Zhou, Shan Zhao, and Zhiping Cai

***Abstract*— Cross-modal retrieval aims to enable a flexible bi-directional retrieval experience across different modalities (*e.g.*, searching for videos with texts). Many existing efforts tend to learn a common semantic representation embedding space in which items of different modalities can be directly compared, wherein the positive global representations of video-text pairs are pulled close while the negative ones are pushed apart via pair-wise ranking loss. However, such a vanilla loss would unfortunately yield ambiguous feature embeddings for texts of different videos, causing inaccurate cross-modal matching and unreliable retrievals. Toward this end, we propose a multimodal contrastive knowledge distillation method for instance video-text retrieval, called MCKD, by adaptively using the general knowledge of self-supervised model (teacher) to calibrate mixed boundaries. Specifically, the teacher model is tailored for robust (less-ambiguous) visual-text joint semantic space by maximizing mutual information of co-occurred modalities during multimodal contrastive learning. This robust and structural inter-instance knowledge is then distilled, with the help of explicit discrimination loss, to a student model for improved matching performance. Extensive experiments on four public benchmark video-text datasets (MSR-VTT, TGIF, VATEX, and Youtube2Text) demonstrate that our MCKD can achieve at most 8.8%, 6.4%, 5.9%, and 5.3% improvement in text-to-video performance by the R@1 metric, compared with 14 SoTA baselines.**

***Index Terms*— Cross-modal retrieval, contrastive learning, knowledge distillation.**

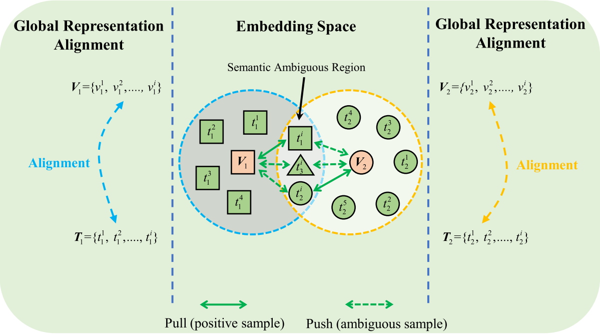

Global Representation Alignment
Embedding Space
Global Representation Alignment
Semantic Ambiguous Region
$V_1=\{v_1^1, v_1^2,...., v_1^i\}$
Alignment
$T_1=\{t_1^1, t_1^2,...., t_1^i\}$
t_1^2
t_1^1
V_1
t_1^3
t_1^4
t_1^i
t_3^i
t_2^i
t_2^4
t_2^3
V_2
t_2^1
t_2^5
t_2^2
$V_2=\{v_2^1, v_2^2,...., v_2^i\}$
Alignment
$T_2=\{t_2^1, t_2^2,...., t_2^i\}$
Pull (positive sample)
Push (ambiguous sample)

Fig. 1. An illustration of potentially ambiguous semantic boundaries between texts to describe different videos. The same shape indicates relevant semantics. Colors represent modalities (*i.e.*, video and text). Assume that $V_1 = \{v_1^1, v_1^2, \ldots, v_1^i\}$ and $V_2 = \{v_2^1, v_2^2, \ldots, v_2^i\}$ represent the video containing i views (sequence frames), whose corresponding text sentence description sets are $T_1 = \{t_1^1, t_1^2, \ldots, t_1^i\}$ and $T_2 = \{t_2^1, t_2^2, \ldots, t_2^i\}$, respectively. If the text annotations are partially aligned with their corresponding video concepts (*a.k.a.*, "Semantic Ambiguous Region" in the Figure), which may bring a spurious correlation between the text and the non-corresponding video.

I. Introduction

With the rapid development of mobile Internet and digital media, multimedia data with video as the carrier is generated in cyberspace (such as YouTube and TikTok platforms) all the time. Video-text cross-modal retrieval [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] as a promising data management technique has attracted attention in both the community of academia and industry. This technique aims to harness one modality as a probe to search instances from another modality: retrieving the videos by a text description (text-based video retrieval) or retrieving the text descriptions that are most relevant to the video content (video-based text retrieval).

In recent years, the pair-wise ranking loss is a popular objective function used in a broad range of tasks about video-text cross-modal retrieval [1], [4], [18], [19], which makes the distance between positive sample pairs smaller than the distance between negative ones by a pre-defined margin. As known, existing methods [1], [3], [4], [6], [7] adopt the pair-wise ranking loss focus on the distance between *global* representations of video and text. However, we argue that vanilla *global* representation alignment using the pair-wise ranking loss is challenging and sometimes ill-suited for video-text retrieval in realistic practice. As semantic concepts in the videos are complex, it commonly exists in video-text retrieval benchmarks [3], [20], [21], [22] that multiple text sentences are able to describe the *same* video from *different* views. Therefore, the multi-view text representations bring a unique challenge to pair-wise ranking loss: semantic boundaries between texts to describe different videos are potentially *ambiguous*, as shown in Fig. 1. It is because if text annotations are only partially aligned with their corresponding video

Manuscript received 23 August 2022; revised 4 December 2022; accepted 4 March 2023. Date of publication 14 March 2023; date of current version 4 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62072465, Grant 62172155, Grant 62102425, and Grant 62201014; in part by the Science and Technology Innovation Program of Hunan Province under Grant 2022RC3061 and Grant 2021RC2071; and in part by the Postgraduate Research and Innovation Project of Hunan Province under Grant CX20210080. This article was recommended by Associate Editor Y. Qin. *(Corresponding authors: Qingchao Chen; Zhiping Cai.)*

Wentao Ma, Tongqing Zhou, and Zhiping Cai are with the College of Computer Science, National University of Defense Technology, Changsha 410073, China (e-mail: wtma@nudt.edu.cn; zhoutongqing@nudt.edu.cn; zpcai@nudt.edu.cn).

Qingchao Chen is with the National Institute of Health Data Science, Peking University, Beijing 100091, China (e-mail: qingchao.chen@pku.edu.cn).

Shan Zhao is with the School of Computer and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: 2022800040@hfut.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3257193.

Digital Object Identifier 10.1109/TCSVT.2023.3257193

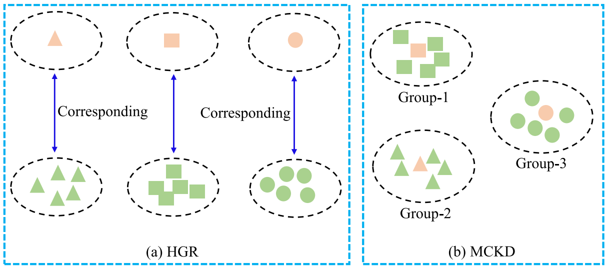


Fig. 2. A simple visual demonstration of the latent embedding spaces learned via different methods. The same shape indicates relevant semantics and colors represent modalities. Different from the existing methods, *i.e.*, HGR [4]), our proposed MCKD defines a "video/text group" as a video with its associated sentences. Therefore, in the training stage, we treat every "video/text group" as a different instance category to yield multimodal contrastive in intra-instance.

concepts, the partial alignment brings a spurious correlation between texts and their non-corresponding videos. That is, text annotations of the corresponding videos are not equally reliable as the "positive pairs" in the pair-wise ranking loss. *Existing solutions relying on unreliable positive pairs, therefore, bring unstable optimization and also collapsed joint visual-text semantic space for visual-text retrieval.*

In light of the above analysis, to maintain the reliable yet discriminative visual-text joint semantic space, it is essential to capture the semantic relationship between the text features and their corresponding visual features, enabling better-curated and reliable positive and negative pairs. By pulling reliable positive sample pairs considering their relationships and pushing unreliable ones, we propose a Multimodal Contrastive Knowledge Distillation (MCKD) model to transfer the reliable and structural inter-instance information, regularizing the cross-modal joint semantic space in parallel with the usage of pair-wise ranking loss. In essence, this procedure can maximize the consistency of mutual information between the intrinsic co-occurrence modalities to bridge the heterogeneous gap between different modalities and excavate the discrimination from intra- and inter-instance.

