Debiased Video-Text Retrieval via Soft Positive Sample Calibration

Huaiwen Zhang, *Member, IEEE*, Yang Yang, *Graduate Student Member, IEEE*, Fan Qi, Shengsheng Qian, *Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

***Abstract*— With the emergence of enormous videos on various video apps, semantic video-text retrieval has become a critical task for improving the user experience. The primary paradigm for video-text retrieval learns the semantic video-text representations in a common space by pulling the positive samples close to the query and pushing the negative samples away. However, in practice, the video-text datasets contain only the annotations of positive samples. The negative samples are randomly drawn from the entire dataset. There may exist soft positive samples, which are sampled as negatives but share the same semantics as positive samples. Indiscriminately enforcing the model to push all the negative samples away from the query leads to inaccurate supervision and then misleads the video-text feature representation learning. In this paper, we introduce debiased video-text retrieval objectives that calibrate the punishment of soft positive samples. In particular, we propose a novel uncertainty measure framework to estimate the credibility of negative samples for each instance. Then, the reliability of negative samples is used to find the soft positive samples and rescale their contribution within video-text retrieval losses, including triplet loss and contrastive loss. Experimental results on five widely used datasets demonstrate that our debiased video-text retrieval objectives achieve significant performance improvements and establish a new state-of-the-art.**

***Index Terms*— Video-text retrieval, debias, soft positive samples.**

Manuscript received 23 September 2022; revised 1 January 2023; accepted 13 February 2023. Date of publication 24 February 2023; date of current version 6 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62206137, Grant 62206200, Grant 62276257, Grant 62036012, and Grant 62066033; in part by the National Natural Science Foundation of Inner Mongolia under Grant 2022MS06025; in part by the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region under Grant NJYT23105; and in part by the Applied Technology Research and Development Program of Inner Mongolia Autonomous Region under Grant 2020GG0046, Grant 2021GG0158, and Grant 2020PT0002. This article was recommended by Associate Editor J. Shen. *(Corresponding author: Shengsheng Qian.)*

Huaiwen Zhang and Yang Yang are with the College of Computer Science, Inner Mongolia University, Hohhot 010021, China, and also with the National and Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian and the Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Hohhot 010021, China (e-mail: huaiwen.zhang@imu.edu.cn; yangyang@mail.imu.edu.cn).

Fan Qi is with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China (e-mail: fanqi@mail.tjut.edu.cn).

Shengsheng Qian and Changsheng Xu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shengsheng.qian@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3248873.

Digital Object Identifier 10.1109/TCSVT.2023.3248873

I. Introduction

NOWADAYS, people are swamped with the massive volume of videos provided by various video apps, such as YouTube and TikTok. To improve the user experience, video retrieval, as a critical topic in the fields of information retrieval and video technology, has attracted increasing research interest.

The critical challenge of video-text retrieval is that the distributions and representations of videos and texts are inconsistent, making it difficult to measure the similarity between different modalities. To tackle this problem, the dominant approaches [1], [2], [3], [4], [5], [6], [7] for video-text retrieval firstly encode different modalities into a common representation space, and then leverage a suitable distance metric to measure the semantic similarities. With such a paradigm, some recent work has focused on designing more complicated encoder [1], [6], [7] to obtain better representations for modalities or more sophisticated matching strategies [2], [3], [5], [8]. For example, Gabeur et al. [2] focus on the multi-modality information in the videos and incorporate multi-layer transformers to learn strong video features. Chen et al. [3] introduce a hierarchical video-text encoder, which factorizes video-text matching into hierarchical levels, including events, actions, and entities.

Existing approaches have achieved remarkable performance by leveraging the strong representation ability of deep neural networks. Meanwhile, few of them are aware that the data preparation process of video-text retrieval brings biases. Annotators of video-text datasets are required to describe the entire untrimmed video in a few sentences [9], [10], [11], [12], [13], resulting in general annotations that disregard video specifics and can be paired with multiple videos. Furthermore, the video-text retrieval datasets only contain annotations for positive video-text pairs, i.e., video v_i and text t_i match in semantics, with no labeled negative pairs, i.e. video v_i and text t_j do not match.

Existing methods [1], [2], [3], [4], [6], [7], [14], [15], [16], [17] randomly sample negatives from the whole distribution, which contains inescapable noise. There are samples that are sampled as negatives but have semantics comparable to the query, termed as "soft positive samples". As shown in Fig.1, given a query t_0 "A man is singing and playing the guitar", only the annotated video v_0 is treated as a positive sample. Except for that, all the other samples are regarded as negative, although some of them (e.g., v_1^- and v_2^-) can also match the query t_0 perfectly.

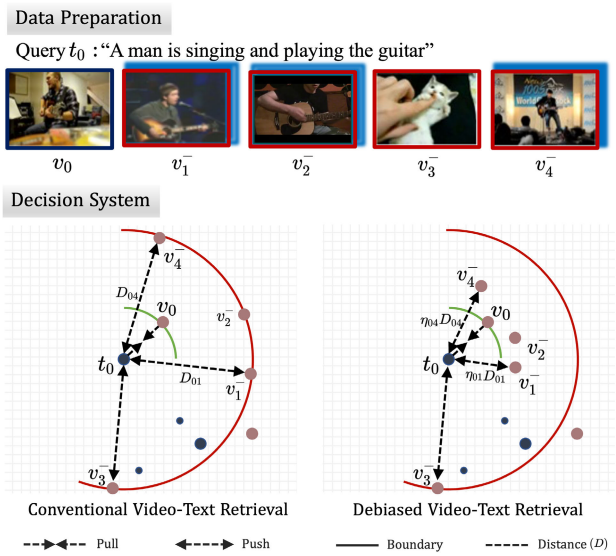

Fig. 1. Training data for video-text retrieval. The dataset only contains positive pair (v_0^+, t_0). The negative samples (v_*^-) are randomly sampled over the whole distribution, which may contain biases. For example, the negative ones in the blue shadow (v_1^-, v_2^-, v_4^-) also have close semantics to the query, which we refer to as "soft positive samples." Indiscriminately enforcing the model to maximize the distance between the query and these soft positive samples leads to biased supervision in optimization. By using our debiased video-text retrieval, the soft positive samples will receive the appropriate supervision based on their uncertainty scores η_{ij}, resulting in a reasonable semantic space for cross-modal matching.

The conventional video-text retrieval methods push all the negative samples away from the query t_0, including soft positive samples v_1^-, v_2^-, and v_4^-. However, indiscriminately enforcing the model to maximize the distance between the query and soft positive samples leads to inaccurate supervision, misleads video-text representation learning, and severely disrupts common space.

To tackle the above challenges, we introduce the Debiased Video-Text Retrieval (DVTR) method that calibrates the biased supervision of soft positive samples. We first introduce the video-text matching uncertainty estimation module, which evaluates the uncertainty [18] of the query and candidate samples to identify the soft positive samples. Specifically, a novel hierarchical probabilistic encoder is introduced for video-text pairs to map them into probabilistic embedding [19]. Then, the heterogeneity-aware multi-modal uncertainty learning strategy is presented to estimate the matching uncertainty of a given video-text pair. The uncertainty is defined as the probability of semantic mismatch of a given video-text pair. A lower uncertainty means the candidate sample has a higher semantic similarity with the query sample. As shown in Fig.1, video v_1^-, v_2^-, and v_4^- may be identified as soft positive samples of query t_0, if the estimated uncertainty score η_{01}, η_{02}, η_{04} are small. Then, we propose the debiased video-text representation learning module, in which we weightedly reduce the penalty of soft positive samples in video-text retrieval losses by their uncertainty scores to address the biased supervision. As shown in Fig.1, the distance between query t_0 and v_1^-, v_2^-, and v_4^- are rescaled by their uncertainty score. In this way, our representation learning module can precisely capture the semantics of videos and texts. Note that the proposed debiased video-text retrieval objectives can be applied to any state-of-the-art video-text retrieval model by simply adjusting their losses. To verify the effectiveness of our proposed method, we conduct extensive experiments on five widely used video-text retrieval benchmark datasets, MSRVTT [9], MSVD [10], VATEX [11], ActivityNet [13] and LSMDC [12]. Extensive experiments with state-of-the-art performance demonstrate the effectiveness of our debiased objectives.

In brief, our contributions can be summarized as follows:

- We propose the novel Debiased Video-Text Retrieval (DVTR) method to alleviate the biased supervision of soft positive samples. DVTR can be integrated into most video retrieval models for better retrieval performance with a few computational costs at training time and no additional time consumption at test time.
- We introduce the uncertainty estimation module to identify the soft positive samples, which precisely estimate the uncertainty of video-text pairs by the novel hierarchical probabilistic encoders and heterogeneity-aware multi-modal uncertainty learning strategy.
- We present the debiased video-text representation learning objectives, in which we calibrate the biased supervision of soft positive samples by reducing their penalty in proportion to their uncertainty scores.
- Extensive experimental results on five widely used datasets demonstrate that our debiased video-text retrieval method achieves significant performance improvements and establishes a new state-of-the-art in video-text retrieval.

II. Related Work

In this section, we briefly review the previous methods most relevant to our work, including video-text retrieval, bias in video-text retrieval, and uncertainty estimation.

A. Video-Text Retrieval

Video-text retrieval aims to perform effective semantic retrieval across video and text data. Recently, semantic video-text retrieval methods [1], [2], [3], [4], [5], [6], [7] aim to develop a powerful encoder to map video and text to a shared embedding space or more sophisticated matching strategies to align video and text at different levels. For example, Gabeur et al. [2] introduce a transformer-based encoder architecture that aggregates multiple modality features extracted from videos to build effective video representations. Dong et al. [1] propose a novel dual network that exploits multi-level encodings to obtain global, local, and temporal patterns in videos and sentences. However, they still suffer from the scale of the training data and have a poor retrieval performance. Most recently, with the promising performance of CLIP [20], some clip-based methods [6], [7], [21] achieve incredibly high performance and outperform other methods by a large margin. Some approaches [17], [22] focus on the large-scale datasets. For example, Bain et al. [17] propose an end-to-end trainable model that is designed to use large-scale image and video captioning datasets for video-text retrieval. Ko et al. [22] introduce a multi-modal self-supervised framework to capture significant information from noisy and weakly correlated large-scale datasets by using a variant of dynamic

time warping. Additionally, some recent works [23], [24] focus on splitting video and text into fine-grained levels to bridge semantic gaps in visual and textual. For example, inspired by the reading strategy of humans, Dong et al. [23] propose a reading-strategy inspired visual representation learning to represent videos. Zhang et al. [24] propose a local-global graph pooling network to disentangle the video and text into four levels with the graph neural networks and exploit a hierarchical pooling strategy to maximize the mutual information between pool features and the corresponding graph node features.

