UATVR: Uncertainty-Adaptive Text-Video Retrieval

Bo Fang\(^1\)* Wenhao Wu\(^2\),* Chang Liu\(^4\)* Yu Zhou\(^1\†\) Yuxin Song\(^3\)

\(^1\)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.

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 Uncertainty-Adaptive Text-Video Retrieval (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 chorographed 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 has 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 $s(\cdot, \cdot)$, 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 $t_i = [w_i^0, w_i^1, \cdots, w_i^N]^T$, $v_i = [f_i^1, f_i^2, \cdots, f_i^M]^T$, 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 $t_i = [w_i^0, w_i^1, \cdots, w_i^N]^T$ and $v_i = [f_i^1, f_i^2, \cdots, f_i^M]^T$. The frame embedding $f_i^m$ 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 $w_i^0$ in Eq. 2. The similarity of the text-video is calculated as the inner production of $t_i, v_i$, c.f. Eq. 3.

\[
v_i = \text{meanPool}([f_i^1, f_i^2, \cdots, f_i^M]^T), \tag{1}
\]

\[
t_i = w_i^0, \tag{2}
\]

\[
s(t_i, v_i) = \langle t_i, v_i \rangle. \tag{3}
\]

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:

\[
L_{t2v} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s(t_i, v_i))}{\sum_{j=1}^{B} \exp(s(t_i, v_j))}, \tag{4}
\]

\[
L_{v2t} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s(v_i, t_i))}{\sum_{j=1}^{B} \exp(s(v_i, t_j))}, \tag{5}
\]

\[
L = L_{t2v} + L_{v2t}. \tag{6}
\]

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:

\[
s(t_i, v_i) = \frac{1}{2} \left( \sum_{n=1}^{N} \max_{m=1}^{M} \langle w_i^n, f_i^m \rangle + \sum_{m=1}^{M} \max_{n=1}^{N} \langle w_i^n, f_i^m \rangle \right), \tag{7}
\]

where $M, N$ denote frame and word number in the $i$th sample pair. $w_i^n, f_i^m$, which are channel-wise normalized before calculating, refer to the $n$th word embedding and $m$th 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, vanilla 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 $\{f_i^m\}_{m=1}^{M}$ and word embeddings $\{w_i^n\}_{n=1}^{N}$ extracted from backbone encoders,
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:

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

We define our dynamic semantic adaption loss $\mathcal{L}_{\text{DSA}}$ 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 $t_i$, $v_i$ 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:

$$p(z|t_i) \sim \mathcal{N}(h^\mu_{t_i}(t_i), \text{diag}(h^\sigma_{t_i}(t_i))),$$
$$p(z|v_i) \sim \mathcal{N}(h^\mu_{v_i}(v_i), \text{diag}(h^\sigma_{v_i}(v_i))),$$

where head module $h^\mu$ is a fully-connected layer followed by LayerNorm [3] and $l_2$ normalization, and head $h^\sigma$ is a separate fully-connected layer without any normalization following [12]. The textual head $h_T$ and visual head $h_V$ share the same parametric structure but optimize independently upon transformer-based feature output. Next, two groups of $K$ probabilistic embeddings $\{t^{(1)}_i, \cdots, t^{(K)}_i\}$ and $\{v^{(1)}_i, \cdots, v^{(K)}_i\}$ 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, formu-
lated as follows:
\[
\begin{align*}
t_i^{(k)} &= \sigma(t_i) \cdot e^k + \mu(t_i), \\
v_i^{(k)} &= \sigma(v_i) \cdot e^k + \mu(v_i),
\end{align*}
\]
where $\epsilon^{(k)} \sim \mathcal{N}(0, I)$ and $\mu, \sigma$ 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 $t_i \in \{t_i^{(k)}\}_{k=1}^K$, we define the positive set $\mathcal{P}_i$ for $t_i$ as all video probabilistic embeddings from $v_i$, formulated as $\mathcal{P}_i = \{v_i^{(k)}\}_{k=1}^K$. The negative set thus is formed as probabilistic embeddings from other videos in the batch, $\mathcal{I}_i = \{v_j^{(k)}\}_{j,k,i}$. We define the distribution-based uncertainty adaptation loss as:
\[
\mathcal{L}_{\text{DUA}} = - \frac{1}{B} \sum_i \log \frac{\sum_{v_i \in \mathcal{P}_i} \exp(s(t_i, v_i))}{\sum_{v_j \in \mathcal{P}_i \cup \mathcal{I}_i} \exp(s(t_i, v_j))}.
\]

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 constrain the learned variances from collapsing to zero, which can be formulated as:
\[
\mathcal{L}_{\text{KL}} = \text{KL}(p(z|t_i), \mathcal{N}(0, I)) + \text{KL}(p(z|v_i), \mathcal{N}(0, I)).
\]

Therefore, the total objectives can be defined as:
\[
\mathcal{L} = \mathcal{L}_{\text{DSA}} + \alpha \cdot \mathcal{L}_{\text{DUA}} + \beta \cdot \mathcal{L}_{\text{KL}},
\]

where $\alpha$ and $\beta$ 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 [39] contains 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 [39] 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.

