UATVR: Uncertainty-Adaptive Text-Video Retrieval

Bo Fang[1]* Wenhao Wu[2,3]* Chang Liu[4]* Yu Zhou[1]† Yuxin Song[3]
Weiping Wang[1] Xiangbo Shu[5] Xiangyang Ji[4] Jingdong Wang[3]
[1]Institute of Information Engineering, Chinese Academy of Sciences [2]The University of Sydney
[3]Baidu Inc. [4]Tsinghua University [5]Nanjing University of Science and Technology
{fangbo,zhouyu,wangweiping}@iie.ac.cn, {songyuxin02,wangjingdong}@baidu.com,
{liuchang2022,xyji}@tsinghua.edu.cn, wenhao.wu@sydney.edu.au, shuxb@njust.edu.cn

Abstract

With the explosive growth of web videos and emerging large-scale vision-language pre-training models, e.g., CLIP, retrieving videos of interest with text instructions has attracted increasing attention. A common practice is to transfer text-video pairs to the same embedding space and craft cross-modal interactions with certain entities in specific granularities for semantic correspondence. Unfortunately, the intrinsic uncertainties of optimal entity combinations in appropriate granularities for cross-modal queries are understudied, which is especially critical for modalities with hierarchical semantics, e.g., video, text, etc. In this paper, we propose an Uncertainty-Adaptive Text-Video Retrieval approach, termed UATVR, which models each look-up as a distribution matching procedure. Concretely, we add additional learnable tokens in the encoders to adaptively aggregate multi-grained semantics for flexible high-level reasoning. In the refined embedding space, we represent text-video pairs as probabilistic distributions where prototypes are sampled for matching evaluation. Comprehensive experiments on four benchmarks justify the superiority of our UATVR, which achieves new state-of-the-art results on MSR-VTT (50.8%), VATEX (64.5%), MSVD (49.7%), and DiDeMo (45.8%). The code is available at https://github.com/bofang98/UATVR.

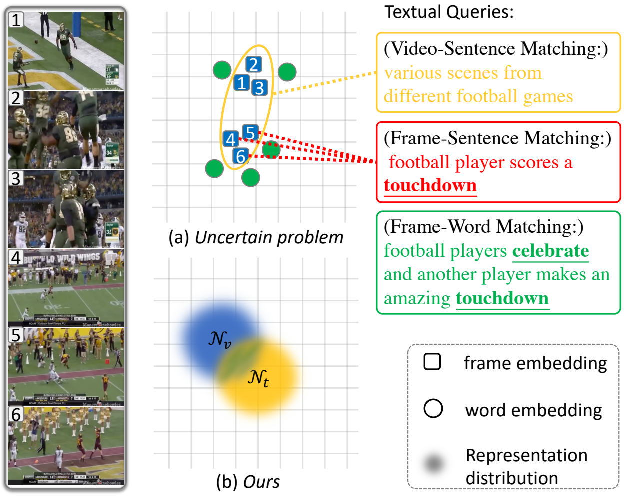

Figure 1. Motivation. A video has numerous descriptions containing different level information, (a) which shows inconsistent text-video correspondences in the common embedding space. The former three frames depict a 'celebrate' action, while the last three frames are about a 'touchdown'. This video diversity thus makes the optimal text-video matching in uncertain granularities, which we call an uncertain matching problem. Moreover, previous deterministic works can only handle one-to-one text-video mappings, yet a realistic relationship between two modalities is one-to-many. (b) The above problems motivate our uncertainty-adaptive model through distribution matching procedures.

1. Introduction

With surging portable filming devices and emerging video media platforms, searching videos of interest with human instructions, typically as texts, has been a part of daily lives, which urgently requires effective and robust text-video retrieval (TVR) techniques. Given a query text (video), TVR aims to find the most relevant video (text) in the database, which is typically overwhelmed with sophisticated vague semantic combinations varying with hierarchical text structures or spatiotemporal video spans.

Recent breakthroughs in the large-scale image and/or text pre-training [48, 22, 61] benefit TVR significantly. A serial of seminal works employ a separated encoder architecture to respectively project texts and videos into a pre-trained joint embedding space for compact cross-modal in-

*Co-first authorship. This work was done when Bo Fang was a research intern at Baidu Inc.

†Corresponding author.

teraction [30, 4, 39, 31].

Since a video inherently contains information beyond texts, simply pooling all frames as a whole video expression brings distraction during matching specific text entities [39, 17]. Therefore, inspired by fine-grained image-text pre-training, *e.g.*, FILIP [59] and ALBEF [32], multi-grained TVR paradigms are introduced to build multi-level cross-modal interactions with *sentence-frame* level [20, 34], *word-frame* level [52], or hierarchical correspondences including *phrase-clip* level [43, 25]. However, these methods are still far from satisfying in handling the intrinsic uncertainties of determining the optimal entity combinations with appropriate granularities during text-video matching.

In Fig. 1, we illustrate the uncertain matching problem in TVR. Since different frame/word combinations can plausibly correspond to semantics in various perspectives, given the same video, successful retrieval can achieve in varying granularities involving discrepant text-video entities, *i.e.*, video-sentence matching, frame-sentence matching, frame-word matching, *etc*. Previous works determine particular cross-modal mapping strategies in certain granularities, yet none have studied the intrinsic uncertainties of optimal text-video entity combinations. Besides, existing deterministic cross-modal retrieval can only handle one-to-one mapping scenarios [12]. However, a video can be described by multiple sentences typically (and vice versa), which formulates realistic one-to-many relationships.

In this paper, we propose a novel TVR framework to tackle the uncertainty problem in cross-modal matching, termed **U**ncertainty-**A**daptive **T**ext-**V**ideo **R**etrieval (**UATVR**). Generally, UATVR models each text-video lookup as a distribution matching procedure in complementary deterministic and probabilistic views. It is materialized upon word-frame token-wise interactions and consists of a dynamic semantic adaptation (DSA) module and a distribution-based uncertainty adaptation (DUA) module.

Concretely, DSA module enhances token-wise matching by introducing additional learnable multi-class tokens. We find these simple-yet-effective tokens can adaptively aggregate multi-grained video (or text) semantics during matching, thus allowing for flexible high-level reasoning. For DUA, we represent samples from each modality as distributions rather than feature points and convert the deterministic matching process to probabilistic distribution alignment. To simulate one-to-many text-video mappings, we pull probabilistic embeddings sampled from each distribution closer via multi-instance contrastive loss [41].