Different from the above work, this paper focuses on addressing the biased supervision of soft positive samples in video-text retrieval representation learning rather than building more complex feature extractors or matching strategies. Furthermore, this work can be equipped with state-of-the-art methods to further improve its performance.

B. The Bias in Video-Text Retrieval

Recently, the bias in video representation learning is attracting more and more attention. Many efforts [4], [25], [26], [27], [28], [29], [30], [31] have been proposed to address a variety of biases. Some approaches [4], [26] focus on the bias of strict assumption of video-text retrieval, i.e., only a single text is relevant to a query video and vice versa [25]. Hence, some approaches [4], [25], [26], [27] are proposed to model the one-to-many or many-to-many correspondences in the retrieval task. For example, Patrick et al. [4] introduce a multi-modal cross-instance text generation task as the auxiliary to extract the inner one-to-many correspondences of instances for video-text retrieval. Chun et al. [26] encode image and text into probability distributions of concepts, and implicitly perform many-to-many matching between those concepts. Some recent work [30], [32], [33] making effort to alleviate biases in other areas. For example, Cheng et al. [32] explore the effectiveness of various video features on visual search and test different search strategies over different types of queries. Liu et al. [33] introduce the cross-modal semantic importance consistency to achieve invariance in the semantics of items during cross-modal aligning. It measures the semantic importance of items and learns a more reasonable representation vector by inter-calibrating the importance distribution. Yang et al. [30] present a novel contrastive self-supervised loss to update features of the foreground in a noise-free manner for instance segmentation. It considers the different roles of noisy labels in different subtasks' loss.

In this paper, we observe the biased supervision of soft positive samples. We introduce Debiased Video-Text Retrieval (DVTR) method, which tackles the bias by directly identifying the soft positive samples in negative samples and correcting their punishment.

C. Uncertainty Estimation

Uncertainty estimation aims to capture what a model does not know or is not confident. Various aspects of uncertainty have been explored [18], [19], [34], [35], [36], including the data-dependent uncertainty, model uncertainty, and the uncertainty on annotation. For example, Oh et al. [19] learn the uncertainty of image representation to obtain a more robust embedding. Chang et al. [34] learn the uncertainty of input data to alleviate the influence of observation noise for better network optimization. Zheng and Yang [35] estimate the uncertainty of the predicted pseudo labels for semantic segmentation. Uncertainty also has much research attention in other areas [37], [38], [39], [40]. Kim et al. [37] consider two types of uncertainty in multispectral pedestrian detection to alleviate the miscalibration of image pairs. Cheng et al. [38] argue that the unlabeled data in deep semi-supervised hashing methods is not always reliable. They introduce the uncertainty-aware and multi-granularity consistent constrained semi-supervised hashing method to alleviate the negative effects of noisy supervised signals, where the uncertainty score is estimated by Monte Carlo dropout. Kim et al. [39] introduce class uncertainty-aware loss for object detectors, in which the uncertainty score of classification is used to modulate the detector loss function. Su et al. [40] introduce an uncertainty-aware loss function in multi-view stereo scenario to measure the reliability of the estimated depth map.

In this work, we explore the reliability estimation of the randomly sampled negative pairs and alleviate the biased supervision of soft positive samples by re-weighting their contributions in video-text representation learning. To the best of our knowledge, this is the first work to utilize uncertainty to resist biases in video-text retrieval.

III. Problem Formulation

Given a video-text retrieval dataset $\mathcal{D} = \{(v_i, t_i)\}_{i=1}^{N}$ of N video-text pairs, where the i-th pair (v_i, t_i) is composed of video v_i and corresponding caption t_i, the task of video-text retrieval is to retrieve the videos (or texts) whose semantics are similar to query text (or video). Here, we take v as a query and t as a candidate, as an example. The primary paradigm of video-text retrieval [1], [2], [3], [4], [6], [7], [14], [15], [16] is to encode the different modalities into a common representation space, then leverage distance metrics to directly compare the semantic similarity of video and text. Specifically, the cross-modal common space is built by ranking losses, in which the representational similarity between a given query v_i and its positive samples $\mathcal{D}_{t_i}^{p}$ is maximized while the similarity with negative samples $\mathcal{D}_{v_i}^{n}$ is minimized.

We denote the uncertainty [18], [19] that v_i and t_j have similar semantic content as $\eta_{ij} \in [0, 1]$. Ideally, none of the negative samples match the query, i.e., $\eta_{ij} = 1, \forall t_j \in \mathcal{D}_{v_i}^{n}$. Unfortunately, there are no negative pair annotations in datasets, standard approaches thus sample negatives t_j for given query v_i from the whole distribution of text $\{t_j\}_{j\neq i}^{N}$ instead. There are negative samples that are sampled as negatives but have semantics comparable to the query ($\eta_{ij} < 1$), termed as "soft positive sample". Indiscriminately enforcing the model to maximize the distance between the query and soft positive samples leads to biased supervision, misleads video-text representation learning. To this end, we propose our debiased video-text retrieval method to mitigate the biased supervision of soft positive samples in the following section.

IV. Methodology

A. Model Overview

To address the biased supervision of soft positive samples in video-text representation learning we propose our Debaised

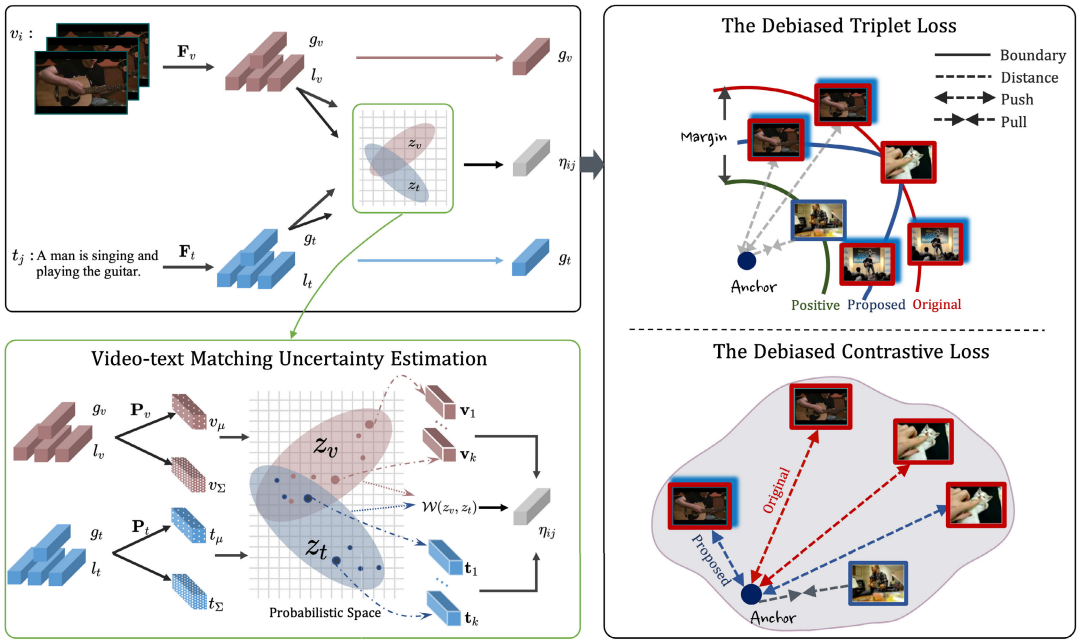

Fig. 2. The framework of the debiased video-text retrieval network. The video-text matching uncertainty estimation module takes global- and local-level representations of each pair (v_i, t_j) as input to obtain its uncertainty score η. Specifically, the pair (v, t) first goes through the feature encoders $\mathbf{F}_v$ and $\mathbf{F}_t$ to extract features. Then the probabilistic encoders $\mathbf{P}_v, \mathbf{P}_t$ project each modality into the probabilistic embeddings z_v and z_t. By joint considering the discrepancy of contributions and the distance of sampled point embeddings, we approximate the uncertainty η_{ij}. The uncertainty is then employed to guide the video-text representation learning by reducing the penalty of the soft positive sample with η_{ij}.

Video-Text Retrieval (DVTR) method. The overview architecture is illustrated in Fig.2. The framework consists of the following components:

- **Basic Video-text Retrieval Backbone**: We construct a basic retrieval backbone to apply video-text retrieval. Specifically, we employ visual and textual encoders to extract video and text representations. Then, by minimizing the ranking loss functions, we could construct a cross-modal common space in which the representation distance between a given query (video or text) and its positive samples is minimized while the distance with negative samples is maximized. Unfortunately, such basic retrieval method can not handle the biased supervision introduced by the soft positive samples.
- **Video-text Matching Uncertainty Estimation**: To identify the soft positive samples, we propose a video-text matching uncertainty estimation module to measure the uncertainties between query and negative samples. We propose a hierarchical probabilistic encoder to map video and text as probability embeddings. In addition, we propose a heterogeneity-aware multi-modal uncertainty learning strategy to comprehensively measure the discrepancy of multi-modal probabilistic distributions. Based on the estimated video-text matching uncertainty, we could precisely detect the soft positive samples.
- **Debiased Video-text representation learning**: The debiased video-text representation learning aims to calibrate the inappropriate supervision for soft positive samples. The uncertainty scores between the query and each negative are first measure by the video-text matching uncertainty estimation module. Based on the estimated uncertainties, we modify the frequently used two ranking loss functions by weighted reducing the penalty of soft positive samples with their uncertainty score.