Concretely, there is ambiguity in semantic boundaries between texts describing different videos, which is obviously undesirable for instance-level cross-modal retrieval tasks. To suppress this unreliable semantic text representation, similar to [19] and [23], our MCKD names the video and its associated text sentence an "video/text group". This is based on the strong assumption that every "video/text group" is a different instance category, namely, it can be treated as a distinct instance, as shown in Fig. 2. Then, considering that multimodal data intrinsically consists of multiple modalities, inspired by recent works [8], [24], [25], we propose to learn feature representations by maximizing consistency between different modalities in the visual-text semantic embedding space, which is a self-supervised manner to implement the classification of video/text groups. Our end is to carry out this model to discriminate between every two videos and two texts from distinct groups. It is beneficial to investigate the stability of video-text semantic space, that is, to eliminate unreliable text representation. It's worth noting that "video/text group" can avoid the risk about the collapse of joint vision-text semantic space, which leads to discriminants of multiple text sentences (different views corresponding to the same video), for example, "two dogs are chasing on the lawn" is semantically equivalent to "two dogs on the lawn" after multimodal contrastive learning. Hence, the pair-wise ranking loss is leveraged to preserve discrimination between texts in intra-instance.

Furthermore, the pair-wise ranking loss employs pre-defined hard similarity to determine positive and negative pairs. However, the hard similarity may discard intra- and inter-instance correlations. Inspired by [18], [19], [23], and [24], our model also discusses the practice of "soft label" in video-text matching. Briefly, we leverage the knowledge distillation module [8], [18], [26], [27], [28] to combine the advantages of multimodal contrastive loss and pair-wise ranking loss. That is, multimodal contrastive loss works by narrowing the heterogeneous gap and then provides the "soft label" supervised signal for pair-wise ranking loss to preserve the discrimination between instance texts. The main contributions of this work are summarized as follows:

- We propose a novel framework for video-text cross-modal retrieval to bridge the gap between distinct modalities. That is, it can effectively learn the discriminative feature representations for heterogeneous data.
- To maintain a reliable yet discriminative visual-text joint semantic space, we propose a multimodal contrastive loss of video-text matching to suppress ambiguous text semantic representations, which can provide robust and structured inter-instance sample signals.
- We propose a pair-wise ranking loss with the "soft label" distillation signal to preserve discrimination between text intra-instance. Namely, the knowledge distillation module is adopted to transfer the reliable and structural inter-instance information, regularizing the cross-modal joint semantic space in parallel with the usage of pair-wise ranking loss.
- We evaluate our MCKD via the comparison with 14 State-of-The-Art (SoTA) baselines and a series of ablation studies. The extensive experimental results show that MCKD can yield promising performance (*a.k.a.*, R@1 of 13.8%, 6.8%, 39.6%, and 19.1% on MSR-VTT, TGIF, VATEX, and Youtube2Text, respectively.)

The remainder of this manuscript is structured as follows. First, we briefly review the most related works to our method in Section II. Section III introduces the multimodal contrastive knowledge distillation model. Then, we present the experimental settings and analysis of the corresponding experimental results in Section IV and V. Finally, conclusions are given in Section VI.

II. Related Work

In this section, we review the video-text retrieval, contrastive learning, and knowledge distillation that are most relevant to our work. Meanwhile, to facilitate a comparison of the proposed MCKD with existing methods, we briefly list the differences in technical components and functional modules, as shown in Appendix A.

A. Video-Text Retrieval

Early concept-based efforts promote the development of video-text retrieval techniques [12], [13], but they are still hard to fully explore the diversity and rich fine-grained semantic representation of video-text within limited concepts. For fine-grained video-text cross-modal retrieval, graph-free [3], [5], [6], [7], [8], [10], [11], [15], [29], [30], [31] and graph-based paradigms [1], [4], [16], [17], [32] are used to jointly encode video and text feature representations into a common embedding semantic space then the effective video-text matching is realized.

1) Graph-Free Paradigm: To capture coarse-to-fine-grained and spatio-temporal feature representations, Dong et al. [5] propose a three-branch method that utilizes mean pooling, BiGRU, and CNN to encode visual-text feature with multi-level granularity. To bridge the heterogeneous gap between visual-text, they present a hybrid embedding space [31], which represents the richness of both modalities coarse-to-fine by multi-level encoding and harnesses hybrid spatial learning to better align cross-modal matching. Yang et al. [7] design a video-text retrieval framework that consists of a tree-augmented text encoder and a temporal attentive video encoder. This method enables better alignment by transforming text containing semantic information into an easy-to-interpret structure. Yu et al. [29] present a JSFusion, which is composed of joint semantic tensor module and convolutional hierarchical decoder module to estimate video-text hierarchical semantic similarity and realize multi-level semantic similarity fusion. Another, Liu et al. [15] propose a PSM model to optimize the visual-text semantic representation space, which employs a progressive learning strategy with a coarse-to-fine architecture to narrow the semantic gap between distinct modalities. Wang et al. [11] design an efficient multi-level alignment model of feature representation for video-text retrieval, called T2VLAD, which enables the meticulous local comparison while reducing the computational overhead of semantic representation interaction between each video-text pair.

2) Graph-Based Paradigm: Alternatively, some researchers focus on adopting graph modeling to generate fine-grained semantic relationships of visual-text representations. Feng et al. [32] propose a visual semantic enhanced inference model, called ViSERN, which utilizes graph convolutional networks based on random walk rules to learn semantic inference. Chen et al. [4] (HGR), Jin et al. [1] (HCGC), and Ma et al. [17] (QAMF) disentangle the feature representation of video-text pair into multiple levels and fuse the video-text matching of different levels respectively. In particular, HGR utilizes a semantic graph, attention layer, and full connection layer to parse video-text pairs into a hierarchical semantic graph including events, actions, and entities respectively. Meanwhile, the multiple embedding common spaces of representation are used to calculate the average similarity and then fused as the final similarity. Considering the graph consistency of multi-level matching, HCGC designs a learning model for multi-level graph consistency to enhance the intra-graph and inter-graph interaction consistency. The HGR and HCGC achieve promising performance in video-text matching, but both efforts fail to realize adaptive fusion, QAMF proposes a query-adaptive fusion mechanism to enable differential fusion of multi-level semantic representation. Meanwhile, to address the asymmetry of video-text feature representations, Wu et al. [16] design a HANet to align the representations of different levels, which realizes semantic coverage cross-modal data from coarse-to-fine. However, these methods achieve remarkable performance via complex graph modeling, they hardly enable high efficiency of video-text matching.

B. Contrastive Learning

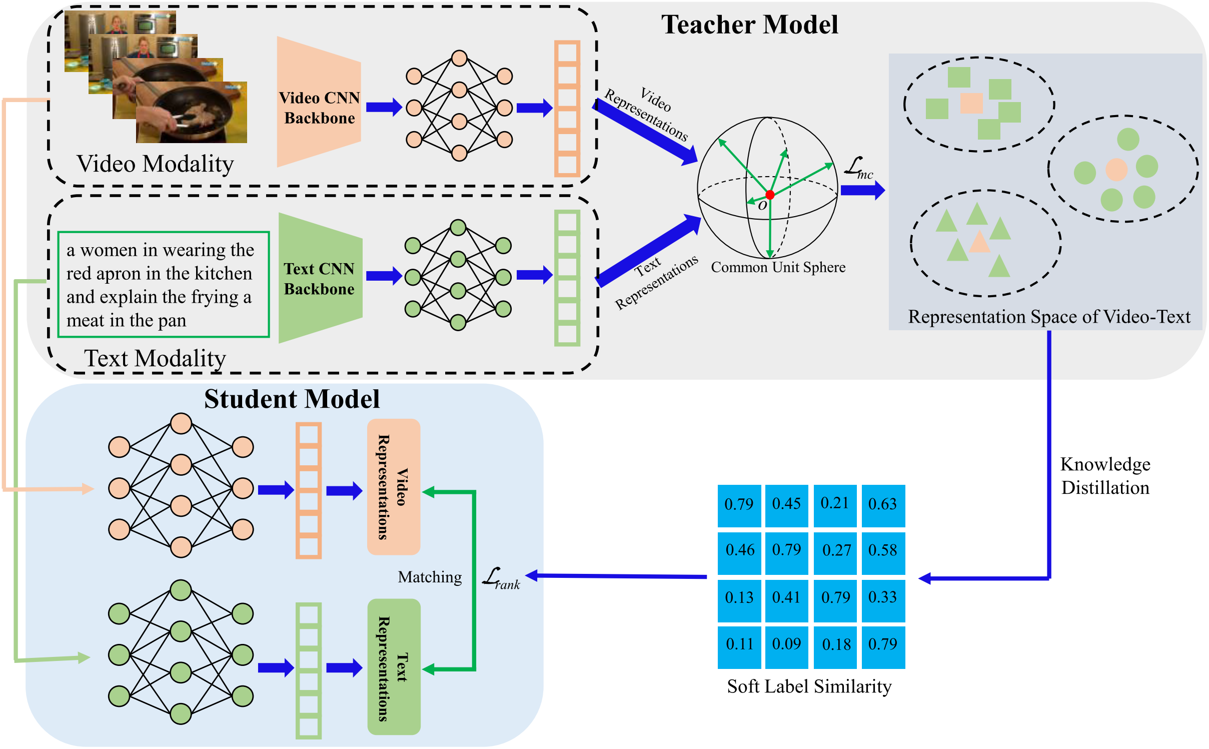
Contrastive learning is a framework that can learn feature representations with strong discriminative to improve the performance of downstream tasks, the core idea of which is to draw the same instance closer to each other under pre-defined marginal metrics and push apart ones between different instances [25], [33].