Our contributions can be summarized as (i) We innovatively model video and text representations as probabilistic distributions and align them through multiple-instance contrast in their common embedding space for uncertainty-adaptive cross-modal matching. (ii) We propose a simple-yet-effective technique for flexible high-level reasoning by adding additional learnable tokens, allowing deterministic semantic uncertainty adaptation in videos/texts. (iii) Comprehensive experimental explorations demonstrate the superiority of our UATVR, which obtains state-of-the-art results across public TVR benchmarks including MSR-VTT [57], MSVD [56], VATEX [54], and DiDeMo [1].

2. Related Work

Vision-Language Pre-training. Cross-modal vision language understanding [51, 8] is a challenging task for both computer vision and natural language processing communities. Recent breakthroughs are large-scale image-text contrastive pre-training, which employs a contrastive loss to jointly align image-text semantics into a unified embedding space, on more than 100M samples [48, 22]. Vision-language pre-training with this paradigm [61, 60, 32, 33, 6] has significantly boosted numerous cross-modal tasks such as VQA [2], image captioning [58], text-image retrieval [27], *etc*. For the video counterparts, large-scale video caption datasets, *e.g.*, HowTo100M [42] and WebVid2M [4], also boost promising cross-modal video understanding. However, due to the high cost of collecting wild videos and huge computing resources requirement, we bootstrap from CLIP like [39, 47] for text-video retrieval.

Text-Video Retrieval is to find the most semantic-relevant video given a text query (text → video). Early research devotes to distilling knowledge from "expert" models based on offline-extracted single-modality features [17, 35, 53, 9, 15, 38]. Recent dominant TVR benefits from end-to-end pre-training on large-scale text-video datasets [42, 4, 41]. Strategies that can improve training efficiency are essential for end-to-end paradigms like ClipBERT [30] and Frozen [4]. In TMVM [34], masked-based prototypes for aggregating video features are proposed, which play a similar role to our DSA tokens. However, only visual RGB frames are modeled in TMVM, ignoring the hierarchical attributes in the textual counterpart.

Another idea of TVR transfers knowledge from publicly available CLIP models pre-trained on large-scale text-image pairs and then align text-video modalities with choreographed mapping strategies [39, 62, 20, 40, 18, 16, 10, 24, 26]. Considering the discrepancy problem that videos always express more information than texts can capture [20], subsequent works devote to crafting cross-modal interactions with certain entities in specific granularities, *e.g.*, sentence-frame level [20, 34], word-frame level [52], and hierarchical level interactions [43, 25, 55]. TS2-Net [37] selects top-k informative tokens per frame, representing salient semantics, for frame-wise cross-modal matching. It is been designed upon a more fine-grained level. Yet none above have studied the intrinsic uncertainties of optimal entity combinations in appropriate granularities, which motivates our uncertainty-adaptive matching model.

Probabilistic Representations. The probabilistic theory has a long history in machine learning [44]. For the vision domain, HIB [45] first introduces probabilistic embeddings to capture the uncertainty of image representations whilst handling the one-to-many correspondences for deep metric learning. Moreover, they have also been successfully applied to other tasks like face recognition [49, 7], pose estimation [50], *etc*. PCME [12] employs probabilistic embeddings for text-image retrieval to perform one-to-many matching between the multiplicity of visual concepts, which inspires us to expand them to videos, as videos typically contain more complex semantic concepts for their temporal dynamics. Moreover, we find that soft contrastive loss [45] used in PCME is sub-optimal for text-video modelling. Instead, we introduce multi-instance contrast for a more appropriate one-to-many relation simulating. From this, our uncertainty-adaptive matching model tackles the uncertain matching problem and remarkably surpasses previous methods.

3. Method

In this section, we first introduce our token-wise word-frame matching baseline. Then we propose two essential modules of UATVR, *e.g.*, dynamic semantic adaptation and adaptive distribution matching, for tackling the uncertainty problem in text-video retrieval.

3.1. Preliminary

Problem Definition. TVR aims to learn a similarity calculation function $s(\cdot)$, which ought to maximize the similarity score of positive cross-modal samples and assign lower similarity for irrelevant pairs. Formally, given a pair of text $t_i \in \mathbb{R}^{N+1}$ and video $v_i \in \mathbb{R}^{M\times 3\times H\times W}$, we formulate them as collections of N words and M frames with $t_i = [w_i^0, w_i^1, w_i^2, \cdots, w_i^N]^T$, $v_i = [f_i^1, f_i^2, \cdots, f_i^M]^T$, where w_i^0 represents the `[CLS]` token and $H \times W$ denotes the resolution. We feed t_i and v_i into a text encoder and a video encoder respectively to get their corresponding embeddings $\mathbf{t}_i = [\mathbf{w}_i^0, \mathbf{w}_i^1, \cdots, \mathbf{w}_i^N]^T$ and $\mathbf{v}_i = [\mathbf{f}_i^1, \mathbf{f}_i^2, \cdots, \mathbf{f}_i^M]^T$. The frame embedding $\mathbf{f}_i^m$ comes from the distinct `[CLS]` token from the transformer-based vision encoder for the *m*th frame. Normally, we represent the whole video by average pooling all frame embeddings in Eq. 1 and represent the sentence with the first `[CLS]` token feature $\mathbf{w}_i^0$ in Eq. 2. The similarity of the text-video is calculated as the inner production of $\mathbf{t}_i, \mathbf{v}_i$, *c.f.* Eq. 3.

$$\mathbf{v}_i = \text{meanPool}([\mathbf{f}_i^1, \mathbf{f}_i^2, \cdots, \mathbf{f}_i^M]^T), \tag{1}$$

$$\mathbf{t}_i = \mathbf{w}_i^0, \tag{2}$$

$$s(\mathbf{t}_i, \mathbf{v}_i) = \langle \mathbf{t}_i, \mathbf{v}_i \rangle. \tag{3}$$