TABLE I
THE MAIN NOTATIONS OF THIS PAPER AND THEIR EXPLANATIONS

Symbol	Explanation
$\mathcal{D}$	Video-text retrieval dataset.
(v_i, t_i)	The i-th video-text pair in dataset.
S_{ij}	Similarity between i-th video and j-th text.
τ	Parameter of temprature in contrastive loss function.
λ	Parameter of margin in triplet loss function.
η_{ij}	Probability of i-th video and j-th text contain similar semantics.
$\mathbf{F}_v, \mathbf{F}_t$	Video and text encoder for extracting representations from raw data.
l_v, g_v	Local and global representations of video v
l_t, g_t	Local and global representations of text t
$\mathbf{P}_v, \mathbf{P}_t$	Probabilistic encoders of video and text.
z_v, z_t	Video and text probabilistic embeddings.
$\mathbf{v}, \mathbf{t}$	Sampled embedding from video and text probabilistic embeddings.
$\mathcal{N}(v_\mu, v_\Sigma)$	Video probabilistic distributions with mean v_μ and variance v_Σ.
$\mathcal{N}(t_\mu, t_\Sigma)$	Text probabilistic distributions with mean t_μ and variance t_Σ.
$\mathcal{W}(z_v, z_t)$	Wasserstein distance between distribution z_v and z_t.
θ_*	Parameter for the specific module.

We introduce the basic video-text retrieval module in Sec.IV-B. The proposed video-text matching uncertainty estimation module is introduced in Sec.IV-C. The debiased loss functions are detailed in Sec.IV-D. Finally, we elaborate on the training and inference flow in Sec.IV-E. The math notations of this paper are summarized in Tab.I.

B. Basic Video-Text Retrieval Backbone

Given a video v and text t, the basic video-text retrieval module aims to encode video and text into the common representation space. The local and global features of video-text

pairs are then generated and fed into the video-text uncertainty estimation module.

1) Video Encoder $\mathbf{F}_v(v;\theta_{\mathrm{F}_v})$*:* Given a video v, the video encoder is used to learn the local- and global-level representations $(l_v, g_v) = \mathbf{F}_v(v;\theta_{\mathrm{F}_v})$, where θ_{F_v} is the parameter of the video encoder. l_v and g_v are the local- and global-level video representations, respectively. $l_v = \{l_v^1, l_v^2, \ldots, l_v^N\}$ N denotes the frame number. The global video representations g_v is obtained by applying a pooling strategy to l_v. Specifically, the video encoder is designed as transformer based architecture. For HiT [16], we use the outputs of the feature-level layer as the local-level representations of video and conduct the mean pooling of the semantic-level to aggregate the global-level representations. For CLIP2Video [7], the vision transformer (ViT) [41] is adopted to encode every frame into features and combines the temporal and spatial information to generate the local-level representations. Following [7], we then apply global average pooling to encode final global-level representations.

2) Text Encoder $\mathbf{F}_t(t;\theta_{\mathrm{F}_t})$*:* Given a text t, text encoder $\mathbf{F}_t(t;\theta_{\mathrm{F}_t})$ is used to encode it as a local- and global-level representation (l_t, g_t), where θ_{F_t} denotes the learnable parameter. Specifically, we adopt the base BERT [42] as the text encoder and fine-tune it. For HiT [16], we use the outputs of the word-level layer as the local-level representations l_t, and perform the average pooling for the semantic-level to aggregate the global-level representations g_t. And for CLIP2Video [7], we obtain the local-level features l_t from the hidden states of BERT and take the highest number in each hidden state as the global-level features g_t.

3) Video-Text Matching: Based on the above video and text encoder, the parameter θ_{F_v} and θ_{F_t} are updated by minimizing Eq.1 or Eq.2 as follows:

$$\mathcal{L}_{TL} = \frac{1}{N}\sum_{i=1}^{N}[S_{ij}^{-} - S_{ii}^{+} + \lambda]_{+} \tag{1}$$

$$\mathcal{L}_{CL} = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{e^{(S_{ii}^{+}/\tau)}}{\Sigma_{k=1}^{B} e^{(S_{ik}/\tau)}}\right] \tag{2}$$

where S_{ij} is the similarity of v_i and t_j, S^{+} are positive pairs, S^{-} are negative pairs, λ is the margin, $[\cdot]_{+} = \mathrm{Max}\{\cdot, 0\}$. λ and τ are the margin and temperature parameters, respectively. B is the batch size.

C. Video-Text Matching Uncertainty Estimation

In this section, we introduce the video-text matching uncertainty estimation module, which evaluates the uncertainty η_{i*}^{-} of the query sample v_i and candidate samples t_* by encoding the video and text into the probabilistic embedding space [19]. Different from common point embedding methods which project input x into an embedding vector z with fixed dimension d, i.e., a point in $\mathcal{R}^d$, probabilistic embedding maps the input x into a probabilistic distribution in $\mathcal{R}^d$ which not only preserves the semantic information but also captures the inherent uncertainty [18] in data. We build the video-text matching uncertainty estimation module by extending probabilistic embedding to the video-text retrieval scenario.

1) Hierarchical Probabilistic Encoders: Given a pair (v_i, t_j) consisting of i-th video and j-th text (for clarity we omit i and j in this section), we first propose the video probabilistic encoder $\mathbf{P}_v(l_v, g_v;\theta_{\mathrm{P}_v})$ and text probabilistic encoder $\mathbf{P}_t(l_t, g_t;\theta_{\mathrm{P}_t})$ to project the local- and global-level representation of video and text as probabilistic embedding z_v and z_t, where $\theta_{\mathrm{P}_*^*}$ denote the parameters.

The local and global features l_v, g_v of video v are first obtained by powerful feature extractors $\mathbf{F}_v(v;\theta_{\mathrm{F}_v})$ to capture the semantic information. Then the video probabilistic encoder projects them into probabilistic embedding $z_v = \mathbf{P}_v(l_v, g_v;\theta_{\mathrm{P}_v})$, where $z_v = \mathcal{N}(v_\mu, v_\Sigma)$ The mean v_μ and the variance v_Σ of video distribution z_v are obtained as follows:

$$\begin{aligned} v_\mu &= \mathbf{P}_v^{\mu}(l_v, g_v;\theta_{\mathrm{P}_v^{\mu}}) \\ &= \mathrm{LN}(g_v + \sigma(\mathrm{MLP}_v^{\mu}\,\mathrm{Attn}_v^{\mu}(l_v))) \end{aligned} \tag{3}$$

$$\begin{aligned} v_\Sigma &= \mathbf{P}_v^{\Sigma}(l_v, g_v;\theta_{\mathrm{P}_v^{\Sigma}}) \\ &= \mathrm{MLP}_v^{\Sigma}(g_v + \mathrm{Attn}_v^{\Sigma}(l_v)) \end{aligned} \tag{4}$$

where $\mathrm{LN}(\cdot)$ is the LayerNorm [43], $\mathrm{Attn}_v^{*}(\cdot)$ is the self-attention layer, MLP_v^{*} indicate the multilayer perception (MLP) layer $\sigma(\cdot)$ means the sigmoid function.

Similar to the video probabilistic encoder, the text probabilistic encoder $\mathbf{P}_t(l_v, g_v;\theta_{\mathrm{P}_t})$ attempts to encode text representations l_t and g_t as a probabilistic embedding $z_t = \mathbf{P}_t(l_v, g_v;\theta_{\mathrm{P}_t})$,where $z_t = \mathcal{N}(t_\mu, t_\Sigma)$. The local and global features l_t, g_t of text t are obtained by $\mathbf{F}_t(t;\theta_{\mathrm{F}_t})$ to capture the semantic information. The mean t_μ and the variance t_Σ of text probabilistic embedding z_t are formulated as follow:

$$\begin{aligned} t_\mu &= \mathbf{P}_t^{\mu}(l_t, g_t;\theta_{\mathrm{P}}^{\mu}) \\ &= \mathrm{LN}(g_t + \sigma(\mathrm{MLP}_t^{\mu}\,\mathrm{Attn}_t^{\mu}(l_t))) \end{aligned} \tag{5}$$

$$\begin{aligned} t_\Sigma &= \mathbf{P}_t^{\Sigma}(l_t, g_t;\theta_{\mathrm{P}}^{\Sigma}) \\ &= \mathrm{MLP}_t^{\Sigma}(g_t + \mathrm{Attn}_t^{\Sigma}(l_t)) \end{aligned} \tag{6}$$

2) Heterogeneity-Aware Multi-Modal Uncertainty Learning: Existing probabilistic embedding methods [19], [26] mainly focus on single-modal data. They estimate the probability that the semantics of z_v and z_t are matched by directly calculating the distance of points sampled from the distributions. We argue that the inherent heterogeneity gap between visual and textual modalities may lead to invalidation of such random sampling measures, especially when the sample size is small. Thus, in this work, we measure the matching probability between two probabilistic distributions by jointly considering the similarity of the point sampling from each distribution and the divergence between distributions:

$$\mathbf{D}(z_v, z_t) = \mathbf{v}^{\mathrm{T}} \cdot \frac{z_v}{z_t}\mathbf{t} \tag{7}$$

where $\mathbf{v}$ and $\mathbf{t}$ are sampled from distribution z_v and z_t by Monte-Carlo sampling. If the two probabilistic distributions are aligned well in the common space, the Eq.7 is degenerated as the standard format: $\mathbf{D}(z_v, z_t) = \mathbf{v}^{\mathrm{T}}\mathbf{t}$, which only considering the distance between point embeddings. Once the two probabilistic distributions are not aligned well, Eq.7 measures the discrepancy between probabilistic embedding by introducing the divergence of distribution.