As we all know, contrastive learning has achieved outstanding results in multimodal applications, such as cross-modal retrieval [10], [24], [34], [35]. Zhang et al. [10] investigate a heuristic model, ReLoCLNet, which leverages two contrastive losses to realize moment retrieval tasks with a decent performance. Nevertheless, this method only considers the hierarchical inter-modal consistency, ignoring the consistency of intra-modal. Another, to mitigate the influence of noisy labels in cross-modal retrieval, Hu et al. [24] propose a novel multimodal robust learning framework for image-text matching with noisy labels via supervised robust clustering and multimodal contrastive. However, this supervised clustering method with noise label suppression also tends to fit the wrong labels to shift cluster centers in the process of model iteration. Thus, Zhang et al. [35] propose an intra-modal contrastive learning criterion for robust feature embedding by maximizing the similarity among samples with the same labels. Although these approaches have achieved non-trivial results by contrastive loss, learning semantic alignments between videos and texts is more challenging since videos are more complex temporal-spatial representation information than images. Reference [34] is the first implementation of the Contrastive Language-Image Pre-training (CLIP) model for video-text cross-modal retrieval, to capture the semantic interaction between video and text modality. This method, however, is a zero-shot task, leveraging the power of CLIP's visual representation without the need for parameter fine-tuning. As a result, there is still much room for improvement in deployment of video-text retrieval tasks based on contrastive learning.

Consequently, inspired by the recent works [10], [24], [34], [35], we present a multimodal contrastive loss, which names the video and its associated text sentence an "video/text group" according to the property of multimodal data. Then it explicitly and fully considers the intra-modal data distribution to compensate for the drawback of pair-wise ranking loss.

C. Knowledge Distillation

Knowledge distillation is essentially a model-agnostic compression strategy used to generate efficient models while



0.79	0.45	0.21	0.63
0.46	0.79	0.27	0.58
0.13	0.41	0.79	0.33
0.11	0.09	0.18	0.79

Fig. 3. An overview of the proposed MCKD model, which consists of two modules: multimodal contrastive learning as teacher model (shown in top part of the figure) and the pair-wise ranking loss as student model (in bottom part of the figure). The teacher model leverages multimodal contrastive loss, which mitigates the cross-modal discrepancy and tries to maximally scatter inter-modal samples while compacting intra-modal points over the common unit sphere/space, thus eliminating the unreliable text representation. Meanwhile, the teacher model transfers the reliable and structural inter-sample information (*a.k.a.*, similarity matrix S), regularizing the cross-modal joint semantic space in parallel with the usage of pair-wise ranking loss.

preserving performances in a teacher-student paradigm, that is, transferring knowledge extracted from trained models to another model as a supervised signal. Since the early success of knowledge distillation in computer vision [27], it has been accepted for a broad range of applications, such as recommendation systems and cross-modal retrieval.

For recommendation systems [26], several works adopt the knowledge distillation to complete downstream tasks. Tang and Wang [26] propose a ranking distillation for a recommendation system, which can generate compact ranking models and improve efficiency. While for cross-modal retrieval [8], [18], [28], [36], some studies use a heuristic teacher-student paradigm that brings a remarkable performance. Hu et al. [28] proposes a novel unsupervised cross-modal hashing, which utilizes the unsupervised teacher model to extract extensive interaction information and transmit robust guidance signals to the student model, obtaining better cross-modal matching performance. Yet, considering that existing methods all require large-scale annotated information, Li et al. [18] designs a framework of cross-modal hashing retrieval based on knowledge distillation, called KDCMH, which extracts similarity supervised signals in an unsupervised manner to guide the student model. The most similar work to ours is Croitoru et al. [8], which investigates a novel generalized distillation method, TEACHTEXT, to explore the effectiveness of large-scale language pre-training models in video-text cross-modal retrieval tasks. However, this method not only brings great computational overhead but also ignores inter-instance consistency. Therefore, in this work, we combine the advantages of pair-wise ranking loss and multimodal contrastive loss via the knowledge distillation module, which builds a robust yet discriminative visual-text joint semantic space, enabling better-curated and reliable positive and negative pairs.

III. The Proposed MCKD Model

In this section, we will elaborate on our proposed MCKD model. First, we briefly introduce the problem definition and notation (Section III-A). Then, we demonstrate the specific pipeline module of MCKD, including distillation supervision signal (Section III-B), reviews of pair-wise ranking loss (Section III-C), and multimodal contrastive loss (Section III-D).

A. Problem Definition and Notation

Without losing the generality of cross-modal matching, our MCKD focuses on feature representation for video-text bimodal data. That is, for the given set of videos (or video clips) and a set of texts, the MCKD is to harness one modality as a query to search all semantically related instances from another modality. Fig. 3 shows the framework of MCKD, which mainly consists of two modules: multimodal contrastive loss (teacher model) and supervised pair-wise ranking loss (student model). Meanwhile, for readability and clarity, some of the notations adopted in our paper and their definitions are described in Table I. We introduce each part in detail in the following.

TABLE I

KEY NOTATIONS

Notations	Description
S	S represents the cosine similarity matrix.
n	n indicates the number of instances in a batch-size.
x, y	x and y indicates the indexes in similarity matrix S.
(V^+, T^+)	(V^+, T^+) indicates the positive video-text pairs in a batch-size.
(V^+, T^-), (T^+, V^-)	(V^+, T^-) and (T^+, V^-) indicate the negative video-text pairs in a batch-size.
$\mathcal{U}$	$\mathcal{U}$ indicates visual-text embedding space.
m	m represents the number of modalities in the data samples.
g_i	g_i indicates the representation function for i-th modality.
$\mathbf{n}_j^i$, $\mathbf{u}_j^i$	$\mathbf{n}_j^i$ and $\mathbf{u}_j^i$ represent the data sample and feature representation of the i-th modality about the j-th instance sample, respectively.

B. Distillation Supervision Signal

The basic working principle of knowledge distillation is to adopt a complex model (teacher) to guide a lighter model (student) for collaborative training. In this work, we propose a multimodal contrastive knowledge distillation module to transfer the reliable and structural inter-instance information S to the student model with pair-wise ranking loss. In terms of cross-modal retrieval, most existing attempts tend to learn a common semantic representation embedding space in which items of different modalities can be compared to each other via a pre-defined margin. Namely, The pre-defined hard label similarity is denoted as $S \in \{0, 1\}^{n\times n}$ wherein n is the number of instances and $S_{xy} = 1$ indicates the corresponding pair under index (x, y) is a positive one, while the soft label similarity is denoted as $S \in [0, 1]^{n\times n}$ with S_{xy} a real value between [0, 1]. An overview of soft label similarity and hard label similarity is shown in Fig. 4. In our MCKD, the supervised signal output from the multimodal contrastive teacher model is the soft label similarity matrix $S \in [0, 1]^{n\times n}$. Hence, compared with the training hard label similarity, the soft label similarity output via the knowledge distillation module of the MCKD not only contains the similarity information of positive instances but also contains the semantic information of negative ones. Overall, we extract supervised signals with intra- and inter-instance correlations from the teacher model via knowledge distillation. Then, the supervised signal $S \in [0, 1]^{n\times n}$ of soft label similarity is utilized to guide student model training to construct a reliable yet discriminative semantic representation space.