In training, a common optimizing method is to use a symmetric cross-entropy loss in both text-to-video and video-to-text directions. Given a batch of B text-video pairs, the model updates its parameters by maximizing the sum of the main diagonal of a $B \times B$ similarity matrix:

$$\mathcal{L}_{t2v} = -\frac{1}{B}\sum_i^B \log\frac{\exp(s(\mathbf{t}_i, \mathbf{v}_i))}{\sum_{j=1}^B \exp(s(\mathbf{t}_i, \mathbf{v}_j))}, \tag{4}$$

$$\mathcal{L}_{v2t} = -\frac{1}{B}\sum_i^B \log\frac{\exp(s(\mathbf{v}_i, \mathbf{t}_i))}{\sum_{j=1}^B \exp(s(\mathbf{v}_i, \mathbf{t}_j))}, \tag{5}$$

$$\mathcal{L} = \mathcal{L}_{t2v} + \mathcal{L}_{v2t}. \tag{6}$$

Fine-grained Interactions. Before the late cross-modal fusion, the key point lies in how to extract accurate video information best described by corresponding textual queries. The naive method pools all frame embeddings equally [39] yet ignores the discrepancy problem that a video contains more information than a single caption can depict [20]. Recent work attempts to ameliorate the above problem by devoting to constructing multi-grained cross-modal interactions, including sentence-frame [20, 34], word-frame [52], or multiple hierarchical interactions [43, 40]. In this paper, we take the token-wise word-frame matching paradigm as a solid baseline due to its successful application in image-text pre-training [59] and strong TVR performance [52]. The text-video similarity thus comes from the mean of the maximum similarities between each frame with all word-level embeddings in bi-directions, formulated as:

$$s(\mathbf{t}_i, \mathbf{v}_i) = \frac{1}{2}\left(\sum_{n=1}^{N}\max_{m=1}^{M}\langle \mathbf{w}_i^n, \mathbf{f}_i^m\rangle + \sum_{m=1}^{M}\max_{n=1}^{N}\langle \mathbf{w}_i^n, \mathbf{f}_i^m\rangle\right), \tag{7}$$

where M, N denote frame and word number in the ith sample pair. $\mathbf{w}_i^n, \mathbf{f}_i^m$, which are channel-wise normalized before calculating, refer to the nth word embedding and mth frame embedding, respectively. Eq.7 would produce larger similarity sums for longer video and text input. Therefore, an average operation is attached before addition.

3.2. Dynamic Semantic Adaptation

To a certain extent, vanila token-wise matching brings more accurate text-video correspondences. However, fine-grained TVR interaction in a deterministic matching granularity does not consider the uncertain matching problem. To tackle the problem, we introduce multiple additional learnable tokens to dynamically aggregate multi-level video and text information while retaining the advance of local context matching in the token-wise baseline, *c.f.* Fig. 2.

Given sequential frame embeddings $\{\mathbf{f}_i^m\}_{m=1}^M$ and word embeddings $\{\mathbf{w}_i^n\}_{n=1}^N$ extracted from backbone encoders,

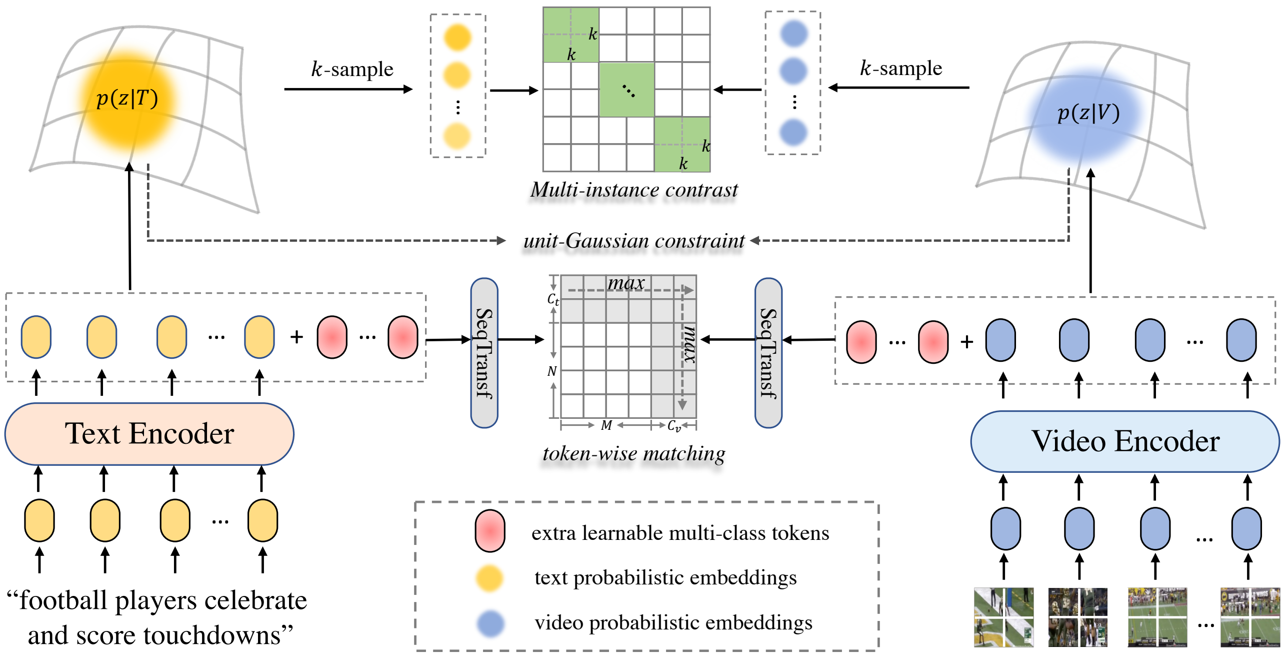

Figure 2. UATVR pipeline. We concatenate numerous extra learnable tokens with sequential frame/word embeddings and feed them into a lightweight SeqTransf [39] for adaptive token-wise matching. Besides, we model all visual and textual tokens as probabilistic distributions and sample k probabilistic embeddings from each distribution with gradient propagating to construct multi-instance contrasts in a batch.

we append additional C_v and C_t learnable multi-class tokens respectively at the sequences' beginning (or end). The extra class tokens are randomly initialized and the position embeddings are omitted for brevity in Fig. 2. Then we feed the enlarged $(M + C_v)$ frame embeddings set into another lightweight sequential Transformer (following the same seqTransf structure in [39]) to further model relations between learnable tokens and frames depending on the corresponding text queries. We adopt the same operation symmetrically for the $(N + C_t)$ word embeddings. All enlarged frame and word embeddings are channel-wise normalized before similarity calculation. Similar to token-wise interactions, we finally calculate our modified text-video similarity function upon the union of frame/word tokens and extra learnable class tokens, formulated as:

$$s = \frac{1}{2}\left(\sum_{n=1}^{N+C_t}\max_{m=1}^{M+C_v}\langle \mathbf{w}_i^n, \mathbf{f}_i^m\rangle + \sum_{m=1}^{M+C_v}\max_{n=1}^{N+C_t}\langle \mathbf{w}_i^n, \mathbf{f}_i^m\rangle\right). \tag{8}$$

We define our dynamic semantic adaption loss $\mathcal{L}_{\mathrm{DSA}}$ in the same formulation as Eq. 4,5,6, in which the modified similarity function in Eq. 8 is employed for text-video cross-modal matching. The additional semantic-aggregated tokens introduce negligible parameters and training overhead, Tab. 2, which is entirely simple yet effective. We give a deep analysis of DSA tokens in Sec. 4.2 and appendix, emphatically interpreting their working mechanisms.