During training, based on the similarity of distribution, the loss function is formulated as follows:

$$\begin{aligned}\mathcal{L}_U &= -E_{z_v} E_{z_t} \log(\mathbf{D}(z_v, z_t)) \\ &= -E_{z_v}[E_{z_t} \log(\mathbf{v}^{\mathrm{T}}\mathbf{t}) - \mathrm{KL}(z_t \| z_v)] \\ &\approx -\frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbf{v}_i^{\mathrm{T}}\mathbf{t}_j + \mathrm{KL}(z_t \| z_v) \end{aligned} \tag{8}$$

where K is the Monte-Carlo sampling number, the $\mathrm{KL}(\cdot\|\cdot)$ is the KL divergence. However, the KL divergence is asymmetric and there is a problem of vanishing gradient. To address it, we adopt the 2-Wasserstein distance [44] to minimize the discrepancy between z_v and z_t. As the z_v and z_t follow Gaussian distribution, the 2-Wasserstein based method is reduced to:

$$\mathcal{W}(z_v, z_t) = \|v_\mu - t_\mu\|_2^2 + \|v_\Sigma^{1/2} - t_\Sigma^{1/2}\|_2^2 \tag{9}$$

Then, we measure the uncertainty that the pair (v_i, t_j) is matched by:

$$\begin{aligned}\eta_{ij} &= \mathbf{U}(z_{v_i}, z_{t_j}; \theta_{\mathrm{U}}) \\ &= \sigma(-\mathrm{MLP}_\eta(\mathcal{W}(z_{v_i}, z_{t_j}))) \end{aligned} \tag{10}$$

where MLP_η denotes a MLP layer. σ is the sigmoid activation function. Furthermore, we optimize $\mathcal{L}_\eta$ to minimizing the matching probability of the:

$$\mathcal{L}_\eta = -\sigma(-\mathrm{MLP}_\eta(\mathcal{W}(z_v, z_t))) \tag{11}$$

Following the aforementioned definition of uncertainty, we can effectively identify the soft positive samples from negative ones since they have a close η_{ij} with the positive samples, a.k.a. the candidate video/text is highly semantically consistent with the query sample.

D. Debiased Video-Text Representation Learning

As shown in Fig.1, we observe that candidate samples with higher semantic similarity to the query have a lower uncertainty. Based on this observation, we propose a novel debiased video-text representation learning module, in which we weightedly reduce the penalty of soft positive samples in ranking losses by their uncertainty scores. Specifically, the two most commonly used ranking losses: triplet loss and contrastive loss, are altered with uncertainty.

1) The Debiased Triplet Loss: Triplet loss is commonly used in video-text retrieval to make the similarity of positive pairs S_{ii}^+ be at least λ larger than negative pairs S_{ij}^-. The conventional equation of triplet loss $\mathcal{L}_{TL}$ is shown in Eq.1.

Minimizing $\mathcal{L}_{TL}$ means maximizing the similarity of positive pairs S_{ii}^+ and minimizing the similarity of negative pairs S_{ij}^-. It is implemented by pulling the representation of the positive sample v_i close to the query t_i, and pushing away the representation of negative sample v_j from the query t_i, until the negative samples narrows the positive one with at least λ margin. However, when facing the soft positive samples, the existing models still try to push them away by imposing a significant penalty, leading to the wrong optimization direction.

In our debiased video-text representation learning, we define the debiased triplet loss as follows:

$$\mathcal{L}_{TL}^U = \frac{1}{N}\sum_{i=1}^{N}[\eta_{ij}^- S_{ij}^- - S_{ii}^+ + \lambda]_+ \tag{12}$$

Algorithm 1 Debiased Video-Text Retrieval Training

Input: dataset $\mathcal{D} = \{(v_i, t_i)\}_{i=1}^N$; max epoch number E
Output: Learned parameters θ_{F_v},θ_{F_t}, θ_{P_v},θ_{P_t}, and θ_{U}

repeat
  // Training the uncertainty estimation module;
  for *sampled batch* $\{(v_i, t_i)\}_{i=1}^B$ *from* $\mathcal{D}$ **do**
    Extract features:
    $(l_v, g_v), (l_t, g_t) = \mathbf{F}_v(v; \theta_{\mathrm{F}_v}), \mathbf{F}_t(t; \theta_{\mathrm{F}_t})$;
    Generate probabilistic embeddings:
    $z_v, z_t = \mathbf{P}_v(l_v, g_v; \theta_{\mathrm{P}_v}), \mathbf{P}_t(l_t, g_t; \theta_{\mathrm{P}_t})$;
    Calculate $\mathcal{L}_U$ and $\mathcal{L}_\eta$;
    Update θ_{F_v}, θ_{F_t}, θ_{P_v}, θ_{P_t}, and θ_{U};
  end
  // Training the debiased retrieval module;
  for *sampled batch* $\{(v_i, t_i)\}_{i=1}^B$ *from* $\mathcal{D}$ **do**
    Obtain negative pairs $\{(v_i, t_j)\}_{i=1,j=1}^B, i \neq j$;
    Calculate η_{ij} for negative pairs ;
    Calculate debiased loss $\mathcal{L}_{TL}^U$ or $\mathcal{L}_{CL}^U$;
    Update θ_{F_v}, θ_{F_t};
  end
until *convergence*;

where for the pairs with low uncertainty scores, such as soft positive samples, a smaller weight η_{ij}^- is used to reduce its penalty, leading to a smaller gradient in optimization. For the pairs with high uncertainty, the $\eta_{ij}^- S_{ij}^-$ stays high, resulting in a high gradient in optimization. Thus, the $\mathcal{L}_{TL}^U$ can effectively prevent the pairs with high semantic similarity from being separated by a far distance in embedding space.

2) The Debiased Contrastive Loss: Contrastive loss is also a widely adopted loss in video-text retrieval, which aims to make the similarity of positive pair S_{ii}^+ account for the largest proportion in the sum of similarity of all pairs in a batch $\sum_{k=1}^B S_{ik}$. The conventional equation of contrastive loss $\mathcal{L}_{CL}$ is shown in Eq.2.

By minimizing $\mathcal{L}_{CL}$, the similarity of positive pairs S_{ii}^+ will approach to 1 and the similarity of all the other negative pairs $\sum_{k\neq i}^B S_{ik}^-$ will approach to 0. It is implemented by pulling the representation of v_i and t_i as close as possible and pushing apart the representation of v_i and t_j as far as possible. Thus, the contrastive loss also cannot handle the soft positive samples well due to the conflict between pulling the similar semantic samples close to the query and pushing away the soft positive samples.

In our debiased video-text representation learning, we define the debiased contrastive loss as follows:

$$\mathcal{L}_{CL}^U = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{e^{(S_{ii}^+/\tau)}}{e^{(S_{ii}^+/\tau)} + \Sigma_{k\neq i}^B \eta_{ik}^- e^{(S_{ik}^-/\tau)}}\right] \tag{13}$$

In $\mathcal{L}_{CL}^U$, we can observe that the issue of soft positive samples can be well addressed, since the lower uncertainty η_{ik}^- the pair has, the smaller the gradient and the contribution of this pair to the optimization.

E. Training and Inference

We summarize our training algorithm in Alg.1. We first learn the uncertainty estimation model with positive pairs. Then, we train the video-text retrieval method by losses $\mathcal{L}_{TL}^{U}$ or $\mathcal{L}_{CL}^{U}$ with the contribution of negative samples calibrated by their uncertainty. Given the positive training pairs $\{(v_i, t_i)\}_{i=1}^{B}$, we first extract the local and global features of video and text by basic retrieval module. Then, these features are fed into the corresponding probabilistic encoder $\mathbf{P}_v$ and $\mathbf{P}_t$, which maps them into probabilistic embeddings. We train the video and text probabilistic encoders by minimizing $\mathcal{L}_U$ and $\mathcal{L}_\eta$. For training the debiased retrieval module, we first obtain negative pairs $\{(v_i, t_j)\}_{i\neq j}^{B}$. Then, we calculate the uncertainty η_{ij} for each negative sample. The debiased $\mathcal{L}_{TL}^{U}$ or $\mathcal{L}_{CL}^{U}$ are then optimized with each sample receiving proper supervision. In the inference stage, given a query sample, we extract its features by $\mathbf{F}_v/\mathbf{F}_t$ and sort the similarities between the query and candidates to choose the best matching samples.

V. Experiments

A. Experimental Details

1) Datasets: To achieve a comprehensive evaluation of our DVTR, we carry out our experiments on five widely used [1], [2], [3], [6], [7], [14], [16] video-text retrieval datasets with various scales and video sources. Table II summarizes the brief statistics of these datasets.

a) MSRVTT [9]: The MSRVTT dataset consists of 10K videos collected from YouTube. Each video lasts 10 to 30s and is annotated with about 20 natural sentences in English. Our results are reported on the train/test splits named Full [9] and 1k-A [45]. Following the Full split, 6,513 videos are used as the training set, 497 for validation, and 2,990 for the testing set. The 1k-A split was introduced by [45] that 9K videos are used for training, 1K for testing and validation.

b) MSVD [10]: The MSVD dataset consists of 80K English sentences for 1,970 videos from YouTube. Each video is described with around 40 sentences. Our results are reported base on the standard split that uses 1,200, 100, and 670 videos for training, validation, and testing.

c) VATEX [11]: VATEX dataset is a multilingual video-text dataset with 34,911 videos. Each video, collected from YouTube, has a duration of 10s and at least 10 English captions. In our work, we only use English annotations. We use the official split, 25,991, 3,000, and 6,000 videos for training, validation, and testing.

d) ActivityNet [13]: The ActivityNet dataset consists of 20,000 YouTube videos. We follow the setting of [51], which concatenates all the captions of a video into a paragraph, and evaluate the model on the "val1" split.

e) LSMDC [12]: The LSMDC dataset contains 118,081 videos and equal captions extracted from 202 movies, with a split of 109,673, 7,408, and 1,000 as the train, validation, and test sets. Every video is selected from movies ranging from 2 to 30 seconds.