C. Ranking Loss of Student Model

To clearly explain the bi-directional retrieval mechanism of video-text matching, we follow the implementation of pair-wise ranking loss in some previous works [1], [4], [17], [19]. Here, in a batch-size, V and T indicate the input feature representation of video and text, respectively. Concretely, for a given quadric input (V^+, T^+, V^-, T^-), which contains visual and textual feature representation vectors, the positive video-text pairs (V^+, T^+) will be pulled closer, while the hard negative video-text pairs (V^+, T^-) and (T^+, V^-) will be pushed further than the pre-defined marginal Δ. That is, the pair-wise ranking loss can be written as:

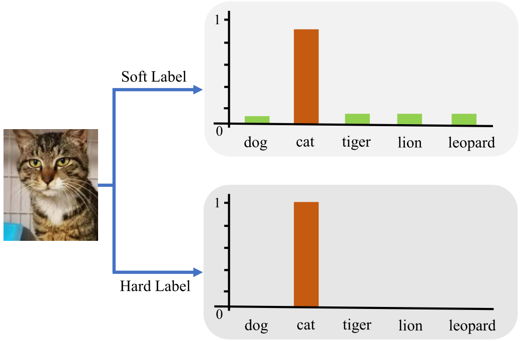


Fig. 4. An overview of examples about soft label similarity and hard label similarity.

$$\begin{aligned}\mathcal{L}_{rank} = &\overbrace{\left[\Delta + S(V^+, T^-) - S(V^+, T^+)\right]}^{\text{video anchor}} \\ &+ \underbrace{\left[\Delta + S(T^+, V^-) - S(T^+, V^+)\right]}_{\text{text anchor}}\end{aligned} \qquad (1)$$

where $S(\cdot, \cdot)$ represents the distance measurement criterion, we adopt cosine similarity in the experiment. For a query V^+ as the "video anchor", whose corresponding text sentence description should have a higher similarity. Also, with a query T^+ as the "text anchor", we expect the semantically relevant video candidates to rank higher. The pair-wise ranking loss is a basic matching strategy, although widely used, it focuses on the distance between global feature representations of video and text (as shown in Eq. (1)), thus it is sometimes unsuitable for video-text retrieval in real-world applications. For example, given several video clips with slightly different semantics, the model may output similar feature representations, resulting in a spurious correlation with the non-corresponding text. Namely, text annotations of the corresponding videos are not equally reliable as the "positive pairs" in the pair-wise ranking loss.

As a result, to maintain the reliable yet discriminative semantic space, inspired by the success of knowledge distillation and contrastive learning in cross-modal retrieval [18], [24], [28], we propose a novel framework for video-text matching, called MCKD, which can transfer robust and structural inter-sample information by knowledge distillation, regularizing the cross-modal joint semantic space in parallel via the pair-wise ranking loss.

D. Multimodal Contrastive Learning of Teacher Model

To learn a reliable visual-text embedding space $\mathcal{U}$, in which the samples of different modalities can be directly compared with each other, the existing methods tend to learn m modality-specific feature representation functions $\{g_i : \mathcal{N}_i \mapsto \mathcal{U}\}_{i=1}^m$ and the g_i can be the feature representation with arbitrary parameters for the i-th modality. Then, suppose that given some data samples $\mathbf{n}_j^i$ (including video, text, audio, and others) of j-th instance, the feature representation vector $\mathbf{u}_j^i$ of all modalities data can be calculated by

$$\mathbf{u}_j^i = g_i\left(\mathbf{n}_j^i\right) \in \mathbb{R}^L \tag{2}$$

where L is the dimension of representation space of visual-textual. It's worth noting that, similar to the prior works, in the case of unimodal data scenario (*e.g.*, video or text), the feature representation can be obtained via Eq. (2). Thus standard contrastive learning loss of each sample instance $\mathbf{n}_j$ can be defined as:

$$\mathcal{L}_c = \frac{\exp\left(\frac{1}{\tau}\left(\mathbf{u}_j^a\right)^T \mathbf{u}_j^b\right)}{\sum_{j=1}^n [\exp\left(\frac{1}{\tau}\left(\mathbf{u}_j^a\right)^T \mathbf{u}_j^a\right) + \exp\left(\frac{1}{\tau}\left(\mathbf{u}_j^a\right)^T \mathbf{u}_j^b\right)]} \tag{3}$$

where τ is a temperature hyperparameter (following [24], [25], [37], [38]), n is the batch-size of sample pairs, $j \in [1, n]$, $\left\{\mathbf{u}_1^a, \ldots, \mathbf{u}_n^a\right\}$ and $\left\{\mathbf{u}_1^b, \ldots, \mathbf{u}_n^b\right\}$ denote the feature representations of two types of data augmentation for sample data $\mathbf{n}_j$. For feature representation $\mathbf{u}_j^a$, its corresponding augmented sample representation is $\mathbf{u}_j^b$ to form positive pairs $(\mathbf{u}_j^a, \mathbf{u}_j^b)$ and leave other pairs to be negative.

As mentioned above, unimodal data contrastive training is implemented in data augmentation. In contrast, multimodal data is intrinsically composed of multiple modalities that can naturally utilize the data of each modality in the instance to maximize mutual information. Inspired by recent works [24], [34], [39], we propose an instance-level multimodal contrastive loss, which explicitly and fully considers the intra-modal distribution to improve mutual information and suppress the unreliable semantic text representation intra-instances. Specifically, according to the assumption of "video/text group" (that is, each "video/text group" is a distinct instance category and each instance contains data for both video and text modalities). We first define the probability that $\mathbf{n}_j^i$ belongs to the j-th in a sample instance containing m modalities data as follows:

$$P\left(j \mid \mathbf{n}_j^i\right) = \frac{\sum_{i=1}^m \exp\left(\frac{1}{\tau}\left(\mathbf{u}_j^i\right)^T \mathbf{u}_j^i\right)}{\sum_{i=1}^m \sum_{j=1}^n \exp\left(\frac{1}{\tau}\left(\mathbf{u}_j^i\right)^T \mathbf{u}_j^i\right)} \tag{4}$$

where τ is a temperature hyperparameter (following [24], [25], [37], [38]), since there are only two modalities of data in our work, namely, videos and texts, the $m = 2$.

Accordingly, to bridge the semantic gap of cross-modal data and excavate the differences between instance samples, our MCKD model makes the multimodal data (including video and text in this work) from the same instance (*e.g.*, $\left\{\mathbf{n}_j^i\right\}_{i=1}^m$ for the j-th instance) close to each other (*a.k.a.*, minimizing the probabilities), while the samples from distinct instances (*e.g.*, $\left\{\mathbf{n}_l^i\right\}_{l \neq j}$ for the j-th instance) push away (*a.k.a.*, maximizing the probabilities). For a simplified formal description, the multimodal contrastive loss of our MCKD model can be formulated as maximizing a joint probability $\prod_{i=1}^m \prod_{j=1}^n P\left(j \mid \mathbf{n}_j^i\right)$, which works by a mechanism equivalent to minimizing the negative log-likelihood estimation [24], [25]:

$$\mathcal{L}_{mc} = -\frac{1}{n}\sum_{i=1}^m \sum_{j=1}^n \log\left(P\left(j \mid \mathbf{n}_j^i\right)\right) \tag{5}$$

Consequently, by minimizing Eq. (5), semantically relevant positive samples will be pulled closer (*a.k.a.*, considered as the data belongs to the same instance, *e.g.*, $\left\{\mathbf{n}_j^i\right\}_{i=1}^m$ for $\mathbf{n}_j^i$), while negative ones will be pushed apart (*a.k.a.*, considered as the data not to belong to an instance, *e.g.*, $\left\{\mathbf{n}_l^i\right\}_{l \neq j}$ for $\mathbf{n}_j^i$) in the joint semantic representation space of visual-text.

E. Objective Function

The overall loss function of our proposed MCKD model can be written as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{rank} + (1 - \alpha)\mathcal{L}_{mc} \tag{6}$$

The MCKD model employs a stochastic gradient descent optimization algorithm, Adam [24], [40], to minimize the joint loss function $\mathcal{L}$ under a batch-to-batch iteration manner. Hence, our model can maximize the consistency of mutual information between the intrinsic co-occurrence modalities to bridge the heterogeneous gap between different modalities and excavate the discrimination from intra- and inter-instance, thus maintaining the reliable yet discriminative joint semantic space of visual-text.

IV. Experimental Settings and Baselines

In this section, we introduce the four public benchmark datasets, evaluation protocols, implementation settings, and baselines.

A. Datasets

The **MSR-VTT** [20] dataset consists of 10k videos in which each video lasts 10 to 30 seconds and corresponds to 20 natural text descriptions. We follow the common splitting for experiments: 6573 videos are used for training, 497 and 2990 for validation and testing respectively.

The **TGIF** [21] dataset is composed of videos in GIF format, where each video corresponds to 1∼3 natural text descriptions. We follow the official splitting experiments: 79451 videos are used for training, 10651 and 11310 for validation and testing respectively.