3.3. Adaptive Distribution Matching

Since deterministic methods can only handle one-to-one mapping scenarios [12], videos typically have multiple descriptions, formulating realistic one-to-many text-video relationships instead. We propose adaptive distribution matching for probabilistic TVR to tackle the inherent inconsistency in text-video distributions, *c.f.* Fig. 2.

Let $\mathbf{t}_i$, $\mathbf{v}_i$ denote the output of each backbone. We represent the text caption t_i and the video v_i as normal distributions $p(z|t_i)$ and $p(z|v_i)$ with mean vectors and diagonal covariance matrices in $\mathbb{R}^D$, respectively:

$$\begin{aligned} p(z|t_i) &\sim \mathcal{N}(h^{\mu}_{\mathcal{T}}(\mathbf{t}_i), \mathrm{diag}(h^{\sigma}_{\mathcal{T}}(\mathbf{t}_i))), \\ p(z|v_i) &\sim \mathcal{N}(h^{\mu}_{\mathcal{V}}(\mathbf{v}_i), \mathrm{diag}(h^{\sigma}_{\mathcal{V}}(\mathbf{v}_i))), \end{aligned} \tag{9}$$

where head module h^{μ} is a fully-connected layer followed by LayerNorm [3] and l_2 normalization, and head h^{σ} is a separate fully-connected layer without any normalization following [12]. The textual head $h_{\mathcal{T}}$ and visual head $h_{\mathcal{V}}$ share the same parametric structure but optimize independently upon transformer-based feature output. Next, two groups of K probabilistic embeddings $\{\mathbf{t}_i^{(1)}, \cdots, \mathbf{t}_i^{(K)}\} \overset{\text{iid}}{\sim} p(z|t_i)$ and $\{\mathbf{v}_i^{(1)}, \cdots, \mathbf{v}_i^{(K)}\} \overset{\text{iid}}{\sim} p(z|v_i)$ are generated by sampling from the distributions of t_i and v_i with gradient propagating. To enable stable training, we use the reparametrization trick [29] during the generation, formulated as follows:

$$\begin{aligned} \mathbf{t}_i^{(k)} &= \sigma(t_i)\cdot\epsilon^k + \mu(t_i), \\ \mathbf{v}_i^{(k)} &= \sigma(v_i)\cdot\epsilon^k + \mu(v_i), \end{aligned} \tag{10}$$

where $\epsilon^{(k)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I)$ and μ, σ denote the mean and the standard deviation of $p(z|t_i)$, $p(z|v_i)$.

Unlike previous methods [45, 12] using soft contrastive loss (a binary classification loss based on the softmax cross-entropy) via Monte-Carlo estimation for distribution alignment, we treat all probabilistic embeddings from a matched text-video pair as positive samples to simulate one-to-many cross-modal matching. We update the model with a Multi-Instance InfoNCE [41] loss, and the training target is to minimize the distance between video (text) probabilistic embeddings and all corresponding text (video) embeddings. Further comparisons are made in the appendix. Given a specific textual probabilistic embedding $\mathbf{t}_i \in \{\mathbf{t}_i^{(k)}\}_{k=1}^K$, we define the positive set $\mathcal{P}_i$ for $\mathbf{t}_i$ as all video probabilistic embeddings from v_i, formulated as $\mathcal{P}_i = \{\mathbf{v}_i^{(k)}\}_{k=1}^K$. The negative set thus is formed as probabilistic embeddings from other videos in the batch, $\widetilde{\mathcal{P}}_i = \{\mathbf{v}_j^{(k)}\}_{j,k}, j \neq i$. We define the distribution-based uncertainty adaptation loss as:

$$\mathcal{L}_{\text{DUA}} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\sum_{\mathbf{v}_i \in \mathcal{P}_i} \exp(s(\mathbf{t}_i, \mathbf{v}_i))}{\sum_{\mathbf{v}_j \in \{\mathcal{P}_i \cup \widetilde{\mathcal{P}}_i\}} \exp(s(\mathbf{t}_i, \mathbf{v}_j))}. \tag{11}$$

3.4. Total Objectives

Following [45], we introduce additional KL divergence loss between the distributions and the unit Gaussian prior $\mathcal{N}(0, I)$ to constraint the learned variances from collapsing to zero, which can be formulated as:

$$\mathcal{L}_{\text{KL}} = \text{KL}(p(z|t_i), \mathcal{N}(0, I)) + \text{KL}(p(z|v_i), \mathcal{N}(0, I)). \tag{12}$$

Therefore, the total objectives can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{DSA}} + \alpha \cdot \mathcal{L}_{\text{DUA}} + \beta \cdot \mathcal{L}_{\text{KL}}, \tag{13}$$

where α and β control the trade-off among three terms.

4. Experiments

We first describe the experimental settings. Then thorough ablation studies are conducted to demonstrate the effectiveness of our proposed UATVR. Finally, we make comparisons of our model to existing state-of-the-art methods on various TVR benchmarks.

4.1. Experimental Settings

Datasets. Experiments are conducted on 4 common video-text retrieval benchmarks: **(a) MSR-VTT** [57] contains 10K video clips in total with 20 captions for each. Following the data splits from [17, 42, 39], we train models on the `Training-9K` set with corresponding captions and report results on the `test 1K-A` set. **(b) MSVD** [56] includes 1,970 videos and 80K captions, with ∼40 captions on average per video. Train, validation, and test set have 1,200, 100, and 670 videos respectively. **(c) DiDeMo** [1] contains 10K videos paired with 40K descriptions. Following previous [39, 4, 30], we concatenate all descriptions of one video to a single query. **(d) VATEX** [54] collects ∼35K videos with multiple annotations for each. There are ∼26K videos for training, 1,500 for validating, and 1,500 for testing.

Methods	R@1	R@5	R@10	MdR↓	MnR↓
CLIP4Clip [39]	47.1	74.1	81.8	2.0	14.9
TI (Token-Wise)	48.4	74.2	83.3	2.0	14.1
+ DSA	49.6	75.5	84.9	2.0	12.5
+ DUA†	50.1	75.8	84.6	1.5	12.8
+ KL† (UATVR)	**50.8**	**76.3**	**85.5**	**1.0**	**12.4**
+ DUA*	50.0	75.8	83.9	1.5	12.9
+ KL*	50.6	75.9	84.9	**1.0**	12.8

Table 1. Ablation study of different components. † denotes the implementation with MIL-NCE contrast and * is implemented with soft contrastive loss via Monte-Carlo estimation [45].