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.
<table>
<thead>
<tr>
<th>Method</th>
<th>Date</th>
<th>Text → Video</th>
<th>Video → Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>MMT [17]</td>
<td>ECCV’20</td>
<td>26.6</td>
<td>57.1</td>
</tr>
<tr>
<td>SupportSet [46]</td>
<td>ICLR’21</td>
<td>30.1</td>
<td>58.5</td>
</tr>
<tr>
<td>Frozen [4]</td>
<td>ICCV’21</td>
<td>32.5</td>
<td>61.5</td>
</tr>
<tr>
<td>BridgeFormer [19]</td>
<td>CVPR’22</td>
<td>37.6</td>
<td>64.8</td>
</tr>
<tr>
<td>TMVM [34]</td>
<td>NeurIPS’22</td>
<td>36.2</td>
<td>64.2</td>
</tr>
</tbody>
</table>

CLIP-ViT-B/32

<table>
<thead>
<tr>
<th>Method</th>
<th>Date</th>
<th>Text → Video</th>
<th>Video → Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>HGR [9]</td>
<td></td>
<td>44.5</td>
<td>71.4</td>
</tr>
<tr>
<td>CLIP4Clip [39]</td>
<td>ArXiv’21</td>
<td>44.7</td>
<td>71.6</td>
</tr>
<tr>
<td>CenterCLIP [62]</td>
<td>SIGIR’22</td>
<td>44.6</td>
<td>72.6</td>
</tr>
<tr>
<td>CLIP2Video [16]</td>
<td>ArXiv’21</td>
<td>45.6</td>
<td>72.6</td>
</tr>
<tr>
<td>X-Pool [20]</td>
<td>CVPR’22</td>
<td>46.9</td>
<td>72.8</td>
</tr>
<tr>
<td>QB-Norm [5]</td>
<td>CVPR’22</td>
<td>47.2</td>
<td>73.0</td>
</tr>
<tr>
<td>EMCL [23]</td>
<td>NeurIPS’22</td>
<td>46.8</td>
<td>73.1</td>
</tr>
<tr>
<td>TS2-Net [37]</td>
<td>ECCV’22</td>
<td>47.0</td>
<td>74.5</td>
</tr>
<tr>
<td>UATVR</td>
<td></td>
<td>47.5</td>
<td>73.9</td>
</tr>
<tr>
<td>UATVR†</td>
<td></td>
<td>49.8</td>
<td>76.1</td>
</tr>
</tbody>
</table>

CLIP-ViT-B/16

<table>
<thead>
<tr>
<th>Method</th>
<th>Date</th>
<th>Text → Video</th>
<th>Video → Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>ClipBERT [30]</td>
<td></td>
<td>48.3</td>
<td>74.6</td>
</tr>
<tr>
<td>CenterCLIP [62]</td>
<td>SIGIR’22</td>
<td>48.4</td>
<td>73.8</td>
</tr>
<tr>
<td>TS2-Net [37]</td>
<td>ECCV’22</td>
<td>49.4</td>
<td>75.6</td>
</tr>
<tr>
<td>UATVR</td>
<td></td>
<td>50.8</td>
<td>76.3</td>
</tr>
<tr>
<td>UATVR†</td>
<td></td>
<td>53.5</td>
<td>79.5</td>
</tr>
</tbody>
</table>


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

Table 7. \( t^2 \)v comparisons on the VATEX [54] dataset.

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

Table 8. \( t^2 \)v comparisons on the DiDeMo [1] dataset.

find that \( t^2 \)v 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-
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 r2v 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 r2v retrieval and 1.5\% in v2t retrieval, yielding a remarkable r2v 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., \(C_v, C_f, K\)) 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

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

Table 9. r2v comparisons on the MSVD \[56\] dataset.

ViT-B/16. Prominent results demonstrate good generalization and robustness of our dynamic semantic-aggregation and distribution-based uncertainty adaption paradigms.

### 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 r2v 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,
References


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


[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


[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


[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