The **VATEX** [3] dataset consists of 34991 videos in which each video has 10 natural text descriptions in English and Chinese, respectively. We follow the common splitting experiments: 25991 videos are used for training, 3000 and 6000 for validation and testing respectively. It is worth noting that, we only employ the English sentence description corresponding to each video in our experimental settings.

The **Youtube2Text** [22] dataset contains 1970 videos in which each video has 40 natural text descriptions. We follow the official splitting experiments: 1200 videos are used for training, 100 and 670 for validation and testing respectively. In our work, we leverage the test set of Youtube2Text to evaluate the generalization performance of the proposed MCKD.

B. Evaluation Protocols

For video-text retrieval on the public benchmark, we employ three common evaluation indicators to measure the performance of this proposed MCKD, namely, Recall at K (R@K), Median Rank (MedR), and Mean Rank (MnR). Here, R@K represents the possibility that the true match occurs in the top-K of the rank list, we set K = 1, 5, and 10 via the following tradition. While the MedR and the MnR are median rank and average rank of the retrieved rank list are closest to the ground truth results respectively, with a lower score being better. We also utilize the sum of all R@K as rsum to measure the overall performance of our MCKD.

$$rsum = \underbrace{R@1 + R@5 + R@10}_{\text{Text}\rightarrow\text{Video}} + \underbrace{R@1 + R@5 + R@10}_{\text{Video}\rightarrow\text{Text}} \tag{7}$$

C. Implementation Details

The experiments are conducted on a platform of configuration with Ubuntu18.04, Intel i7-9700KF CPU@3.60GHz, and two Nvidia GeForce RTX-2080Ti GPUs. Similar to [1], [4], [16]: for video encoding, we employ the pre-trained ResNet [41] as the backbone network to extract frame-level feature representations of MSR-VTT and TGIF, adopting the I3D [42] video feature representation provided by the official of VATEX. For text encoding, we follow the practice of prior work, which sets the word embedding size to 300, while adopting the pre-trained glove embedding [43] as the backbone of the encoder for initialization.

For training, we adopt the general optimizer, ADAM [24], [40], to train our model and set the initial learning rate to 1×10^{-3}. Furthermore, similar to [19], we implement a two stages training strategy in which the model trained within each step is used as the warm-up model for the next step. Specifically,

- Stage I: We freeze the pre-trained weights of the video-text dual-tower backbone network and fine-tune the remaining parts by only employing the proposed $\mathcal{L}_{mc}$. This strategy can suppress ambiguous text semantic representations, which can provide reliable and structured inter-instance signals.
- Stage II: Next, when the Stage I converges, we fine-tune our MKCD model via interaction distillation (that is, combining $\mathcal{L}_{mc}$ and $\mathcal{L}_{rank}$) on the video-text matching. Namely, the knowledge distillation module is adopted to transfer the reliable and structural inter-instance information.

For testing, we respectively select the epoch numbers (*a.k.a.*, epoch=37 on MSR-VTT, epoch=46 on TGIF, and epoch=43 on VATEX) with the best rsum on the three validation sets for inference.

D. Baselines

We evaluate the superiority of this proposed MCKD model by comparing it with two paradigms (*a.k.a.*, graph-free and graph-based) that include 14 SoTA methods in video-text matching. Briefly, the differences in technical components and functional modules are shown in Appendix A.

1) Graph-Free Paradigm: VSE [44]: It is a SoTA cross-modal retrieval model and is also regarded as a strong baseline in text-video or text-image retrieval tasks.

VSE++ [45]: An improved version of VSE, which utilizes a novel loss based on augmented data and fine-tuning to significantly improve cross-modal retrieval performance.

W2VV [30]: W2VV can transform natural language statements into meaningful visual feature representations, that is, the relevant video-text pairs will be pulled closer, while irrelevant ones will be pushed apart.

DualEn [5]: Mean pooling, biGRU, and CNN are leveraged to realize the visual-text pairs coarse-to-fine-grained and spatial-temporal feature representations.

S^{2}Bin [46]: Since video contains complex spatial-temporal features, an effective spatial-temporal video encoder and text encoder are designed in S^{2}Bin to learn fine-grained video cues information and text discrimination information, respectively.

DualEn* [31]: An improved version of DualEn [5], which employs a better sentence encoding strategy and an improved triplet ranking loss.

PSM [15]: It leverages progressive semantic matching to optimize the visual-text joint semantic representation space.

T2VLAD [11]: The global-to-local alignment framework is proposed, which enables the fine-grained feature representation compact and also reduces the complexity and computational cost of the interactions between each video-text pair.

2) Graph-Based Paradigm: ViSERN [32]: It utilizes graph convolutional networks based on random walk rules to learn semantic inference for video-text cross-modal retrieval, and then improves the alignment of video-text representations.

HGR [4]: The graph convolutional network is used to model the hierarchical representations of video and text respectively and the alignment of video-text pairs is implemented at three hierarchies of visual-text embedding common space.

HANet [16]: It employs multi-level video-text alignment to compensate for the asymmetry of cross-modal feature representation.

HCGC [1]: Multi-hierarchy graph consistency learning is leveraged to bridge the semantic gap between video-text cross-modal retrieval.

QAMF [17]: To realize adaptive fusion for video-text retrieval tasks, QAMF proposes a query-adaptive fusion mechanism to enable differential fusion of multi-level semantic representation.

V. Experimental Results and Analysis

In this section, we present the experimental results on four datasets to assess our MCKD by making comparisons with various SoTA baselines. To be specific, we attempt to answer 5 research questions (RQs) as our evaluation goals in experiments:

- **RQ1**: What is the performance of our MCKD as compared to various SoTA baseline methods?
- **RQ2**: How do the two stages training strategy, the $\mathcal{L}_{rank}$ and the $\mathcal{L}_{mc}$ in our MCKD affect the video-text matching performance?
- **RQ3**: How does the proposed $\mathcal{L}_{mc}$ influence the distribution of positive pairs and negative pairs?

TABLE II

VIDEO-TEXT BI-DIRECTIONAL CROSS-MODAL RETRIEVAL COMPARED WITH SOTA BASELINES ON MSR-VTT. HERE, "-" DENOTES THAT NO EXPERIMENTAL RESULTS WITH THE SAME SETTINGS ARE AVAILABLE. THE HIGHEST SCORE IS SHOWN IN **BOLD**

Methods	Text-to-Video Retrieval					Video-to-Text Retrieval					rsum
	R@1	R@5	R@10	MedR	MnR	R@1	R@5	R@10	MedR	MnR	
VSE [44]	5.0	16.4	24.6	47	215.1	7.7	20.3	31.2	28	185.8	105.2
VSE++ [45]	5.7	17.1	24.8	65	300.8	10.2	25.4	35.1	25	228.1	118.3
W2VV [30]	6.1	18.7	27.5	45	-	11.8	28.9	39.1	21	-	132.1
DualEn [5]	7.7	22.0	31.8	32	-	13.0	30.8	43.3	15	-	148.6
S^{2}Bin [46]	7.9	22.5	32.2	31	-	13.3	32.5	43.7	15	-	152.1
ViSERN [32]	7.9	23.0	32.6	30	178.7	13.1	30.1	43.5	15	119.1	151.1
HGR [4]	9.2	26.2	36.5	24	164.0	15.0	36.7	48.8	11	90.4	172.4
HANet [16]	9.3	27.0	38.1	20	-	16.1	39.2	52.1	9	-	181.8
HCGC [1]	9.7	28.0	39.2	19	129.5	17.1	40.5	53.2	9	58.2	187.7
CE [6]	10.0	29.0	41.2	16	86.8	15.6	40.9	55.2	8.3	38.1	191.9
QAMF [17]	11.0	28.4	39.2	22	150.3	16.2	37.9	50.9	11	89.3	183.6
DualEn* [31]	11.6	30.3	41.3	17	-	22.5	47.1	58.9	7	-	211.7
PSM [15]	12.0	31.7	43.0	16	-	**22.8**	48.0	61.0	6	-	218.5
T2VLAD [11]	12.7	34.8	47.1	12	-	20.7	48.9	62.1	6	-	226.3
Our MCKD	**13.8**	**37.9**	**49.2**	**10**	**93.2**	20.3	**49.4**	**62.6**	**6**	**45.2**	**233.2**

- **RQ4**: What is the generalization capacity of our proposed MCKD model on the unseen dataset, that is, zero-shot tasks on the Youtube2Text dataset?
- **RQ5**: In the overall loss function, what is the influence of different α values on the performance of our MKCD model?