Methods	Time Complexity	Params	Time
CLIP4Clip [39]	$\mathcal{O}(B^2)$	162.3M	75.04h
TI (Token-Wise)	$\mathcal{O}(B^2MN)$	162.8M	82.00h
w/ DSA	$\mathcal{O}(B^2M'N')$	162.8M	82.24h
w/ DUA	$\mathcal{O}(B^2(M'N'+K^2))$	164.9M	84.08h

Table 2. Comparisons of different components. B denotes sample size. M, N denote the length of frame tokens and text tokens resp. M', N' are slightly enlarged with additional learnable tokens. K is the number of probabilistic embeddings. Training time denotes GPU hours calculated by a single P40 card.

Evaluation Metrics. For brevity, we abbreviate Recall at K to R@K ($K = 1, 5, 10$) upon all datasets, which calculates the percentage of correct videos among the top K retrieved videos given textual queries (Text→Video, and vice versa). MdR, Median Rank, calculates the median of the ground truth in the retrieval ranking list. MnR, Mean Rank, calculates the mean rank of the correct results in the retrieval ranking list. Note that for MdR and MnR, the lower score means the better (indicated as ↓).

Implementation Details. We initialize our visual and language backbone with CLIP [48] pre-trained weight. Following [39], we further use a four-layer lightweight sequential transformer to encode extra learnable class tokens with frame and word embeddings. In ablations, we take ViT/B-16 by default. The textual token length is 32 and the frame length is 12 for all datasets except DiDeMo (64 max query words and 64 frames). A uniform frame sampling strat-

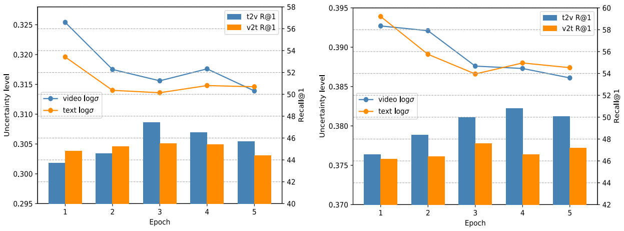

Figure 3. Uncertainty level versus R@1 on MSR-VTT dataset. (Left) Results on ViT-B/32. (Right) Results on ViT-B/16.

Extra-Tokens #		R@1	R@5	R@10	MdR↓	MnR↓
baseline	0	48.4	74.2	83.3	2.0	14.1
	1	48.9	75.3	84.6	2.0	12.3
	2	**49.6**	75.5	84.6	2.0	12.5
C_v	3	**49.6**	**75.7**	84.6	2.0	**11.9**
$(C_t = 0)$	4	49.3	75.2	84.2	2.0	12.4
	8	48.8	75.2	84.7	2.0	12.5
	12	49.2	75.4	84.2	2.0	12.4
C_t	2	**49.6**	75.5	**84.9**	2.0	12.5
$(C_v = 3)$	4	49.1	75.4	84.4	2.0	13.3

Table 3. Ablation study for the number of extra learnable tokens.

egy with one frame per second sampling rate is employed. The dimension of video (text) distributions is 512 by default. Following [39, 20, 37], we train UATVR model for 5 epochs with Adam [28] optimizer and adopt a warmup [21] strategy. We set the batch size as 64 and the initial learning rate as 5e-5. The coefficient is 1e-2 for α and 1e-4 for β.

4.2. Ablation Study

We evaluate the effectiveness of different components in UATVR by comprehensive experiments. The default visual encoder is ViT-B/16 [14] and the *t2v* retrieval results are reported on the widely-used MSR-VTT [57] dataset.

Uncertainty-Adaptive Matching. The baseline of UATVR is fine-grained token-wise interaction (TI), which provides minimal granularity tokens for retrieval. As shown in Tab. 1, with extra multi-class learnable tokens appended, we observe a 1.2% R@1 improvement (48.4% *vs.* 49.6%) and a lower MnR (14.1 *vs.* 12.5) compared to the baseline. We explain that additional tokens can aggregate multi-grained information extracted from video frames and textual words, which adapts to flexible cross-modal matching in different granularities. Moreover, distribution-based uncertainty adaptation with KL divergence constraint further obtains the highest 50.8% R@1 and the lowest 1.0 MdR and 12.4 MnR, which demonstrates the effectiveness of the proposed distribution alignment mechanism. Also, we formulate a soft contrastive loss following [45], which is a binary classification loss based on the softmax cross-entropy via Monte-Carlo estimation (marked as *). We observe very close but slightly lower performance than MIL-NCE (multi-instance contrast, marked as †). Further analysis is shown in the appendix. All the above results prove the significance of distribution-based cross-modal matching in tackling the proposed uncertainty problem.

Prob-Embeds #		R@1	R@5	R@10	MdR↓	MnR↓
$(C_v = 3, C_t = 2)$		49.6	75.5	84.9	2.0	12.5
	1	49.6	76.5	84.3	2.0	12.5
	3	49.8	76.1	84.9	2.0	12.9
K	5	50.5	**77.1**	84.4	**1.0**	12.6
	7	**50.8**	76.3	**85.5**	**1.0**	**12.4**
	9	-	-	-	-	-

Table 4. Ablation for the number of probabilistic embeddings.

Frames #	R@1	R@5	R@10	MdR↓	MnR↓
4	44.9	74.0	82.2	2.0	15.1
8	50.6	76.0	83.9	**1.0**	12.9
12	50.8	76.3	**85.5**	**1.0**	**12.4**
16	**51.0**	**76.4**	**85.5**	**1.0**	13.5
20	50.8	76.0	84.9	**1.0**	13.4

Table 5. Impact of visual frame numbers.

Tab. 2 compares time complexity and params for each component, in which feature dimension D is omitted for brevity. Despite the relatively larger time complexity taken by DSA and DUA, it is still limited in quadratic time for one text-video pair. Note that our extra learnable tokens do not bring more model parameters, meanwhile having negligible additional training cost. A similar conclusion can be drawn for the distribution-based matching module. Therefore, our proposed UATVR is simple-yet-effective.

Dynamic Semantic Adaptation Tokens. In Tab. 3, we study the impact of additional appended C_v and C_t multi-class learnable tokens. It shows distinct *t2v* retrieval improvements once extra visual tokens are added (*i.e.*, $C_v > 0$), which reflects that this simple-yet-effective technology can aggregate multi-grained video semantics depending on the uncertain captions. $C_v = 3$ is the best among all. When C_v is larger than 3, the performance starts to degrade. We observe a similar phenomenon in TMVM [34] that more video prototypes would significantly degrade the performance. Too many extra tokens would become noise rather than representative, negatively influencing the normal matching process. In subsequent experiments, we set final $C_v = 3$ and $C_t = 2$. Additional C_t text tokens further promote *v2t* R@1 to 47.8%. We analyze the corresponding *v2t* retrieval results in the appendix.