2) Implementation Details: Follow Alg.1, we apply our debiased video-text retrieval objectives to the HiT [16] and CLIP2Video (C2V) [7] to obtain our DVTR + HiT model and DVTR + C2V model, respectively. We follow the HiT [16] to conduct feature extraction for DVTR + HiT model, including the audio features from VGGish [54], appearance features from SENet-154 [55], motion features from S3D [56]. For MSRVTT, MSVD, and LSMDC, we use all the audio, appearance, and motion pre-extracted features. For ActivityNet, we use the motion and audio pre-extracted features. For VATEX, we use motion and the official I3D [57] features. We use 30 and 25 as the frame length and caption token numbers in the DVTR model. The initial learning rate is set to 2e-5 and the network is optimized by the AdamW [58] optimizer. We use the 10% proportion of warm up and cosine decay for scheduling the learning rate. The batch size is 128 and we train 40 epochs. We follow the CLIP2Video [7] to set the DVTR + C2V model. We initialize the spatial transformer (ViT) with CLIP (ViT-B/16) [20] by reusing parameters of similar dimensions in CLIP. We use 12 and 32 as the frame length and caption token number in the DVTR + C2V model. We fine-tune the model with the Adam optimizer. For the learning schedule, we follow the cosine schedule of CLIP [20] to decay the learning rate. The learning rate is set as 1e-7 for both video encoder and text encoder, and 2e-5 for our uncertainty estimation module. The batch size is 128 and running 5 epochs. The sample number of Monte-Carlo sampling is set to 7 for both DVTR + HiT and DVTR + C2V. We set the λ as 0.5 for Eq.12 and τ as 0.07 for Eq.13.

TABLE II
Brief Introduction of Five Publication Datasets Used in Our Experiments: MSRVTT, MSVD, VATEX, ActivityNet, and LSMDC.For a Comprehensive Evaluation, We Conduct Experiments on All Splits of MSRVTT

Datasets	#Videos			#Sentences		
	train	val	test	train	val	test
MSRVTT [9]						
- Full split [9]	6,513	497	2,990	130,260	9,940	59,800
- 1k-A split [45]	7,010	1,000	1,000	140,200	1,000	1,000
MSVD [10]	1,200	100	670	48,774	4,290	27,763
VATEX [11]	25,991	1,500	1,500	259,910	15,000	15,000
ActivityNet [13]	10,009	-	4,917	10,009	-	4,917
LSMDC [12]	109,673	7,408	1,000	109,673	7,408	1,000

3) Evaluation Metrics: We adopt the common metrics to report retrieval performance, including Recall at K (R@K), Median Rank (MedR). R@K is the fraction of queries that correctly retrieve desired items in the top K of the ranking list. Following the tradition, K = 1,5,10 are adopted. Especially, for ActivityNet, K=1,5,50. Therefore, a higher score of R@K means better performance of the retrieval methods. The MedR computes the median rank of the correct targets for a query, where a lower score indicates a better performance. Furthermore, rsum is also considered as the evaluation metric on the overall perspective, which is the sum of the R@K.

4) Compared Methods: To validate the effectiveness of our DVTR, we choose baselines from the following aspects to compare.

a) Conventional Video-Text Retrieval Models: The conventional video-text matching methods [2], [3], [16] focus on

TABLE III
VIDEO-TEXT RETRIEVAL COMPARISON WITH STATE-OF-THE-ART METHODS ON MSRVTT DATASETS

Methods	Split	Text → Video				Video → Text				rsum
		R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
Dual [1]	Full	7.7	22.0	31.8	32.0	13.0	30.8	43.3	15.0	148.6
HGR [3]	Full	9.2	26.2	36.5	24.0	15.0	36.7	48.8	11.0	172.4
CE [14]	Full	10.0	29.0	41.2	16.0	15.6	40.9	55.2	8.3	191.9
HiT [16]	Full	10.5	29.4	41.2	15.0	20.3	44.3	57.3	7.0	203.0
T2VLAD [5]	Full	12.7	34.8	47.1	**12.0**	20.7	48.9	62.1	6.0	226.3
DVTR + HiT	Full	**13.8**	**35.9**	**48.5**	**12.0**	**26.2**	**52.2**	**64.4**	**5.0**	**241.0**
CLIP [21]	Full	21.4	41.1	50.4	10.0	40.3	69.7	79.2	2.0	302.1
MDMMT [46]	Full	23.1	49.8	61.8	6.0	-	-	-	-	-
CLIP2Video [7]	Full	29.8	55.5	66.2	4.0	**54.6**	82.1	90.8	**1.0**	379.0
DVTR + C2V	Full	**32.5**	**58.5**	**68.5**	**3.0**	**54.6**	**83.5**	**91.0**	**1.0**	**388.6**
CE [14]	1k-A	20.9	48.8	62.4	6.0	20.6	50.3	64.0	5.3	267.0
MMT [2]	1k-A	24.6	54.0	67.1	4.0	24.4	56.0	67.8	4.0	293.9
SSB [4]	1k-A	27.4	56.3	67.7	3.0	26.6	55.1	67.5	**3.0**	300.6
HiT [16]	1k-A	27.7	59.2	72.0	3.0	28.8	60.3	72.3	**3.0**	320.3
T2VLAD [5]	1k-A	29.5	59.0	70.1	4.0	31.8	60.0	71.1	**3.0**	321.5
DVTR + HiT	1k-A	**30.5**	**60.7**	**72.6**	**2.0**	**32.0**	**62.3**	**73.8**	**3.0**	**331.9**
CLIP [21]	1k-A	31.2	53.7	64.2	4.0	27.2	51.7	62.6	5.0	290.6
SSB (pretrained) [4]	1k-A	30.1	58.5	69.3	3.0	28.5	58.6	71.6	3.0	316.6
HiT (pretrained) [16]	1k-A	30.7	60.9	73.2	2.6	32.1	62.7	74.1	3.0	333.7
MCQ (pretrained) [47]	1k-A	37.6	64.8	75.1	3.0	-	-	-	-	-
CLIP4CLIP-seqTransf [6]	1k-A	44.5	71.4	81.6	2.0	42.7	70.9	80.6	2.0	391.7
CLIP2Video [7]	1k-A	45.6	72.6	81.7	2.0	43.5	72.3	82.1	2.0	397.8
CAMoE [48]	1k-A	44.6	72.6	81.8	2.0	45.1	72.4	83.1	2.0	399.6
CAMoE + DSL [48]	1k-A	47.3	74.2	84.5	2.0	49.1	74.3	84.3	2.0	413.7
X-Pool [49]	1k-A	46.9	72.8	82.2	2.0	44.4	73.3	84.0	2.0	403.6
QB-NORM+C2V [50]	1k-A	47.2	73.0	83.0	2.0	-	-	-	-	-
DVTR + C2V	1k-A	**52.0**	**77.3**	**85.2**	**1.0**	**50.2**	**76.5**	**85.0**	**1.0**	**426.2**

mining the multi-modality information from video and text to improve the retrieval performance.

- HGR [3] proposes a hierarchical graph reasoning module that decomposes video-text matching into global-to-local levels.
- MMT [2] presents a multi-modal transformer to jointly encode the different modality features in video and allows them to make hierarchical interaction with the text feature.
- HiT [16] proposes a hierarchical transformer for video-text retrieval. It performs hierarchical cross-modal contrastive matching at both feature and semantic levels and achieves a multi-grained matching between video and text modality.

b) The Pretrained Model based Video-Text Retrieval Models: The pre-trained model based video-text retrieval methods [6], [7], [21], [49] transfer the ability of the pre-trained model to the cross-modal retrieval task by fine-tuning in the downstream datasets.

- CLIP-straight [21] directly adopts CLIP [20] to obtain video and text represnetaions for video-text retrieval.
- CLIP4Clip [6] aims to transfer the knowledge of the CLIP model to video-text retrieval and introduces several cross-modal fusion modules to investigate an appropriate cross-modal matching strategy.
- CLIP2Video (C2V) [7] presents a temporal difference block to capture motions at fine temporal video frames, and a temporal alignment block to re-align the token of video clips and phrases and improve the multi-modal matching.
- X-Pool [49] focuses on the difference of information between video and text and proposes an x-pool strategy that main mechanism is a scaled dot product attention for a text to attend to its most semantically similar frames.

c) The Debiased Video-text Retrieval Models: The debiased cross-modal retrievals [15], [27], [48], [50] reveal and alleviate the bias in retrieval datasets. And all above them could be applied directly to many different video-text retrieval for improving the retrieval performance.

- TT [15] aims to alleviate the bias in captions and introduces multiple text encoders as complementary cues to provide an enhanced supervisory for the retrieval model.
- CMGSD [27] proposes an adaptive margin changed with the distance between positive and negative pairs to solve the influence of soft negative samples.
- CAMoE [48] introduces a alignment strategy named dual softmax, which could rectify the similarity matrix by dual soft max to avoid the one-way optimum-match in cross-modal matching.
- QB-NORM [50] presents a re-normalize strategy to alleviate impacts of hub embedding that is close to many queries in common space.

In the following, the best performance is highlighted in **bold**, "-" means no result reported.

TABLE IV
VIDEO-TEXT RETRIEVAL COMPARISON WITH STATE-OF-THE-ART METHODS ON MSVD DATASETS

Methods	Text → Video				Video → Text				rsum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
VSE++ [52]	15.4	39.6	53.0	9.0	-	-	-	-	-
CE [14]	19.8	49.0	63.8	6.0	-	-	-	-	-
SSML [53]	20.3	49.0	63.3	6.0	-	-	-	-	-
SSB [4]	23.0	52.8	65.8	5.0	27.3	50.7	60.8	5.0	280.4
DVTR + HiT	**26.2**	**58.5**	**71.4**	**4.0**	**35.4**	**61.5**	**70.0**	**3.0**	**323.0**
SSB (pretrained) [4]	28.4	60.0	72.9	4.0	34.7	59.9	70.0	3.0	325.9
CLIP [21]	37.0	64.1	73.8	3.0	59.9	85.2	90.7	**1.0**	410.7
CLIP4CLIP-seq [6]	45.2	75.5	84.3	2.0	62.0	87.3	92.6	**1.0**	446.9
CLIP2Video [7]	47.0	76.8	85.9	2.0	58.7	85.6	91.6	**1.0**	445.6
CAMoE [48]	46.9	76.1	85.5	-	-	-	-	-	-
CAMoE + DSL [48]	49.8	79.2	87.0	-	-	-	-	-	-
X-Pool [49]	47.2	77.4	86.0	2.0	66.4	90.0	94.2	**1.0**	461.2
QB-NORM + C2V [50]	47.6	77.6	86.1	2.0	-	-	-	-	
DVTR + C2V	**50.7**	**80.8**	**87.9**	**1.0**	**70.4**	**92.4**	**96.9**	**1.0**	**479.1**