A. Comparison With the SoTAs

To answer **RQ1**, we compare the performance of our MCKD with all the above baselines in video-text bi-directional retrieval on different datasets, as shown in Table II and Table III. For fairness of comparison, we implement the code and feature representations released by some methods. Meanwhile, we directly cite numbers from the original paper whenever appropriate. Table II demonstrates the results of our MCKD and the compared counterparts on MSR-VTT, we can draw the following two observations: 1) Our MCKD yields SoTA performance over all baselines, including the traditional and promising video-text retrieval methods. Compared with two promising counterparts, HGR and HCGC, our proposal is outstanding to them. Both methods implement hierarchical graph reasoning for fine-grained video-text matching. Yet, the HGR can hardly explore the video-text hierarchical matching strategy. In our MCKD, the multimodal contrastive loss is adopted to model the invariance of multimodal data intra-instance. While the HCGC jointly models multiple graph consistency learning in video-text cross-modal matching, its improved performance clearly demonstrates the advantages of relying on inter-modal and intra-modal relationship interactions. 2) The MCKD can also outperform the SoTA competitors, DualEn*, PSM, and T2VLAD in all indicators that include R@1, R@5, R@10, and rsum. Particularly, the rsum index reflecting the overall retrieval quality of the model is boosted by a large margin, relative +21.5%, +14.7%, and +6.9%, respectively.

As shown in Table III, the performance comparison between our MCKD and other methods on TGIF and VATEX. One can see that the MCKD consistently achieves the best performance compared to its counterparts on TGIF. It's worth noting that the same methods have lower metrics than those in Table II, which means the TGIF dataset is more complex than MSR-VTT. Even so, the MCKD can yield R@K (K = 1, 5, 10) of 6.8%, 18.7% and 25.6%, respectively. On VATEX, the MCKD outperforms all listed competitors again and keeps the performance of 39.6%, 77.4%, and 85.4% in R@K (K = 1, 5, 10), compared to 36.8%, 73.6%, and 83.7% of DualEn* [31].

TABLE III

TEXT-TO-VIDEO RETRIEVAL COMPARISON WITH SOTA BASELINES ON THE TGIF AND VATEX DATASETS. HERE, "-" DENOTES THAT NO EXPERIMENTAL RESULTS WITH THE SAME SETTINGS ARE AVAILABLE. THE HIGHEST SCORE IS SHOWN IN **BOLD**

Datasets & Methods		R@1	R@5	R@10	MedR
TGIF	DeViSE [47]	0.8	3.5	6.0	379
	VSE++ [45]	0.4	1.6	3.6	692
	Order [48]	0.5	2.1	3.8	500
	Corr-AE [49]	0.9	3.4	5.6	365
	PVSE [50]	2.2	7.8	12.3	155
	HGR [4]	4.5	12.4	17.8	160
	HCGC [1]	6.3	16.2	22.9	**79**
	QAMF [17]	6.7	14.9	20.7	159
	Our MCKD	**6.8**	**18.7**	**25.6**	103
VATEX	W2VV [30]	14.6	36.3	46.1	-
	VSE++ [45]	31.3	65.8	76.4	-
	CE [6]	31.1	68.7	80.2	3
	DualEn [5]	31.1	67.4	78.9	3
	W2VV++ [51]	32.0	68.2	78.8	-
	HGR [4]	35.1	73.5	83.5	2
	HANet [16]	36.4	74.1	84.1	2
	DualEn* [31]	36.8	73.6	83.7	-
	Our MCKD	**39.6**	**77.4**	**85.4**	**2**

B. Ablation Study

To answer **RQ2**, we conduct a series of ablation studies to investigate the contributions of different components (*i.e.*, two-stage training strategy, $\mathcal{L}_{rank}$, and $\mathcal{L}_{mc}$) on MSR-VTT. The experimental results are shown in Table IV that one can draw the following two conclusions:

- Regarding the training strategy, we respectively adopt $\mathcal{L}_{mc}$ and $\mathcal{L}_{rank}$ to evaluate performance under freezing pre-trained weights of dual-tower backbone in Stage I. As we can see from the top two rows of Table IV, the

TABLE IV

COMPARISON BETWEEN OUR MCKD (FULL VERSION) AND ITS FOUR COUNTERPARTS (NAMELY, TWO VARIATIONS UNDER STAGE I AND TWO VARIATIONS OF MCKD UNDER STAGE II) ON THE MSR-VTT DATASET. THE HIGHEST SCORE IS SHOWN IN **BOLD**

Methods	Stage	Text-to-Video Retrieval					Video-to-Text Retrieval					rsum
		R@1	R@5	R@10	MedR	MnR	R@1	R@5	R@10	MedR	MnR	
Only $\mathcal{L}_{rank}$	I	5.8	18.4	27.4	60	279.6	10.7	28.3	38.5	22	199.7	129.1
Only $\mathcal{L}_{mc}$	I	7.6	21.4	31.2	38	213.4	12.7	31.8	43.5	16	132.8	148.2
Only $\mathcal{L}_{rank}$	II	9.3	26.8	38.4	20	141.7	15.3	38.4	51.3	10	60.2	179.5
Only $\mathcal{L}_{mc}$	II	11.6	30.9	41.2	17	118.4	18.7	40.3	55.2	7	64.1	197.9
Full MCKD (with $\mathcal{L}_{rank}$ and $\mathcal{L}_{mc}$)	II	**13.8**	**37.9**	**49.2**	**10**	**93.2**	**20.3**	**49.4**	**62.6**	**6**	**45.2**	**233.2**

TABLE V

GENERALIZATION CAPACITY ON UNSEEN YOUTUBE2TEXT TEST SET UTILIZING VARIANT MODELS ON MSR-VTT. THE HIGHEST SCORE IS SHOWN IN **BOLD**

Methods	Text-to-Video Retrieval					Video-to-Text Retrieval					rsum
	R@1	R@5	R@10	MedR	MnR	R@1	R@5	R@10	MedR	MnR	
VSE [44]	11.0	28.6	39.9	18	48.7	15.4	31.0	42.4	19	128.0	168.3
VSE++ [45]	13.8	34.6	46.1	13	48.4	20.8	37.6	47.8	12	108.3	200.6
DualEn [5]	12.7	32.0	43.8	15	52.7	18.7	37.2	45.7	15	142.6	190.0
HGR [4]	16.4	38.3	49.8	11	49.2	23.0	42.2	53.4	8	77.8	223.2
HCGC [1]	17.4	39.6	52.6	9	42.1	24.2	47.9	56.4	6	77.4	238.1
QAMF [17]	18.5	40.7	51.2	10	50.4	23.7	48.4	55.3	8	80.2	237.8
Our MCKD	**19.1**	**41.8**	**54.3**	**8**	**40.7**	**25.7**	**49.2**	**59.8**	**5**	**63.9**	**249.9**

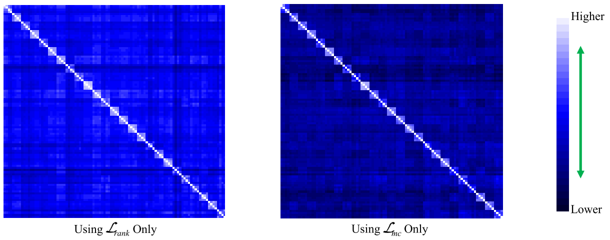


Fig. 5. We extract feature representations from randomly selected 100 video–text pairs in MSR-VTT, only using the $\mathcal{L}_{rank}$ model and only using the $\mathcal{L}_{mc}$ model under Stage II, respectively.

$\mathcal{L}_{mc}$ achieves more promising results. Since the $\mathcal{L}_{rank}$ focuses on the distance between global representations of video and text, there is ambiguity in semantic boundaries between texts describing different videos, which may bring a spurious correlation between texts and their non-corresponding videos. As we expected, the $\mathcal{L}_{mc}$ can maximize the consistency of mutual information between the intrinsic co-occurrence modalities to bridge the heterogeneous gap.

- In Stage II, only using $\mathcal{L}_{rank}$ or $\mathcal{L}_{mc}$ continuously outperforms Stage I. Even better than some SoTA baselines, which are DualEn [5], S^2Bin [46], and HGR [4]. Furthermore, compared with models only using $\mathcal{L}_{rank}$ or $\mathcal{L}_{mc}$, the full MCKD with two losses provides higher performance, which indicates that multimodal contrastive loss can maintain the reliable yet discriminative visual-text joint semantic space and transfer the reliable and structural inter-instance information.