Distribution-Based Uncertainty Adaptation. A larger number of probabilistic embeddings can better simulate video and caption distributions but can also lead to more computing requirements. In Tab. 4, we report the performance according to the number of sampled probabilistic embeddings K based on our optimal DSA branch. We

Method	Date	Text → Video					Video → Text				
		R@1	R@5	R@10	MdR↓	MnR↓	R@1	R@5	R@10	MdR↓	MnR↓
MMT [17]	ECCV'20	26.6	57.1	69.6	4.0	-	27.0	57.5	69.7	3.7	21.3
SupportSet [46]	ICLR'21	30.1	58.5	69.3	3.0	-	28.5	58.6	71.6	3.0	-
Frozen [4]	ICCV'21	32.5	61.5	71.2	3.0	-	-	-	-	-	-
BridgeFormer [19]	CVPR'22	37.6	64.8	75.1	-	-	-	-	-	-	-
TMVM [34]	NeurIPS'22	36.2	64.2	75.7	3.0	-	34.8	63.8	73.7	3.0	-
CLIP-ViT-B/32											
CLIP4Clip [39]	ArXiv'21	44.5	71.4	81.6	2.0	15.3	42.7	70.9	80.6	2.0	11.6
CenterCLIP [62]	SIGIR'22	44.2	71.6	82.1	2.0	15.1	42.8	71.7	82.2	2.0	10.9
CAMoE [11]‡	ArXiv'21	44.6	72.6	81.8	2.0	13.3	45.1	72.4	83.1	2.0	10.0
CLIP2Video [16]	ArXiv'21	45.6	72.6	81.7	2.0	14.6	43.5	72.3	82.1	2.0	10.2
X-Pool [20]	CVPR'22	46.9	72.8	82.2	2.0	14.3	-	-	-	-	-
QB-Norm [5]‡	CVPR'22	47.2	73.0	83.0	2.0	-	-	-	-	-	-
EMCL [23]	NeurIPS'22	46.8	73.1	83.1	2.0	-	46.5	73.5	83.5	2.0	-
TS2-Net [37]	ECCV'22	47.0	**74.5**	**83.8**	2.0	13.0	45.3	**74.1**	83.7	2.0	9.2
UATVR		**47.5**	73.9	83.5	2.0	**12.3**	**46.9**	73.8	**83.8**	2.0	**8.6**
UATVR‡		**49.8**	**76.1**	**85.5**	2.0	**12.9**	**51.1**	**74.8**	**85.1**	**1.0**	**8.3**
CLIP-ViT-B/16											
CLIP2TV [18]	ArXiv'21	48.3	74.6	82.8	2.0	14.9	46.5	75.4	84.9	2.0	10.2
CenterCLIP [62]	SIGIR'22	48.4	73.8	82.0	2.0	13.8	47.7	75.0	83.3	2.0	10.2
TS2-Net [37]	ECCV'22	49.4	75.6	85.3	2.0	13.5	46.6	75.9	84.9	2.0	8.9
UATVR		**50.8**	**76.3**	**85.5**	**1.0**	**12.4**	**48.1**	**76.3**	**85.4**	2.0	**8.0**
UATVR‡		**53.5**	**79.5**	**88.1**	**1.0**	**10.2**	**54.5**	**79.1**	**87.9**	**1.0**	**7.6**

Table 6. *t2v* and *v2t* comparisons on MSR-VTT [57]. ‡ denotes using inverted dual softmax [11] or QB-Norm [5] for post-processing.

Method	R@1	R@5	R@10	MdR↓	MnR↓
HGR [9]	35.1	73.5	83.5	2.0	-
CLIP [48]	39.7	72.3	82.2	2.0	12.8
SUPPORT [46]	44.9	82.1	89.7	1.0	-
CLIP4Clip [39]	55.9	89.2	95.0	1.0	3.9
Clip2Video [16]	57.3	90.0	95.5	1.0	3.6
QB-Norm [5]	58.8	88.3	93.8	1.0	-
TS2-Net [37]	59.1	90.0	95.2	1.0	3.5
UATVR(ViT-B32)	61.3	91.0	95.6	1.0	3.3
UATVR(ViT-B16)	**64.5**	**92.6**	**96.8**	1.0	**2.8**

Table 7. *t2v* comparisons on the **VATEX** [54] dataset.

Method	R@1	R@5	R@10	MdR↓	MnR↓
ClipBERT [30]	20.4	48.0	60.8	6.0	-
TT-CE [13]	21.6	48.6	62.9	6.0	-
Frozen [4]	31.0	59.8	72.4	3.0	-
TMVM [34]	36.5	64.9	75.4	3.0	-
CLIP4Clip [39]	42.8	68.5	79.2	2.0	18.9
TS2-Net [37]	41.8	71.6	82.0	2.0	14.8
UATVR(ViT-B32)	43.1	71.8	82.3	2.0	15.1
UATVR(ViT-B16)	**45.8**	**73.7**	**83.3**	2.0	**13.5**

Table 8. *t2v* comparisons on the **DiDeMo** [1] dataset.

find that *t2v* retrieval performance increases as K increases. When K is larger than 7, the performance starts to saturate. Considering the computational costs, we choose $K = 7$ finally. Moreover, we have attempted sampling different K for text and video distributions in the appendix and obtain a similar conclusion to Tab. 4. Overall, our DUA surpasses the baseline by a large margin, which demonstrates the effectiveness of the distribution alignment mechanism.

In Fig. 3, we measure the inherent uncertainty (geometric mean over the $\sigma \in \mathbb{R}^D$) of test set texts and videos and report the R@1 performance in each epoch. We show comparisons on two visual encoders to analyze the correlation between the uncertainty and the discriminability of learned representations. Generally, we observe performance improvements with decreasing uncertainty, which verifies the positive effects of distribution-based uncertain adaptation.

The Number of Visual Frames. The impact of frame numbers is studied in Tab. 5. UATVR achieves a decent 50.6% R@1 with only 8 frames. The performance starts to saturate with more than 12 frames. Here, we only use 12 frames by default for fair comparisons with others.

4.3. Comparison with State-of-the-arts

To evaluate the generalization of our uncertainty-adaptive models, we compare UATVR with SOTA methods on various text-video retrieval benchmarks, including MSR-VTT [57], MSVD [56], VATEX [54], and DiDeMo [1].