TABLE V
VIDEO-TEXT RETRIEVAL COMPARISON WITH STATE-OF-THE-ART METHODS ON VATEX DATASETS

Methods	Text → Video				Video → Text				rsum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
Dual [1]	31.1	67.4	78.9	3.0	-	-	-	-	-
VSE++ [52]	33.7	70.1	81.0	2.0	-	-	-	-	-
HGR [3]	35.1	73.5	83.5	2.0	-	-	-	-	-
SSB [4]	44.6	81.8	89.5	**1.0**	58.1	83.9	90.9	**1.0**	448.8
DVTR + HiT	**47.2**	**83.4**	**90.3**	**1.0**	**63.2**	**87.0**	**94.3**	**1.0**	**465.4**
SSB (pretrained) [4]	45.9	82.4	90.4	**1.0**	61.2	85.2	91.8	**1.0**	456.9
CLIP [21]	39.7	72.3	82.2	2.0	52.7	88.8	94.9	**1.0**	430.6
CLIP4CLIP-seq [6]	55.9	89.2	95.0	**1.0**	73.2	97.1	99.1	**1.0**	509.5
CLIP2Video [7]	57.3	90.0	95.5	**1.0**	76.0	97.7	**99.9**	**1.0**	516.4
QB-NORM + C2V [50]	58.8	88.3	93.8	**1.0**	-	-	-	-	
DVTR + C2V	**65.6**	**92.8**	**96.8**	**1.0**	**81.6**	**98.9**	99.7	**1.0**	**535.4**

B. Results

1) Comparison of State-of-The-Arts: Tab.III, Tab.IV, Tab.V, Tab.VI and Tab.VII show the performance comparison results between our model and state-of-the-art methods on the five benchmark datasets. Our performance surpasses all state-of-the-art methods on five common datasets across most evaluation metrics for both text-to-video and video-to-text retrieval. We compare our model with other state-of-the-art methods on the MSRVTT Full and 1k-A partitions, respectively. Our retrieval performance at rsum exceeds recent state-of-the-art methods T2VALD [5] by over 10 points. On the MSVD, VATEX, ActivityNet and LSMDC, the DVTR + HiT also outperforms comparison methods by a large margin. The significant improvements achieved by DVTR indicate that the samples which belong to negative but have a close semantic distance with positive ones have seriously disrupted video-text representation learning in the state-of-the-art methods.

We also compare our DVTR + C2V model with the state-of-the-art methods that are pretrained with extra training data, such as pretrained on HowTo100M [59] or adopting the pretrained features from CLIP [20]. On the MSRVTT, MSVD, VATEX, and ActivityNet datasets, we achieve state-of-the-art performance improvements compared with all baselines. On the LSMDC dataset, we outperform the state-of-the-art model on most of the metrics. The results show that pretrained models are still suffering from the negative impact of soft positive samples in the transfer to downstream tasks, althought they maintain a powerful ability and achieve a significant retrieval performance. We notice that the increase brought by DVTR is not as large as the non-pretrain methods. This may be because pretrained models are trained on vast and comprehensive datasets, in which the probability η^+ that one sample has similar semantics to another random sample is smaller in that dataset, thus suffering minor biases.

2) Comparison With Other Denoising Methods: Tab.VIII shows the performance comparison between our proposed DVTR and other denoising methods on the full split of MSRVTT. Following the CMGSD [27] and TT [15], we adopt CE [14] as our backbone model, keep all the settings unchanged and apply our DVTR to it. Specifically, CMGSD gives samples a dynamic margin to reduce their optimization time. But it neglects the soft positive samples, and still provide the same supervision information as the ordinary negative samples. TT uses multiple text encoders to provide abundant text supervision while never considering the impacts introduced by soft positive samples. This improvement can be attributed to the additional supervision information provided by multiple text encoders. The results show that our method

TABLE VI
VIDEO-TEXT RETRIEVAL COMPARISON WITH STATE-OF-THE-ART METHODS ON ACTIVITYNET DATASETS

Methods	Text → Video				Video → Text				rsum
	R@1	R@5	R@50	MedR	R@1	R@5	R@50	MedR	
MMT [2]	22.7	54.2	93.2	5.0	22.9	54.8	93.1	4.3	340.9
T2VLAD [5]	23.7	55.5	93.5	4.0	24.1	56.6	94.1	4.0	347.5
SSB [4]	26.8	58.1	93.5	**3.0**	25.5	57.3	93.5	**3.0**	354.7
DVTR + HiT	**27.1**	**59.3**	**95.0**	**3.0**	**26.8**	**59.7**	**95.1**	**3.0**	**363.0**
MMT (pretrained) [2]	28.7	61.4	94.5	3.3	28.9	61.1	94.3	4.0	368.9
SSB (pretrained) [4]	29.2	61.6	94.7	3.0	28.7	60.8	94.8	**2.0**	369.8
CLIP4CLIP-meanP [6]	40.5	72.4	98.2	**2.0**	-	-	-	-	-
DVTR + C2V	**47.6**	**77.8**	**98.8**	**2.0**	**47.6**	**78.3**	**99.0**	**2.0**	**449.1**

TABLE VII
VIDEO-TEXT RETRIEVAL COMPARISON WITH STATE-OF-THE-ART METHODS ON LSMDC DATASETS

Methods	Text → Video				Video → Text				rsum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
CE [14]	11.2	26.9	34.8	25.3	-	-	-	-	-
MMT [2]	13.2	29.2	38.8	21.0	12.1	29.3	37.9	22.5	
T2VLAD [5]	14.3	32.4	42.2	16.0	14.2	33.5	41.7	17.0	178.3
DVTR + HiT	**16.2**	**34.2**	**44.0**	**15.0**	**16.0**	**34.0**	**43.7**	**15.0**	**190.8**
CLIP-straight [21]	11.3	22.7	29.2	56.5	-	-	-	-	-
MMT (pretrained) [2]	12.9	29.9	40.1	19.3	12.3	28.6	38.9	20.0	162.7
MCQ (pretrianed) [47]	17.9	35.4	44.5	15.0	-	-	-	-	
CLIP4CLIP-seqTransf [6]	22.6	41.0	49.1	11.0	-	-	-	-	-
CAMoE [48]	22.5	42.6	50.9	-	-	-	-	-	-
CAMoE + DSL [48]	**25.9**	46.1	53.7	-	-	-	-	-	-
X-Pool [49]	25.2	43.7	53.5	**8.0**	22.7	42.6	51.2	10.0	238.9
DVTR + C2V	25.2	**46.5**	**54.0**	8.5	**22.9**	**44.3**	**51.4**	**9.0**	**244.2**

TABLE VIII
VIDEO-TEXT RETRIEVAL COMPARISON WITH OTHER DENOISING METHODS ON MSRVTT FULL DATASETS

Methods	R@1	R@5	R@10	MedR
CE [14]	10.0	29.0	41.2	16.0
CE + CMGSD [27]	11.8	32.5	44.2	14.0
CE + TT [15]	11.8	32.7	45.3	**13.0**
CE + DVTR	**12.7**	**34.6**	**45.9**	**13.0**

surpasses CMGSD [27] and TT [15] and further upgrades the retrieval performances of the baseline by considering the semantics of soft positive samples. Specifically, our method gains 2.7, 5.6, and 4.7 on R@1, R@5, and R@10 for CE, respectively. The experimental results show that our model can effectively fix the bias and improve the retrieval performance by identifying the soft positive samples and correcting their inaccurate supervision.

3) Comparison With Different Probabilistic Embedding: In this work, we theoretically analyze the biased supervision problem in Sec.III and find the root causes that the existing methods draw negative samples from the whole dataset, which contains biases. Inspired by the probabilistic embedding [19], we propose an innovative and effective way to solve the problem: locate the biased samples and rescale their contributions by their uncertainty score, as shown in Eq.12 and Eq.13. As reviewed in related work, the PCME [26] also tackles the biased supervision problem, they conjecture that the problem is that many-to-many relationships are not modeled. Thus, they introduce probabilistic embedding to capture many-to-many relationships. The uncertainty in PCME is a by-product of providing interpretability for retrieval results.

In this section, we compare our method with the PCME, and extend the PCME model by replacing our video-text matching uncertainty estimation module in DVTR with it (DVTR$_{pcme}$ + HiT and DVTR$_{pcme}$ + C2V). Results are shown in Tab.IX. According to the results, the PCME model performs poorly in video-text retrieval tasks in R@K metrics. This may be because the probability embedding may better capture relations rather than represent samples. The DVTR$_{pcme}$ + HiT and DVTR$_{pcme}$ + C2V performed well, which demonstrated the effectiveness of our proposed debiased framework in identifying the soft positive samples and calibrating the biased supervision. The DVTR + * models outperform the DVTR$_{pcme}$ + * indicating the effectiveness of the proposed hierarchical probabilistic encoder and heterogeneity-aware multi-modal uncertainty learning. Furthermore, it also demonstrated that the proposed video-text matching uncertainty estimation module estimates the uncertainty score more accurately than PCME.