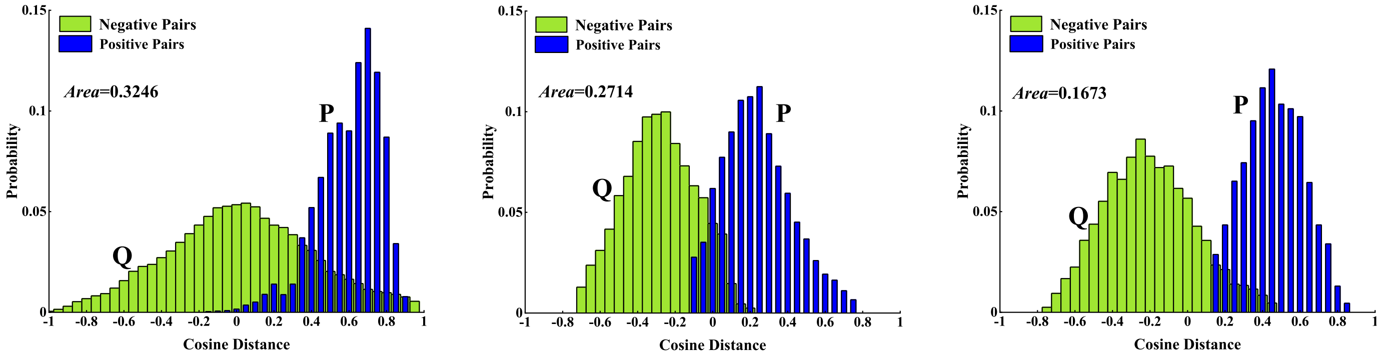
C. Dual-Loss

To answer **RQ3**, we compare the distribution of video-text feature representation from dual-loss (*a.k.a.*, $\mathcal{L}_{rank}$ and $\mathcal{L}_{mc}$), to investigate that the $\mathcal{L}_{mc}$ can learn intra-modal discriminative feature representations and transfer the reliable and structural inter-instance information for $\mathcal{L}_{rank}$. As shown in Fig. 5 and Fig. 6, we can draw the following two observations:

- In training Stage II, we randomly selected 100 video-text pairs from the MSR-VTT dataset and used $\mathcal{L}_{rank}$ and $\mathcal{L}_{mc}$ to extract the feature representations, respectively. Meanwhile, as shown in Fig. 5, Pearson's correlation visualization was carried out for the feature representation of video and text two modalities. Namely, the lower Pearson's correlation between two modalities of feature representations indicates higher orthogonality. Since the proposed $\mathcal{L}_{rank}$ explicitly takes into account the distance of inter-instance sample, we observe that after $\mathcal{L}_{rank}$ training, the Pearson's correlation between two modalities of feature representations is small. In effect, $\mathcal{L}_{rank}$ encourages the model to find fine-grained details information, to distinguish "video/text group" with similar semantics.
- In addition, to demonstrate the distribution of positive sample pairs and negative sample pairs in the semantic space provided by different loss functions, following the prior work [19], we also quantitatively visualize the distribution P of intra-instance similarity and the distribution Q of inter-instance similarity on MSR-VTT. Since the ambiguous semantic boundaries between texts to describe different videos can bring spurious correlation between texts and their non-corresponding videos, therefore, only using $\mathcal{L}_{rank}$ will achieve a relatively large margin between the positive pairs and negative pairs (that is, there exist many "hard" negative pairs with high similarity in visual-text joint semantic space), as shown in Fig. 6 (a). To be specific, we leverage the quantitative indicator function (defined in [19], lower is

TABLE VI
COMPARISON WITH PREVALENT BASELINES IN TERMS OF TECHNICAL COMPONENTS AND FUNCTIONAL MODULES

Methods		Multi-granularity Fusion	Video/Text Group	Knowledge Distillation	Consistency Learning	
					Inter-modal	Intra-modal
Graph-free Paradigm	VSE [44]	No	No	No	Yes	No
	VSE++ [45]	No	No	No	Yes	No
	W2VV [30]	No	No	No	Yes	No
	DualEn [5]	Yes	No	No	Yes	No
	S^{2}Bin [46]	No	No	No	Yes	Yes
	CE [6]	No	No	No	Yes	No
	DualEn* [31]	Yes	No	No	Yes	No
	PSM [15]	Yes	No	No	Yes	No
	T2VLAD [11]	Yes	No	No	Yes	No
Graph-based Paradigm	ViSERN [32]	No	No	No	Yes	No
	HGR [4]	Yes	No	No	Yes	No
	HANet [16]	Yes	No	No	Yes	No
	QAMF [17]	Yes	No	No	Yes	No
	HCGC [1]	Yes	No	No	Yes	Yes
Our MCKD (Graph-free Paradigm)		No	Yes	Yes	Yes	Yes



(a) Using $\mathcal{L}_{rank}$ Only (b) Using $\mathcal{L}_{mc}$ Only (c) Full MCKD (Using Dual-Loss)

Fig. 6. The similarity (cosine distance) distribution of the positive pairs P and negative pairs Q on MSR-VTT. We show the result obtained by (a) using $\mathcal{L}_{rank}$ only, (b) using $\mathcal{L}_{mc}$ only, and (c) full MCKD model (using dual-loss, namely, with $\mathcal{L}_{rank}$ and $\mathcal{L}_{mc}$), respectively. Indicator $Area$ is calculated as the overlapping area between P and Q (defined in [19], lower is better).

better) to respectively calculate the index scores under $\mathcal{L}_{rank}$ only using, $\mathcal{L}_{mc}$ only using, and full MCKD (with $\mathcal{L}_{rank}$ and $\mathcal{L}_{mc}$) model: $Area_{(\mathcal{L}_{rank})} = 0.3246$, $Area_{(\mathcal{L}_{mc})} = 0.2714$, and $Area_{(MCKD)} = 0.1673$. That is, the extent of feature representations separability can be formalized as $Area_{(MCKD)} > Area_{(\mathcal{L}_{mc})} > Area_{(\mathcal{L}_{rank})}$ in different embedding spaces. As a result, the proposed MCKD model can provide a reliable yet discriminative semantic representation space for video-text bi-directional matching.

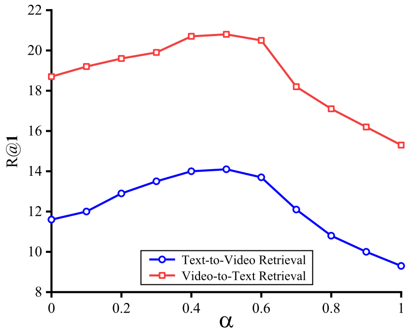


Fig. 7. Cross-modal retrieval performance of our MCKD in terms of R@1 scores versus different values of α on the MSR-VTT datasets.

D. Generalization on Unseen Dataset

To answer **RQ4**, we assess the generalization capacity of our proposed MCKD on the unseen dataset, that is, zero-shot tasks on Youtube2Text. As we all know, the most promising video-text retrieval models are mainly evaluated on test sets derived from the original dataset. However, in real scenarios, generalizing the trained model to out-of-domain (never seen) data is also a crucial index to evaluate performance.

As such, we train models on the MSR-VTT dataset and then test the trained models on the Youtube2Text testing split [22]. As shown in Table V, one can see that the performance of our MCKD on Youtube2Text is still outstanding. Particular, compared with the results in Table II, both DualEn [5] and VSE++ [45] have achieved promising performance on MSR-VTT, which hardly generalize well on a new dataset. Furthermore, a similar phenomenon can be observed in HGR [4] model and HCGC [1] model. While our MCKD can yield consistent benefits across different datasets (both in-domain and out-of-domain) compared with other methods. Since our proposal combines the advantages of two loss functions via knowledge distillation, the model can improve the generalization capacity of new compositions.

E. Parameter Analysis

To answer **RQ5**, in training Stage II of the experiment, we have tried to manually adjust the weight ratios between $\mathcal{L}_{rank}$ and $\mathcal{L}_{mc}$ at a step size of 0.1, such as 0.1 and 0.9, 0.2 and 0.8, *etc.*, to evaluate the impact of the trade-off hyper-parameter ratio. As shown in Fig. 7, the results show that our method can obtain stable performance in a relatively dense range (*i.e.* 0.4 and 0.6, 0.5 and 0.5, 0.4 and 0.6) on the MSR-VTT dataset. Therefore, the default weight hyper-parameter ratio of 1:1 is adopted in our work.

VI. Conclusion

In this paper, we propose the MCKD model for instance-level video-text retrieval, which generates a reliable and structural inter-sample soft-label signal by multimodal contrastive loss and then transfers this signal to guide the pair-wise ranking loss. Compared to models that only adopt pair-wise ranking loss, our MCKD combines the advantages of two-loss functions via a knowledge distillation module, which can capture the semantic relationship between the textual features and their corresponding visual features to build a robust yet discriminative representation space. Extensive experiments on four public datasets demonstrate the strength of our MCKD, and we report competitive results compared with the SoTA counterparts.