Tab. 6 shows detailed comparisons on MSR-VTT `test 1k-A` set. We divide current approaches into Training-

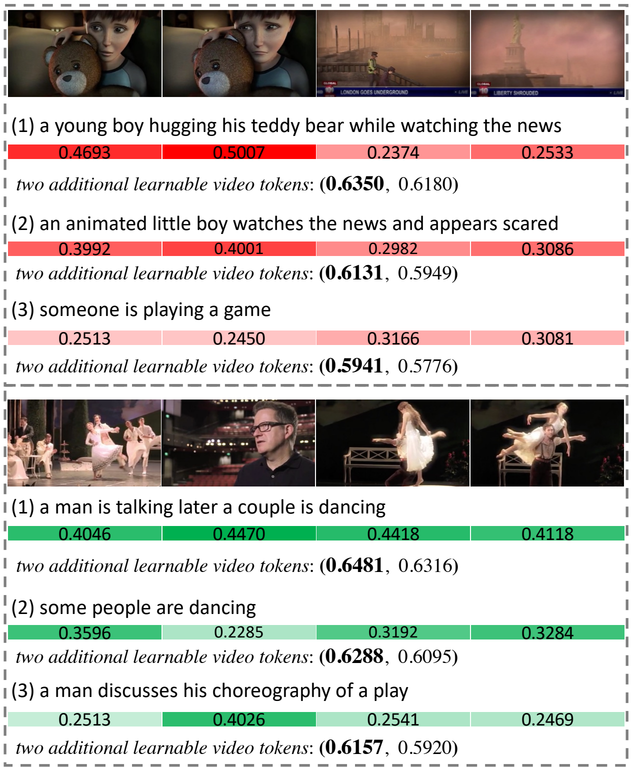
(1) a young boy hugging his teddy bear while watching the news
0.4693 0.5007 0.2374 0.2533
two additional learnable video tokens: (0.6350, 0.6180)
(2) an animated little boy watches the news and appears scared
0.3992 0.4001 0.2982 0.3086
two additional learnable video tokens: (0.6131, 0.5949)
(3) someone is playing a game
0.2513 0.2450 0.3166 0.3081
two additional learnable video tokens: (0.5941, 0.5776)
(1) a man is talking later a couple is dancing
0.4046 0.4470 0.4418 0.4118
two additional learnable video tokens: (0.6481, 0.6316)
(2) some people are dancing
0.3596 0.2285 0.3192 0.3284
two additional learnable video tokens: (0.6288, 0.6095)
(3) a man discusses his choreography of a play
0.2513 0.4026 0.2541 0.2469
two additional learnable video tokens: (0.6157, 0.5920)

Figure 4. Impact of additional learnable tokens. Darker background color denotes higher frame attention.

from-scratch (upper rows) and CLIP-Driven [48]. Transferring knowledge from CLIP has distinctly surpassed models w/o initialization, which demonstrates that spatial semantics learned from image-text pairs are essential for the TVR task. Our proposed UATVR falls into the CLIP-Driven paradigm. For the ViT-B/32 encoder, UATVR obtains higher R@1 performance than the previous best method (47.5% vs. 47.2% in *t2v* retrieval, and 46.9% vs. 46.5% in *v2t* retrieval). The improvement is more significant on the ViT-B/16 backbone. Specifically, UATVR outperforms previous best TS2-Net [37] by 1.4% in *t2v* and 1.5% in *v2t* retrieval, yielding a remarkable *t2v* R@1 50.8%. Notice that our method firstly reduces the MdR metric from 2.0 to 1.0 and has the lowest 12.4 MnR, which means UATVR is more robust to wrong retrieval samples. In the appendix, we further report results with dual softmax learning (DSL) operation. Our results again surpass methods with post-processing operations like QB-Norm [5] and CAMoE [11].

Moreover, we conduct evaluations on multiple other TVR benchmarks, including VATEX [54] in Tab. 7, DiDeMo [1] in Tab. 8, and MSVD [56] in Tab. 9. Despite possible sub-optimal hyper-parameters (e.g., C_v, C_t, K) for the specific dataset, UATVR achieves consistent improvements across various datasets, *e.g.*, 61.3% *vs.* 59.1% for VATEX, and 43.1% *vs.* 42.8% for DiDeMo. UATVR outperforms SOTAs by a large margin with better-performed ViT-B/16. Prominent results demonstrate good generalization and robustness of our dynamic semantic-aggregation and distribution-based uncertainty adaption paradigms.

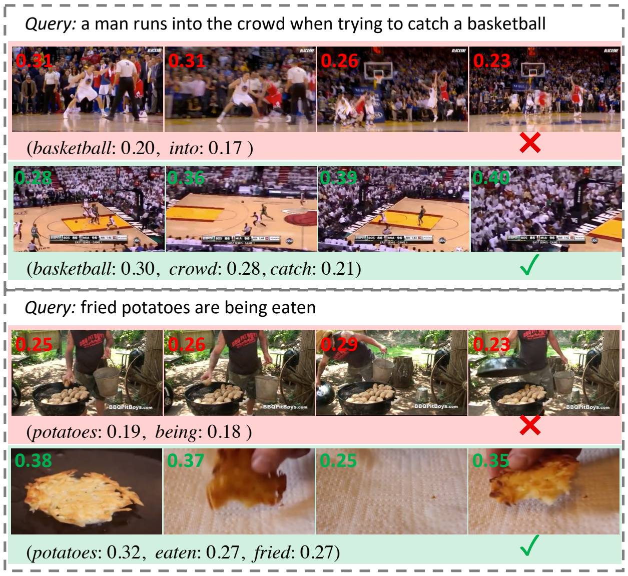
Query: a man runs into the crowd when trying to catch a basketball
0.31 0.31 0.26 0.23
(basketball: 0.20, into: 0.17)
✗
0.28 0.36 0.39 0.40
(basketball: 0.30, crowd: 0.28, catch: 0.21)
✓
Query: fried potatoes are being eaten
0.25 0.26 0.29 0.23
(potatoes: 0.19, being: 0.18)
✗
0.38 0.37 0.25 0.35
(potatoes: 0.32, eaten: 0.27, fried: 0.27)
✓

Figure 5. Visualization of TVR results and attention weights for each frame and significant words. Red: incorrect results of the token-wise baseline model. Green: correct results of our UATVR.

Method	R@1	R@5	R@10	MdR↓	MnR↓
CE [36]	19.8	49.0	63.8	6.0	-
SUPPORT [46]	28.4	60.0	72.9	4.0	-
CLIP [48]	37.0	64.1	73.8	3.0	-
Frozen [4]	33.7	64.7	76.3	3.0	-
TMVM [34]	36.7	67.4	81.3	2.5	-
CLIP4Clip [39]	45.2	75.5	84.3	2.0	10.3
X-Pool [20]	47.2	77.4	86.0	2.0	9.3
UATVR(ViT-B32)	46.0	76.3	85.1	2.0	10.4
UATVR(ViT-B16)	**49.7**	**79.0**	**87.3**	2.0	**8.9**

Table 9. *t2v* comparisons on the **MSVD** [56] dataset.

4.4. Qualitative Results

To better understand the impact of additional learnable tokens, we show specific attention weights computed by Eq.8 for each frame and the extra tokens under different text descriptions in *v2t* retrieval. As shown in Fig. 4, our model shows higher weights on text-related frames, such as "*some people are dancing*" and "*a man discusses his choreography of a play*", resulting in quite a contrary frame attention in the second video. Moreover, our additional tokens assign higher attention scores to more accurate texts, benefiting text-depended video semantic aggregation during cross-modal matching. In Fig. 5, we show correct UATVR *t2v* retrieval results compared to the token-wise baseline. Attention weights for each frame and the most significant words are highlighted. Due to the matching uncertainty, token-wise baseline easily falls into local context matching with specific queries like '*basketball*' and '*potatoes*'. Nevertheless, UATVR retrieves correct videos with multi-grained high-level reasoning, showing advance in recognizing subtle clues and global semantics simultaneously. We show further visualization and analysis in the appendix.

5. Conclusion

In this work, we analyze the uncertain matching problem in existing multi-grained text-video retrieval and propose a novel uncertainty-adaptive matching framework (UATVR) in complementary deterministic and probabilistic views. We model each text-video lookup as a distribution matching procedure by introducing semantic aggregation learnable tokens and distribution-based probabilistic embeddings. UATVR adaptively addresses the uncertain matching problem and formulates realistic one-to-many text-video correspondences. Thorough ablation studies and remarkable performance demonstrate the effectiveness of UATVR. We leave more sophisticated and refined distribution modelling, like a Mixture of Gaussians, as part of future work.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. 2, 5, 7, 8

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 2

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 2, 5, 7, 8

[5] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *CVPR*, pages 5194–5205, 2022. 7, 8

[6] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *CVPR*, pages 2917–2927, 2022. 2

[7] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *CVPR*, pages 5710–5719, 2020. 3

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021. 2

[9] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, pages 10638–10647, 2020. 2, 7

[10] Yizhen Chen, Jie Wang, Lijian Lin, Zhongang Qi, Jin Ma, and Ying Shan. Tagging before alignment: Integrating multi-modal tags for video-text retrieval. *arXiv preprint arXiv:2301.12644*, 2023. 2

[11] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 7, 8

[12] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, pages 8415–8424, 2021. 2, 3, 4, 5

[13] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *ICCV*, pages 11583–11593, 2021. 7

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[15] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Dongliang He, and Weipinng Wang. Mamico: Macro-to-micro semantic correspondence for self-supervised video representation learning. In *ACM MM*, pages 1348–1357, 2022. 2

[16] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2, 7

[17] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, pages 214–229. Springer, 2020. 2, 5, 7

[18] Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*, 2021. 2, 7

[19] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *CVPR*, pages 16167–16176, 2022. 7

[20] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, pages 5006–5015, 2022. 2, 3, 6, 7, 8

[21] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 6

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1, 2

[23] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-

language representations. *arXiv preprint arXiv:2211.11427*, 2022. 7

[24] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *CVPR*, pages 2472–2482, 2023. 2

[25] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video retrieval with disentangled conceptualization and set-to-set alignment. *arXiv preprint arXiv:2305.12218*, 2023. 2

[26] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. *arXiv preprint arXiv:2303.09867*, 2023. 2

[27] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[30] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021. 2, 5, 7

[31] Hao Li, Peng Jin, Zesen Cheng, Songyang Zhang, Kai Chen, Zhennan Wang, Chang Liu, and Jie Chen. Tg-vqa: Ternary game of video question answering. *arXiv preprint arXiv:2305.10049*, 2023. 2

[32] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705, 2021. 2

[33] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2

[34] Chengzhi Lin, Ancong Wu, Junwei Liang, Jun Zhang, Wenhang Ge, Wei-Shi Zheng, and Chunhua Shen. Text-adaptive multiple visual prototype matching for video-text retrieval. *arXiv preprint arXiv:2209.13307*, 2022. 2, 3, 6, 7, 8

[35] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 2

[36] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 8

[37] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, pages 319–335. Springer, 2022. 2, 6, 7, 8

[38] Dezhao Luo, Yu Zhou, Bo Fang, Yucan Zhou, Dayan Wu, and Weiping Wang. Exploring relations in untrimmed videos for self-supervised learning. *TOMM*, 18(1s):1–21, 2022. 2

[39] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2, 3, 4, 5, 6, 7, 8

[40] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, pages 638–647, 2022. 2, 3

[41] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020. 2, 5

[42] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019. 2, 5

[43] Shaobo Min, Weijie Kong, Rong-Cheng Tu, Dihong Gong, Chengfei Cai, Wenzhe Zhao, Chenyang Liu, Sixiao Zheng, Hongfa Wang, Zhifeng Li, et al. Hunyuan_tvr for text-video retrivial. *arXiv preprint arXiv:2204.03382*, 2022. 2, 3

[44] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 3

[45] Seong Joon Oh, Kevin P Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C Gallagher. Modeling uncertainty with hedged instance embeddings. In *ICLR*, 2018. 3, 5, 6

[46] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 7, 8

[47] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer, 2021. 2

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 5, 7, 8

[49] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, pages 6902–6911, 2019. 3

[50] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *ECCV*, pages 53–70. Springer, 2020. 3

[51] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2

[52] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022. 2, 3

[53] Wenzhe Wang, Mengdan Zhang, Runnan Chen, Guanyu Cai, Penghao Zhou, Pai Peng, Xiaowei Guo, Jian Wu, and Xing Sun. Dig into multi-modal cues for video retrieval with hierarchical alignment. In *IJCAI*, pages 1113–1121, 2021. 2

[54] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4581–4591, 2019. 2, 5, 7, 8

[55] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *CVPR*, pages 10704–10713, 2023. 2

[56] Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. Deep learning for video classification and captioning. In *Frontiers of multimedia research*, pages 3–29. 2017. 2, 5, 7, 8

[57] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 2, 5, 6, 7

[58] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057. PMLR, 2015. 2

[59] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2021. 2, 3

[60] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[61] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1, 2

[62] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *SIGIR*, pages 970–981, 2022. 2, 7