VI. ABLATION STUDIES

A. Debiased Loss Functions

Tab.X shows the results of ablation studies on the MSRVTT Full datasets of video-text retrieval task. The $\mathcal{L}_{TL}^{U}$ and $\mathcal{L}_{CL}^{U}$ represent the debiased triplet ranking loss and debiased

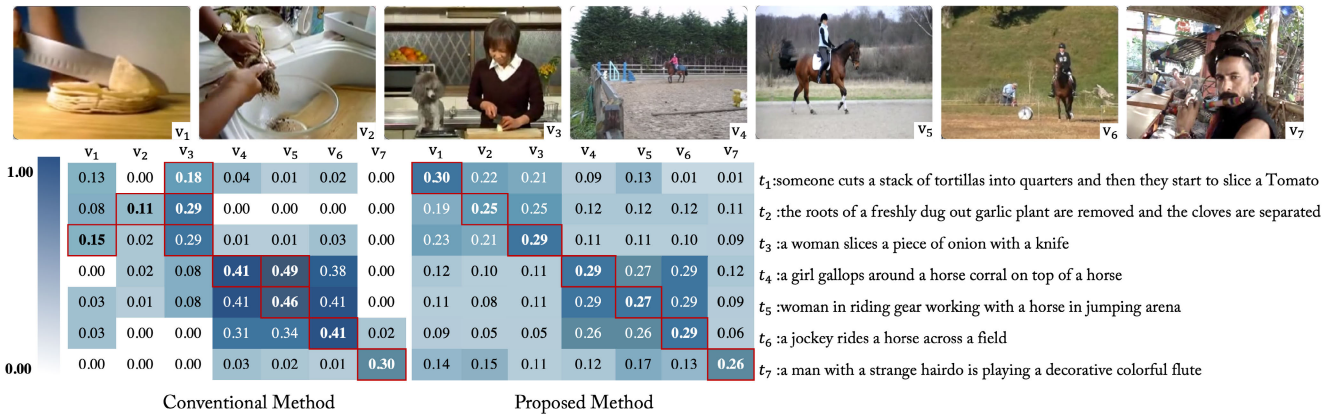

Fig. 3. Visualization of the similarity matrix on a batch of MSVD test datasets The max value on each row and column is highlighted with the red border. The diagonal is the ground-truth pair (v_*, t_*). When obtaining text-to-video retrieval, the model returns the max similar video on each row. For the video-to-text retrieval, the most similar texts over columns are returned.

TABLE IX
VIDEO-TEXT RETRIEVAL COMPARISON WITH OTHER PROBABILISTIC EMBEDDING METHODS ON MSRVTT FULL DATASETS

Methods	R@1	R@5	R@10	MedR
PCME [26]	10.5	29.4	41.2	-
DVTR$_{pcme}$ [26] + HiT	12.1	33.0	47.2	**12.0**
DVTR + HiT	**13.8**	**35.9**	**48.5**	**12.0**
DVTR$_{pcme}$ [26] + C2V	30.5	56.5	67.5	4.0
DVTR + C2V	**32.5**	**58.5**	**68.5**	**3.0**

TABLE X
ABLATION STUDIES ON MSRVTT FULL DATASETS TO INVESTIGATE CONTRIBUTIONS OF DIFFERENT DEBIASED LOSSES

$\mathcal{L}_{TL}^{U}$	$\mathcal{L}_{CL}^{U}$	R@1	R@5	R@10	MedR
✗	✗	10.5	29.4	41.2	15.0
✓	✗	13.6	35.5	47.9	**12.0**
✗	✓	**13.8**	**35.9**	**48.5**	**12.0**

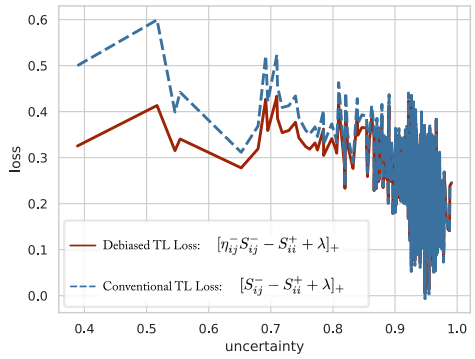

Fig. 4. Visualization of the soft positive sample calibration.

contrastive learning loss, respectively. According to Tab.X, we can find that both $\mathcal{L}_{TL}^{U}$ and $\mathcal{L}_{CL}^{U}$ can help the model achieve a better performance. This illustrates that the debiased loss functions help models to learn a better representation of videos and texts. Furthermore, the debiased contrastive loss functions yield a better performance, indicating that bias may affect conventional contrastive loss functions more easily.

TABLE XI
PARAMETER ANALYSIS FOR THE SAMPLE NUMBER K ON TEXT-TO-VIDEO RETRIEVAL ON MSVD DATASET

K	2	3	5	7	9	11
Geo-Mean	65.68	66.33	67.44	**67.90**	67.72	67.67

TABLE XII
PARAMETER ANALYSIS OF MARGIN λ ON MSVD DATASET

λ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
rsum	319.2	320.4	321.3	321.1	**322.2**	321.4	321.4	320.6	318.5

B. Monte-Carlo Sampling Numbers

Tab.XI reports the effect of the number of samples on the retrieval results by Geometric-Mean of R@1-R@5-R@10 at text-to-video retrievals. In these experiments, we only modify the number of samples of Monte-Carlo Sampling and keep other parameters unchanged. The results show that, the retrieval performance grows with the number of samples. Due to the computation limits, we choose the $K = 7$ as the number of the sample finally.

C. Hyperparameter of Triplet Loss

λ is the key parameter of the triplet loss function. To explore the suitable value of margin λ, sufficient experiments are conducted at Tab.XII. According to Tab.XII, when the margin λ is 0.5, the model achieves the best performance.

D. Hyperparameter of Contrastive Loss

The temperature τ in contrastive loss is a sensitive parameter. To further analyze the effect of τ, we present sufficient experiments and show results at Tab.XIII. A fine-grained step length 0.01 is used to explore the most appropriate τ value. As Tab.XIII shows, the best retrieval performance can be achieved when τ is set to 0.07.

VII. QUALITATIVE RESULTS

A. Visualization of the Similarity Matrix

In this section, we study how the debiased objectives work on retrieval by visualization of the output of the similarity

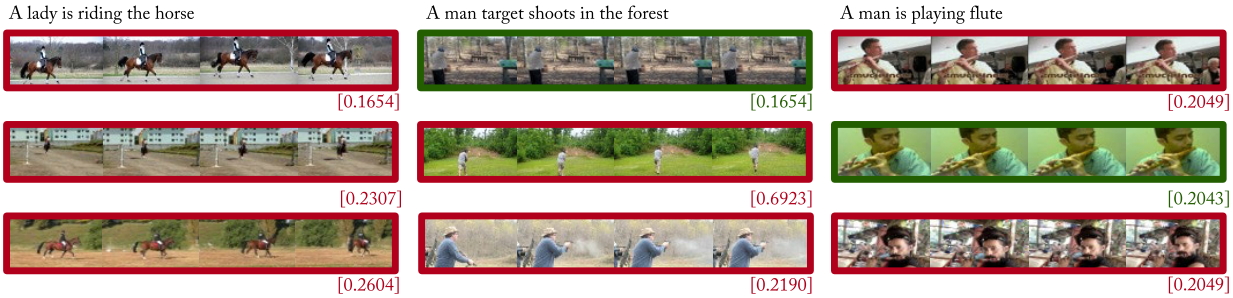

Fig. 5. Visualization of text to video retrieval results on MSVD test dataset. The ground-truth is shown in green.

TABLE XIII
PARAMETER ANALYSIS OF TEMPERATURE τ ON MSVD DATASET

τ	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
rsum	288.5	309.4	318.1	318.9	321.6	320.5	**323.0**	319.4	318.9	317.9

matrix. Fig.3 visualized the similarity scores between the query and candidate samples in the video-text retrieval. From Fig.3, we conclude that: (1) The conventional methods fail in most retrieval cases since they have trouble with the bias of soft positive samples, while the proposed method performs well. (2) Both the conventional method and our proposed method encourage the model to give a high similarity score to the diagonal, i.e., the positive samples. (3) For the negative ones that are not diagonal, our debiased video-text retrieval objectives weightedly reduce their punishment, resulting in an accurate similarity score for the soft positive samples and some minor scores for the true negatives.

B. Visualization of the Soft Positive Sample Calibration

To show the effectiveness of debiasing more clearly, we randomly select a query and show its training loss along with the uncertainty on the whole MSVD dataset. From Fig.4, we can conclude that: (1). Over 95% of negative samples have an uncertainty of over 0.9, which indicates that they are the real negative ones. (2). The model can quickly fit the real negative samples, so they have lower losses. (3). The model has trouble with the biased supervision problem. The wrong supervision forces the model to give a low similarity score for the soft positive samples by imposing significant losses. However, the soft positive samples do have a similar semantic with the query. (4). The proposed negative sample reweighting can effectively reduce the punishment for soft positive samples.

C. Visualization of the Uncertainty Estimation Results

Fig.5 and Fig.6 show the retrieval results in our video-to-text retrieval with the output of the uncertainty estimation module. For both visualization examples, we display the top 3 retrieved results for analysis, among which, the correct results are marked in green and the wrong ones in red. The uncertainty between the retrieved results and the query sample is shown beside or at the bottom of the retrieved results. We can find that although the retrieved results may have no positive candidate from the retrieved results, the proposed debiased retrieval model can still return the most relevant result, which has close semantics to the query. Furthermore, our uncertainty estimation module can also identify them by giving an accurate uncertainty score.

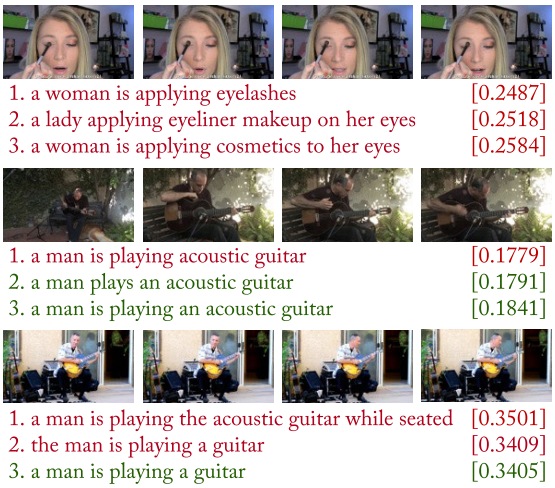

Fig. 6. Visualization of video to text retrieval results on MSVD test dataset. The ground-truth is shown in green.

D. Time Complexity

The uncertainty estimation module in DVTR is based on the features of the video-text retrieval backbone. The additional time is required when DVTR projects v and t as the probabilistic embeddings. At the projection stage, the main time consuming operator is the attention mechanism, which needs additional time of $O(N^3)$. Numerous methods [4], [5], [16] adopt transformers or attention mechanisms to learn better representations of video and text. Thus, compared to other methods, no additional time consumption is required in our DVTR.