Appendix A
Comparison With the Baseline Methods

In order to facilitate the comparison of our MCKD model with the baseline methods, we briefly list the differences in some technical components and functional modules. The comparison involves the type of paradigm, the multi-granularity fusion, the video/text group, the knowledge distillation, and the consistency learning. As shown in Table VI, one can draw the following observations:

- Our MCKD model belongs to the graph-free paradigm, which does not rely on hand-crafted multi-granularity representation fusion or complex graph reasoning, it bridges the heterogeneous gap of cross-modal data via an end-to-end framework.
- Our MCKD model combines inter-modal and intra-modal consistency learning through knowledge distillation to construct a robust yet discriminative visual-text joint semantic space, enabling better-curated and reliable positive and negative pairs.

References

[1] W. Jin, Z. Zhao, P. Zhang, J. Zhu, X. He, and Y. Zhuang, “Hierarchical cross-modal graph consistency learning for video-text retrieval,” in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1114–1124.

[2] A. Araujo and B. Girod, “Large-scale video retrieval using image queries,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1406–1420, Jun. 2018.

[3] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, “VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4581–4591.

[4] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, “Fine-grained video-text retrieval with hierarchical graph reasoning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10638–10647.

[5] J. Dong et al., “Dual encoding for zero-example video retrieval,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 9346–9355.

[6] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, “Use what you have: Video retrieval using representations from collaborative experts,” 2019, *arXiv:1907.13487*.

[7] X. Yang, J. Dong, Y. Cao, X. Wang, M. Wang, and T.-S. Chua, “Tree-augmented cross-modal encoding for complex-query video retrieval,” in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1339–1348.

[8] I. Croitoru et al., “TeachText: CrossModal generalized distillation for text-video retrieval,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11583–11593.

[9] J. Gao and C. Xu, “Learning video moment retrieval without a single annotated video,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1646–1657, Mar. 2022.

[10] H. Zhang et al., “Video corpus moment retrieval with contrastive learning,” in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 685–695.

[11] X. Wang, L. Zhu, and Y. Yang, “T2VLAD: Global-local sequence alignment for text-video retrieval,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5079–5088.

[12] X. Chang, Y. Yang, A. G. Hauptmann, E. P. Xing, and Y.-L. Yu, “Semantic concept discovery for large-scale zero-shot event detection,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2234–2240.

[13] A. Habibian, T. Mensink, and C. G. M. Snoek, “Composite concept discovery for zero-shot video event detection,” in *Proc. Int. Conf. Multimedia Retr.*, Apr. 2014, pp. 17–24.

[14] J. Dong et al., “Reading-strategy inspired visual representation learning for text-to-video retrieval,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5680–5694, Aug. 2022.

[15] H. Liu, R. Luo, F. Shang, M. Niu, and Y. Liu, “Progressive semantic matching for video-text retrieval,” in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5083–5091.

[16] P. Wu, X. He, M. Tang, Y. Lv, and J. Liu, “HANet: Hierarchical alignment networks for video-text retrieval,” in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3518–3527.

[17] W. Ma, Q. Chen, F. Liu, T. Zhou, and Z. Cai, “Query-adaptive late fusion for hierarchical fine-grained video-text retrieval,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 25, 2022, doi: 10.1109/TNNLS.2022.3214208.

[18] M. Li and H. Wang, “Unsupervised deep cross-modal hashing by knowledge distillation for large-scale cross-modal retrieval,” in *Proc. Int. Conf. Multimedia Retr.*, Aug. 2021, pp. 183–191.

[19] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, “Dual-path convolutional image-text embeddings with instance loss,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, 2020.

[20] J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A large video description dataset for bridging video and language,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.

[21] Y. Li et al., “TGIF: A new dataset and benchmark on animated GIF description,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4641–4650.

[22] S. Guadarrama et al., “YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2712–2719.

[23] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, “Cross-media semantic representation via bi-directional learning to rank,” in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 877–886.

[24] P. Hu, X. Peng, H. Zhu, L. Zhen, and J. Lin, “Learning cross-modal retrieval with noisy labels,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5403–5413.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[26] J. Tang and K. Wang, “Ranking distillation: Learning compact ranking models with high performance for recommender system,” in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2289–2298.

[27] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Mar. 2021.

[28] H. Hu, L. Xie, R. Hong, and Q. Tian, "Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3123–3132.

[29] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 471–487.

[30] J. Dong, X. Li, and C. G. M. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3377–3388, Dec. 2018.

[31] J. Dong et al., "Dual encoding for video retrieval by text," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4065–4080, Aug. 2021.

[32] Z. Feng, Z. Zeng, C. Guo, and Z. Li, "Exploiting visual semantic reasoning for video-text retrieval," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1005–1011.

[33] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" 2020, *arXiv:2005.10243*.

[34] J. A. Portillo-Quintero, J. C. Ortiz-Bayliss, and H. Terashima-Marín, "A straightforward framework for video retrieval using CLIP," in *Proc. Mex. Conf. Pattern Recognit.* Berlin, Germany: Springer, 2021, pp. 3–12.

[35] T. Xu, X. Liu, Z. Huang, D. Guo, R. Hong, and M. Wang, "Early-learning regularized contrastive learning for cross-modal retrieval with noisy labels," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 629–637.

[36] J. Liu, M. Yang, C. Li, and R. Xu, "Improving cross-modal image-text retrieval with teacher-student learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3242–3253, Aug. 2021.

[37] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Conf. Neural Inf. Process. Syst.*, 2020, pp. 9912–9924.

[38] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2020, pp. 776–794.

[39] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," 2020, *arXiv:2010.00747*.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[42] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[43] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[44] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*.

[45] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*.

[46] M. Qi, J. Qin, Y. Yang, Y. Wang, and J. Luo, "Semantics-aware spatial–temporal binaries for cross-modal video retrieval," *IEEE Trans. Image Process.*, vol. 30, pp. 2989–3004, 2021.

[47] A. Frome et al., "Devise: A deep visual-semantic embedding model," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[48] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," 2015, *arXiv:1511.06361*.

[49] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 7–16.

[50] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1979–1988.

[51] X. Li, C. Xu, G. Yang, Z. Chen, and J. Dong, "W2VV++: Fully deep learning for ad-hoc video search," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1786–1794.

Wentao Ma received the B.S. degree in electronic science and technology from Northwestern Polytechnical University (NPU), Xi'an, China, in 2017, and the M.S. degree in information and communication engineering from the Central South University of Forestry and Technology (CSUFT), Changsha, China, in 2020. He is currently pursuing the Ph.D. degree in computer science and technology from the National University of Defense Technology (NUDT), Changsha. His main research interests include multimodal representation learning and cross-modal retrieval. He is the recipient of Outstanding M.S. Dissertation Award of Hunan, China.

Qingchao Chen received the B.S. degree in telecommunication engineering from the Beijing University of Post and Telecommunication (BUPT), Beijing, China, in 2013, and the Ph.D. degree from University College London (UCL), London, U.K, in 2018. He is currently an Assistant Professor with the National Institute of Health Data Science, Peking University. His main research interests include radar sensor design and processing, computer vision and machine learning, and multimodal learning for clinical applications.

Tongqing Zhou received the B.S., M.S., and Ph.D. degrees in computer science and technology from the National University of Defense Technology (NUDT), Changsha, in 2012, 2014, and 2018, respectively. He is currently an Assistant Researcher with the College of Computer, NUDT. His main research interests include crowdsensing and data privacy. He is the recipient of Outstanding Ph.D. Dissertation Award and Outstanding Postdoctoral, both of Hunan, China.

Shan Zhao received the Ph.D. degree from the College of Computer, National University of Defense Technology (NUDT), Changsha, China, in 2021. He is currently an Associate Professor with the Hefei University of Technology (HFUT), China. He has published several papers in refereed journals and conferences. His research interests include natural language processing and multimodal information extraction.

Zhiping Cai received the B.S., M.S., and Ph.D. degrees in computer science and technology from the National University of Defense Technology (NUDT), Changsha, in 1996, 2002, and 2005, respectively. He is currently a Full Professor with the College of Computer, NUDT. His current research interests include network security and big data. He is a Senior Member of the China Computer Federation. His Ph.D. dissertation received the Outstanding Dissertation Award of the Chinese PLA.