E. Space Complexity

The common method of point embedding requires $O(N)$ space to store the features in the joint embedding space. In our DVTR, extra spaces are used at the stages of probabilistic embedding and Monte-Carlo sampling. For probabilistic embedding stage, our DVTR projects v and t as probabilistic distributions. μ and Σ of video and text need to be stored beforehand, resulting in the doubled space requirement. For Monte-Carlo sampling, our DTVR needs K^2 storage by sampling K points from video and text distributions, respectively. Thus, the additional space requirements of DVTR are $O(2N * K^2)$.

VIII. Conclusion

In this work, we tackle the biased supervision of soft positive samples in video-text retrieval learning and propose the novel Debiased Video-Text Retrieval (DVTR) method to alleviate the biased supervision of soft positive samples. We first introduce the novel video-text matching uncertainty estimation module, which identify the soft positive samples by evaluates the uncertainty of the query and candidate samples with probabilistic embeddings. Then, a debiased video-text representation learning objective is employed to fix the inaccurate supervision by weightedly reducing the penalty of soft positive samples in ranking losses. DVTR can be integrated into most video retrieval models for better retrieval performance with a few computational costs at training time and no additional time consumption at test time. Comprehensive experimental results on five widely used datasets demonstrate the superiority of the proposed method compared with other state-of-the-art video-text retrieval methods.

References

[1] J. Dong et al., "Dual encoding for zero-example video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9346–9355.

[2] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *Computer Vision—ECCV 2020*. Glasgow, U.K.: Springer, Aug. 2020, pp. 214–229.

[3] 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.

[4] M. Patrick et al., "Support-set bottlenecks for video-text representation learning," in *Proc. ICLR*, 2021, pp. 1–18.

[5] 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.

[6] H. Luo et al., "CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, Oct. 2022.

[7] H. Fang, P. Xiong, L. Xu, and Y. Chen, "CLIP2Video: Mastering video-text retrieval via image CLIP," *CoRR*, vol. abs/2106.11097, pp. 1–10, Jun. 2021.

[8] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10941–10950.

[9] 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.

[10] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. ACL*. Stroudsburg, PA, USA: Association for Computer Linguistics, 2011, pp. 190–200.

[11] 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. 4580–4590.

[12] A. Rohrbach, M. Rohrbach, and B. Schiele, "The long-short story of movie description," in *Pattern Recognition*. Aachen, Germany: Springer, Oct. 2015, pp. 209–221.

[13] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*. Venice, Italy: IEEE Computer Society, Oct. 2017, pp. 706–715.

[14] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," in *Proc. BMVC*. Cardiff, U.K.: BMVA Press, 2019, p. 279.

[15] I. Croitoru et al., "TeachText: CrossModal generalized distillation for text-video retrieval," in *Proc. ICCV*, Oct. 2021, pp. 11583–11593.

[16] S. Liu, H. Fan, S. Qian, Y. Chen, W. Ding, and Z. Wang, "HiT: Hierarchical transformer with momentum contrast for video-text retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 11895–11905.

[17] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 1708–1718.

[18] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst.*, vol. 30, I. Guyon et al., Eds., Long Beach, CA, USA, Dec. 2017, pp. 5574–5584.

[19] S. J. Oh, A. C. Gallagher, K. P. Murphy, F. Schroff, J. Pan, and J. Roth, "Modeling uncertainty with hedged instance embeddings," in *Proc. ICLR*, 2019, pp. 1–17.

[20] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, M. Meila and T. Zhang, Eds., Jul. 2021, pp. 8748–8763.

[21] J. A. Portillo-Quintero, J. C. Ortiz-Bayliss, and H. Terashima-Marín, "A straightforward framework for video retrieval using CLIP," in *Pattern Recognition* (Lecture Notes in Computer Science), vol. 12725, E. Roman-Rangel, Á. F. K. Morales, J. F. M. Trinidad, J. A. Carrasco-Ochoa, and J. A. Olvera-López, Eds. Mexico City, Mexico: Springer, 2021, pp. 3–12.

[22] D. Ko et al., "Video-text representation learning via differentiable weak temporal alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 5006–5015.

[23] 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.

[24] P. Zhang, Z. Zhao, N. Wang, J. Yu, and F. Wu, "Local-global graph pooling via mutual information maximization for video-paragraph retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7133–7146, Oct. 2022.

[25] M. Wray, H. Doughty, and D. Damen, "On semantic similarity in video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3650–3660.

[26] S. Chun, S. J. Oh, R. Sampaio de Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. New York, NY, USA: Computer Vision Foundation, Jun. 2021, pp. 8415–8424.

[27] F. He et al., "Improving video retrieval by adaptive margin," in *Proc. SIGIR*, Jul. 2021, pp. 1359–1368.

[28] J. Li, H. Yong, F. Wu, and M. Li, "Online multi-view subspace learning with mixed noise," in *Proc. 28th ACM Int. Conf. Multimedia (MM)*, C. W. Chen et al., Eds., Seattle, WA, USA, Oct. 2020, pp. 3838–3846.

[29] X. He and Y. Peng, "Fine-grained visual-textual representation learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 520–531, Feb. 2020.

[30] L. Yang, H. Li, F. Meng, Q. Wu, and K. N. Ngan, "Task-specific loss for robust instance segmentation with noisy class labels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 213–227, Jan. 2023.

[31] L. Zhang et al., "Multimodal marketing intent analysis for effective targeted advertising," *IEEE Trans. Multimedia*, vol. 24, pp. 1830–1843, 2022.

[32] Z. Cheng, X. Li, J. Shen, and A. G. Hauptmann, "Which information sources are more effective and reliable in video search," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven, and J. Zobel, Eds., Pisa, Italy, Jul. 2016, pp. 1069–1072.

[33] Z. Liu, F. Chen, J. Xu, W. Pei, and G. Lu, "Image-text retrieval with cross-modal semantic importance consistency," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Nov. 7, 2022, doi: 10.1109/TCSVT.2022.3220297.

[34] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Seattle, WA, USA: Computer Vision Foundation, Jun. 2020, pp. 5709–5718.

[35] Z. Zheng and Y. Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1106–1120, 2021.

[36] Y. Wang, J. Peng, and Z. Zhang, "Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9092–9101.

[37] J. U. Kim, S. Park, and Y. M. Ro, "Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1510–1523, Mar. 2022.

[38] S. Cheng et al., "Uncertainty-aware and multigranularity consistent constrained model for semi-supervised hashing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6914–6926, Oct. 2022.
[39] J. U. Kim, S. T. Kim, H. J. Lee, S. Lee, and Y. M. Ro, "CUA loss: Class uncertainty-aware gradient modulation for robust object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3529–3543, Sep. 2021.
[40] W. Su, Q. Xu, and W. Tao, "Uncertainty guided multi-view stereo network for depth estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7796–7808, Nov. 2022.
[41] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Virtual Event, Austria, May 2021, pp. 1–22.
[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, vol. 1, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.
[43] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, pp. 1–14, Jul. 2016.
[44] C. Villani, *Topics in Optimal Transportation*, vol. 58. Providence, RI, USA: American Mathematical Society, 2021.
[45] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 471–487.
[46] M. Dzabraev, M. Kalashnikov, S. Komkov, and A. Petiushko, "MDMMT: Multidomain multimodal transformer for video retrieval," in *Proc. CVPR Workshops*, Jun. 2021, pp. 3354–3363.
[47] Y. Ge et al., "Bridging video-text retrieval with multiple choice questions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 16146–16155.
[48] X. Cheng, H. Lin, X. Wu, F. Yang, and D. Shen, "Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss," 2021, *arXiv:2109.04290*.
[49] S. K. Gorti et al., "X-pool: Cross-modal language-video attention for text-video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5006–5015.
[50] S.-V. Bogolin, I. Croitoru, H. Jin, Y. Liu, and S. Albanie, "Cross modal retrieval with querybank normalisation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5194–5205.
[51] B. Zhang, H. Hu, and F. Sha, "Cross-modal and hierarchical modeling of video and text," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 374–390.
[52] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*.
[53] E. Amrani, R. Ben-Ari, D. Rotman, and A. Bronstein, "Noise estimation using density estimation for self-supervised multimodal learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 8, pp. 6644–6652.
[54] S. Hershey et al., "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.
[55] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[56] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 305–321.
[57] 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)*. Honolulu, HI, USA: IEEE Computer Society, Jul. 2017, pp. 4724–4733.
[58] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019, pp. 1–19.
[59] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2630–2640.

Huaiwen Zhang (Member, IEEE) received the B.E. degree from Inner Mongolia University, Hohhot, Inner Mongolia, China, in 2016, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2021. He is currently a Professor with the National and Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Inner Mongolia University. His research interests include social multimedia analysis and multimedia computing.

Yang Yang (Graduate Student Member, IEEE) received the B.E. degree from Inner Mongolia University, Hohhot, China, in 2019. He is currently pursuing the Ph.D. degree with the National and Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Inner Mongolia University. His research interests include video-text retrieval and cross-modal matching.

Fan Qi received the Ph.D. degree in computer science and technology from the Hefei University of Technology in 2021. She is currently an Associate Professor at the Tianjin University of Technology. Her research interests include multimedia sentiment analysis and computer vision.

Shengsheng Qian (Member, IEEE) received the B.E. degree from Jilin University, Changchun, China, in 2012, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include social media data mining and social event content analysis.

Changsheng Xu (Fellow, IEEE) is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and the Executive Director of the China-Singapore Institute of Digital Media. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. He has hold 30 granted/pending patents and published over 200 refereed research papers in these areas. He is a fellow of IAPR and ACM Distinguished Scientist. He received the Best Associate Editor Award of the *ACM Transactions on Multimedia Computing, Communications and Applications* in 2012 and the Best Editorial Member Award of *ACM/Springer Multimedia Systems* Journal in 2008. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing, Communications and Applications*, and *ACM/Springer Multimedia Systems* journal. He has served as the Program Chair for ACM Multimedia 2009. He has served as an associate editor, the guest editor, the general chair, the program chair, the area/track chair, a special session organizer, the session chair, and a TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops.