Text-Adaptive Multiple Visual Prototype Matching for Video-Text Retrieval

Anonymous Author(s)
Affiliation
Address
email

Abstract

Cross-modal retrieval between videos and texts has gained increasing interest because of the rapid emergence of videos on the web. Generally, a video contains rich instance and event information and the query text only describes a part of the information. Thus, a video can have multiple different text descriptions and queries. We call it the Video-Text Correspondence Ambiguity problem. Current techniques mostly concentrate on mining local or multi-level alignment between contents of video and text (e.g., object to entity and action to verb). It is difficult for these methods to alleviate video-text correspondence ambiguity by describing a video using only one feature, which is required to be matched with multiple different text features at the same time. To address this problem, we propose a Text-Adaptive Multiple Visual Prototype Matching Model. It automatically captures multiple prototypes to describe a video by adaptive aggregation on video token features. Given a query text, the similarity is determined by the most similar prototype to find correspondence in the video, which is called text-adaptive matching. To learn diverse prototypes for representing the rich information in videos, we propose a variance loss to encourage different prototypes to attend to different contents of the video. Our method outperforms state-of-the-art methods on four public video retrieval datasets.

1 Introduction

Cross-modal retrieval has gained popularity as a means of matching semantically similar samples across multiple modalities (14; 13; 2). With the rapid increase in the number of videos, Text-Video Retrieval (TVR) has garnered a lot of interest. Given a text description/video, we aim to retrieve the corresponding video/text description in the gallery. TVR remains a challenging task since the video and text have different semantical descriptions which results in modality gap.

To overcome the modality gap issue, most existing methods generally follow a dual encoding network paradigm which encodes video and text through a visual encoder and a language encoder into the same embedding space, respectively (14; 13; 2; 10; 15). And then they focus on learning a similarity function between the representations of text and video, which provides the basis for ranking the gallery video/text. However, the modality gap issue is still not well solved due to the misalignment. Some methods aim to extract additional information from the video to get more fine-grained video features to be aligned with corresponding text features (10; 29; 16). They focus on properly combining different video modalities, such as appearance, audio, speech, etc. Some methods pay attention to multi-level alignment, including text entity to video object, text verb to video action, and text phrase to video event(15; 31). However, most of them ignore the fact that a video contains rich instance and event information, while the query text only describes a part of the information, which leads to video-text correspondence ambiguity problem. Concretely, as shown in Figure 1, due to the multiple characters in the video, the video has different text descriptions: “a woman does something”, “‘a man

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

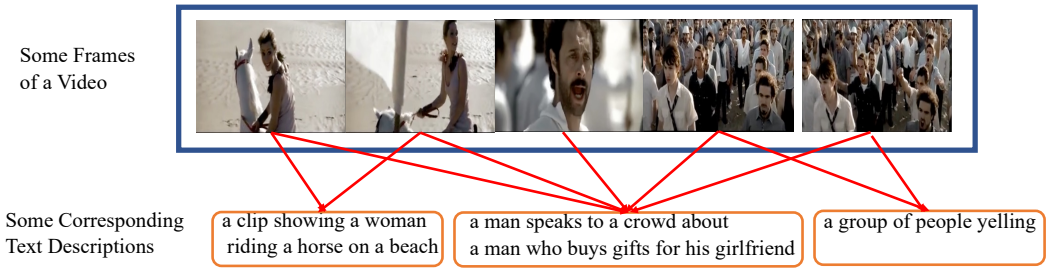

Figure 1: An example of text-to-video retrieval. It shows a video and different text descriptions. The characters in the text descriptions are different. It is because different person focus on different characters in the video and provide different text descriptions. In this paper, multiple video prototypes matching is utilized to help mitigate this issue.

does something" or "a group does something". The text descriptions are very diverse and scattered in the feature space. With diverse video contents , it is infeasible to align a single video feature with the diverse text features. Neither additional information-based methods nor multi-level alignment-based methods consider such correspondence ambiguity since both of them match a single video feature with multiple text features.

In this work, we focus on video-text retrieval problem and aim to mine fine-grained and diverse event descriptions in a video and adaptively match them with multiple text descriptions to mitigate correspondence ambiguity issue for better alignment. Specifically, we automatically capture multiple prototypes for each video to account for its rich instance and event information by aggregating on token features. Based on multiple prototypes of a video, we aim at finding many-to-many correspondence between prototypes and text representations to reduce ambiguity. To this end, we propose a Text-Adaptive Multiple Visual Prototype Matching Model. We follow the dual encoding network paradigm as previous methods(2). First, we input video into visual encoder to get the video token features. Then we employ a token-adaptive aggregation method to generate several visual prototypes from the final video token features. As shown in Figure 3, one prototype is the class token and the others are the linear combinations of all the token features. The text feature encoded by the language encoder is matched with the most similar video prototype. This text-adaptive matching can help find text-to-prototype correspondence in video and alleviate the correspondence ambiguity. To further force the prototypes to represent rich information in the video, we propose a variance loss to enhance the diversity of prototypes to attend to different contents of the video.

Extensive experiments were conducted on the MSR-VTT (30), DIDEMO(11), MSVD(5), and LSMDC (24) to evaluate video-text bi-directional retrieval performance. Experimental results show that our proposed model can achieve state-of-the-art results on all the above datasets. Ablation studies were carried out to evaluate the effectiveness of each single part of our model.

The contributions of our work are three folds. 1) We propose to automatically generate multiple prototypes for video to account for its rich information and introduce a text-adaptive matching strategy to adaptively find correspondence between texts and prototypes. 2) We further design a variance loss to effectively encode diverse information into prototypes to attend to different contents in the video and boost alignment. 3) Our proposed model can achieve new state-of-the-art results on four benchmark datasets. Extensive experiments demonstrate its effectiveness and generalization.

2 Related Work

Text-Video Retrieval. The methods for text-video retrieval can be classified into two categories: pre-extracted feature based (16; 29; 15; 19) and end-to-end training (2; 13). Given the small size of most video-text retrieval datasets, the dominant paradigm for video retrieval has been to combine pre-extracted features from "expert" models, such as models trained for a variety of tasks and on multiple modalities such as face, scene, and object recognition, action classification, and sound classification. CE (16) and MMT (10) compute the overall similarity for a video-text pair obtained as a weighted sum of each expert's similarity with the text. DMM (29) uses hierarchical video-text alignment to explore the diverse properties in multi-modal cues for fine-grained alignment. Another

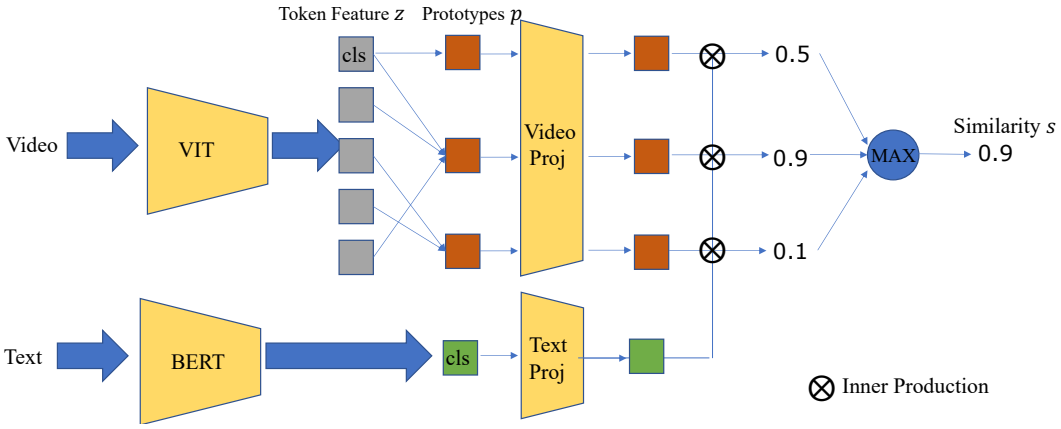

Figure 2: Workflow of our method. We feed video into vision transformer to get the token features. Then we employ a token-adaptive aggregation method to generate several visual prototypes from the final video token features. The text feature from the language model is matched with the video prototype which is most similar to text after the projection layer.

important paradigm is to first train on a large-scale vision-and-language dataset and then finetune on downstream dataset. ClipBERT (13) proposes an efficient end-to-end technique based on sparse sampling and demonstrates that the pretrained model by image-language dataset enables superior video-text retrieval initialization. Frozen (2) proposes an end-to-end trainable model that can benefit from large-scale image and video captioning datasets. Most of them, however, ignore the fact that a video contains rich instance and event information, whereas the query text only describes a portion of the information, resulting in video-text correspondence ambiguity. We propose a Text-Adaptive Multiple Visual Prototype Matching Model to help mitigate this issue.

Vision Transformer. Now vision transformer becomes the visual basic backbone in most of the state-of-the-are methods for vision-language tasks (22; 13; 2; 14). BLIP(14) makes effective use of noisy web data by bootstrapping captions, in which a captioner generates synthetic captions and a filter removes the noisy ones. Frozen(14) makes a minor modification to the Divided Space-Time attention introduced by (4). Our model is based on the dual framework of Frozen.

3 Method

A video, in general, contains a lot of rich instance and event information, and the text description only describes a portion of it. A video can have many characters/events/actions, whereas a text typically only describes one character, action, or event. So mining representation of multiple essential events facilitates the text-video matching by reducing the modality gap properly. As a result, a model that generates multiple visual prototypes is better than that generates one feature. With this rationale in mind , we propose our Text-Adaptive Multiple Visual Prototype Matching Model. To generate several visual prototypes from video, we use a token-adaptive aggregation method on the final video token features. Given a query text, the most similar prototype is used to find correspondence in the video, a process called as text-adaptive matching. We also propose a variance loss to encourage different prototypes to attend to different contents of the video in order to learn diverse prototypes for representing the rich information in videos. We define the problem in Section 3.1, present Text-Adaptive Multiple Visual Prototype Matching Model in Section 3.2, describe our variance loss in Section 3.3, and explain our dual encoder in Section 3.4.

3.1 Problem Definition

For the video-text retrieval task, we are given M videos $V = \{V_i\}_{i=0}^{M-1}$ and N text descriptions $T = \{T_i\}_{i=0}^{N-1}$. This function should maximize positive cross-modal sample similarity while minimizing negative cross-modal sample similarity. The goal of this task is to learn a similarity function $S : (T, V) \rightarrow \mathbf{R}$. It should maximize positive cross-modal sample similarity while minimizing

negative cross-modal sample similarity. In general, we fed video and text to into an vision encoder and language encoder to get normalized feature f^v, f^v. Then the similarity function S is defined by

$$S_{t,v} = \langle f^t, f^v \rangle. \tag{1}$$

The similarity of text-video is calculated as the inner production of f^v, f^t.

3.2 Text-Adaptive Multiple Visual Prototype Matching Model

The similarity function of Equation 1 ignores the Video-Text Correspondence Ambiguity problem: some videos have multiple different text descriptions and queries. To furthermore illustrate this problem, we show the text descriptions diversity of videos by showing the inter-text similarity and the minimum of intra-similarity distribution in 3.3. We define inter-text similarity as the similarity of different text descriptions for different videos, intra-text similarity as the similarity of different text descriptions for same video. The less the minimum of intra-text similarity, the diversity of the text descriptions for a video. As shown in Figure 3.3, at least one-third of the minimum of intra-text similarity is less than the mean of inter-text similarity on features from both models. It demonstrates that many videos have diverse captions and contents. So a single feature in some videos cannot be aligned with multiple text features at the same time. As a result, using multiple visual prototypes to align with different text is critical.

An intuitive idea for producing multiple visual prototypes is to generate a feature for each frame/clip of video as a prototype. However, one frame/clip can contain multiple characters and events at the same time. So a better way is to generate many potential prototypes for each frame, then we aggregate them together into multiple visual prototypes. Vision Transformer (2; 13) have shown it's effectiveness on learning multi-modality shared representation. So we take the final layer token features $z \in \mathbf{R}^{B \times D}$ of vision transformer as the potential prototypes. B, D represents the number of tokens, the dimension of token feature.

There are many methods to aggregate the token features, such as transformer decoder, graph neural network, LSTM etc. However, we find that the mask-based production of kth is enough. The kth prototype

$$p_k = \sum_{j=1}^{B} z^j \times f^{mask}(z^j)_k \tag{2}$$

, where z^j, f^{mask} represents the jth token feature of z, mask-generating function. f^{mask} is implemented by a linear layer, followed by relu function. In general, the class token feature is regarded as the final output of vision transformer (2; 13). So we expand $p_{k+1} = z^{cls}$, which means we have $K + 1$ visual prototypes. For comparing visual prototypes and text feature on same embedding space, we input p to a projection linear layer, do normalization and get multiple vision features $f^{ve} \in \mathbf{R}^{(K+1) \times D}$. Matching text to video becomes a matter of matching one text feature to multiple visual prototypes. The text is aligned with the video if one of the prototypes is very close to it. Based on this text-adaptive matching, we have new similarity function as

$$S_{t,v}^{\text{TMVM}} = \max_{k=1}^{K+1} \langle f^t, f_k^{ve} \rangle. \tag{3}$$

3.3 Variance Loss

Define mask $m_{i,k}^j = f^{mask}(z^j)_{i,k}$, which represents the mask weight on jth token from kth prototype of ith sample . It's an intuitive idea to different prototypes of same video to attend to different contents and the masks have different attention on different tokens. However, as shown in Figure 3.3,in some videos, the text descriptions are very similar. The videos may be too short and only have a character, event or action. The prototypes and masks are similar . Instead, we'd like all masks of all videos to pay attention to different token positions. Inspired by the variance loss on feature of (3), we design the variance loss on mask as

$$\mathcal{L}_{var} = \frac{1}{B} \sum_{j=1}^{B} \max(0, \gamma - D(m^j, \epsilon)) \tag{4}$$

,where D is the regularized standard deviation defined by

$$D(x, \epsilon) = \sqrt{\text{Var}(x) + \epsilon} \tag{5}$$

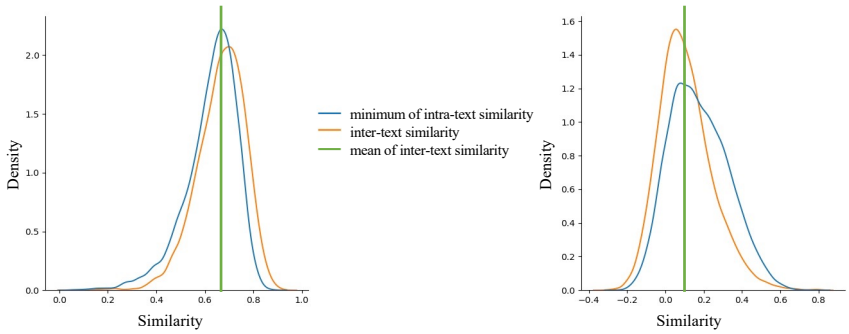

Figure 3: Inter-text similarity and minimum of intra-text similarity distribution on MSRVTT sample features. The features of the the left and right graph come from CLIP(22) and SentenceBertFormer(23) respectively. Intra-text similarity is defined as the similarity of different text descriptions for same video. Inter-text similarity is defined as the similarity of different text descriptions for different videos. As a result, the diversity of video contents can be represented by the minimum of intra-text similarity. It can be seen that at least one-third of the minimum of intra-text similarity is less than the mean of inter-text similarity on both graphs. It demonstrates that many videos have diverse captions and contents and it is necessary to use multiple visual prototypes to align with different text descriptions.

,γ is a constant target value for standard deviation, ϵ is a small scalar to prevent numerical instabilities. The variance loss forces the variance of all prototype mask values for each token to be large. It also encourages diverse prototypes for representing the rich information in videos.

3.4 Dual-Encoder Framework

We use a end-to-end training dual-encoder framework. Frozen (2), a basic and successful dual-encoder architecture (separate visual encoder and text encoder), serves as our foundation in this study. It's designed for video-language pretraining to match raw-pixel video and raw text with a contrastive loss in dual-encoder framework. The language and vision encoder is multi-layer bidirectional transformer and Timesformer (4) with some changes.

To train this dual-encoder framework, the matched text-video pairings in the batch are regarded as positive, while all other pairwise combinations in the batch are deemed negative. Each video appears at most once in each batch to prevent duplicates. With L text-video pairs in a batch, the symmetrical contrastive loss is calculated as follows

$$\mathcal{L}_{t,v} = \frac{1}{L}\sum_{i}^{L} -\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{L}\exp(s_{ij}/\tau)} - \log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{L}\exp(s_{ji}/\tau)} \tag{6}$$

where τ is the temperature and s_{ii}, s_{ij} is the similarity of ith paired text-video, ith text to jth video. Then the ranking of videos/texts is determined by this similarity calculation for single text/video query. So overall loss is defined by

$$\mathcal{L} = \mathcal{L}_{t,v} + \alpha \times \mathcal{L}_{var} \tag{7}$$

where α controls the weight of $\mathcal{L}_{var}$.

4 Experiment

We evaluate our model's performance on several public datasets. In the sections that follow, we first introduce the datasets and metrics for comparing results, as well as implementation details. Then we demonstrating the overall performance of our model on the datasets. Finally, we present ablation studies to evaluate the effectiveness of each component of our model.

Table 1: Retrieval performance on the MSR-VTT dataset

Method	Text-to-Video				Video-to-Text				SumR
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
ActBERT (34)	8.6	23.4	33.1	36.0	-	-	-	-	-
HowTo100M (19)	14.9	40.2	52.8	9.0	16.8	41.7	55.1	8.0	221.5
NoiseEstimation (1)	17.4	41.6	53.6	8.0	-	-	-	-	-
UviVL (17)	21.2	49.6	63.1	6.0	-	-	-	-	-
AVLnet (26)	27.1	55.6	66.6	4.0	28.5	54.6	65.2	4.0	297.5
HGR (6)	21.7	47.4	61.1	6.0	20.4	47.9	60.6	6.0	259.1
TT-CE+ (7)	29.6	61.6	74.2	3.0	-	-	-	-	-
Dual Encoding (8)	21.1	48.7	60.2	6.0	21.7	49.4	61.6	6.0	262.7
TACO (31)	24.5	52.8	65.5	5.0	-	-	-	-	-
SUPPORT_SET (21)	30.1	58.5	69.3	3.0	28.5	58.6	71.6	3.0	316.6
HiT (15)	30.7	60.9	73.2	**2.6**	32.1	62.7	74.1	3.0	333.7
CE (16)	20.9	48.8	62.4	6.0	20.6	50.3	64.0	5.3	267.0
MMT (10)	26.6	57.1	69.6	4.0	27.0	57.5	69.7	3.7	307.5
DMM (29)	31.2	62.8	**76.4**	3.0	29.4	63.4	**75.0**	**3.0**	338.2
ClipBERT (13)	22.0	46.8	59.9	6.0	-	-	-	-	-
Frozen (2)	31.0	59.5	70.5	3.0	-	-	-	-	-
Ours	**36.2**	**64.2**	75.7	3.0	**34.8**	**63.8**	73.7	**3.0**	**348.4**

Table 2: Retrieval performance on the LSMDC dataset

Method	Text-to-Video				Video-to-Text				SumR
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
JSFusion (32)	9.1	21.2	34.1	27.0	-	-	-	-	-
HiT (15)	14.0	31.2	41.6	18.5	-	-	-	-	-
TT-CE+ (7)	17.2	36.5	**46.3**	13.7	-	-	-	-	-
MEE (18)	9.3	25.1	33.4	25.3	-	-	-	-	-
MEE-COCO (18)	10.1	25.6	34.6	21.0	-	-	-	-	-
CE (16)	11.2	26.9	34.8	25.3	-	-	-	-	-
MMT (10)	13.2	29.2	38.8	21.0	12.1	29.3	37.9	22.5	160.5
DMM (29)	15.8	34.1	43.6	14.3	14.3	33.7	43.6	15.5	185.1
Frozen (2)	15.0	30.8	39.8	20.0	-	-	-	-	-
Ours	**17.8**	**37.1**	45.9	**13.5**	**16.5**	**34.3**	**44.6**	**14.0**	**196.2**

4.1 Datasets and Metrics

MSR-VTT

MSR-VTT(30) contains ten thousand YouTube videos, each of which is accompanied by 20 natural sentences that describe it. Following (2), we train on 9K train+val videos and report results on the 1K-A test set.

MSVD

MSVD (5) consists of 80K English descriptions for 1,970 videos from YouTube, with each video containing 40 sentences each. Following (2), we use the standard split of 1200, 100, and 670 videos for training, validation, and testing.

DIDEMO

DIDEMO (11) contains 10K Flickr videos annotated with 40K sentences. Following (2), we evaluate paragraph- to-video retrieval, where all sentence descriptions for a video are concatenated into a single query.

LSMDC

LSMDC (24) consists of 118,081 video clips sourced from 202 movies. The validation set contains 7,408 clips and evaluation is done on a test set of 1,000 videos from movies disjoint from the train and val sets. This follows the protocol outlined in (25).

Table 3: Retrieval performance on the DIDEMO dataset

Method	Text-to-Video			
	R@1	R@5	R@10	MedR
S2VT (28)	11.9	33.6	-	13.0
FSE (33)	13.9	36.0	-	11.0
TT-CE+ (7)	21.6	48.6	62.9	6.0
CE (16)	16.1	41.1	-	8.3
ClipBERT (13)	20.4	48.0	60.8	6.0
Frozen (2)	31.0	59.8	72.4	3.0
Ours	**36.5**	**64.9**	**75.4**	**3.0**

Table 4: Retrieval performance on the MSVD dataset

Method	Text-to-Video			
	R@1	R@5	R@10	MedR
VSE (12)	12.3	30.1	42.3	14.0
VSE++ (9)	15.4	39.6	53.0	9.0
TT-CE+ (7)	25.4	56.9	71.3	4.0
SUPPORT_SET (21)	28.4	60.0	72.9	4.0
MC (20)	20.3	47.8	61.1	6.0
CE (16)	19.8	49.0	63.8	6.0
Frozen (2)	33.7	64.7	76.3	**3.0**
Ours	**36.7**	**67.4**	**81.3**	**2.5**

Metrics

We evaluate retrieval performance using common information retrieval metrics such as Recall at K (R@K and K=1, 5, 10) and Median Rank (MedR). R@K is the proportion of test queries for which at least one relevant item is found among the top-K retrieved results. The MedR calculates the median rank of correct items in the returned ranking list, with a lower score indicating a better model. To reflect overall retrieval performance, we also take the sum of all R@K as SumR.

4.2 Implementation Details

We fintune the pretrained model from Frozen (2) on text-video retrieval dataset. The visual encoder have 12 attention blocks, patch size P = 16, sequence dimension D = 768, and 12 heads. The text encoder takes the architecture of DistilBERT base-uncased (27). The dimensionality of the common text-video space is set to 256. We take some training settings same with Frozen (2): we use 8 frames per video when training model; The temperature hyperparameter σ for the loss defined in Equation 6 is set to 0.05; We randomly crop and horizontally flip during training and center crop the maximal square crop at test time for visual augmentation; All videos are resized to 224$\times$224 as input; We use text augmentation during training for paragraph-retrieval settings, randomly sampling and concatenating a variable number of corresponding captions per video;The optimizer is Adam. We use a different learning rate scheduler: warmup for 5 epochs with a linearly growing learning rate from 0.0 to 0.00003, then use cosine decay scheduler to guide the learning rate from 0.00003 to 0.0 for other 45 epochs. The batch size is set as 64. The number of extra prototypes besides class token K is set as 3. The variance loss parameter γ, ϵ are set as 0.75, 1e-4. Variance loss weight α is set as 5.0.

4.3 Comparison to the State of the Art

The Table 1, 2, 3, 4 present the retrieval results of HiT on MSR-VTT, LSMDC, DIDEMO and MSVD. We also compare TMVM with other state-of-the-art methods: improving the robustness of pretrained model (34; 19; 1; 17; 26); designing better architecture or loss (6; 7; 8; 31; 21; 15; 28; 33; 32) on pre-extracted features; focusing on mining information of different visual modalities(10; 29; 18);end-to-end trainable model (13; 2).

As demonstrated by the results, TMVM outperforms all comparison methods clearly. We present video-to-text and text-to-video retrieval results for MSR-VTT and LSMDC. For MSR-VTT, our retrieval performance at SumR, in particular, is 348.4, outperforming recent state-of-the-art methods (29) by a margin of 10.2. For LSMDC, our retrieval performance at SumR also outperforms other methods by a clear margin. It accurately reflects TMVM's overall retrieval quality. We report retrieval performance in terms of text-to-video retrieval for DIDEMO and MSVD. TVMM continues to outperform comparison methods. Our method doesn't outperform R@10 of some other methods in Table 1, 2 . It is possibly because the methods use multi modality information of video, such as speech, audio and appearance. Our method is based on Frozen and only use the frames.

4.4 Ablation Study

We discover that the proposed components, such as Multiple Visual Prototypes and Variance Loss, help to improve retrieval performance. We exhaustively and comprehensively ablate two components to demonstrate their effectiveness and robustness.

Table 5: Ablation study on MSR-VTT to investigate the contributions of multiple visual prototypes matching and variance loss.

Method	Multiple Prototypes	Variance Loss	Text-to-Video			
	K		R@1	R@5	R@10	SumR
Baseline	0	✗	34.1	61.0	72.8	167.9
TVMM-Part	3	✗	34.7	61.7	72.8	169.2
TVMM-Decoder	3	✗	35.1	61.8	74.0	170.9
TVMM-Mask	3	✗	35.9	63.3	74.5	173.2
TVMM-Mask	3	✓	36.2	**64.2**	**75.7**	**176.1**
TVMM-Mask	1	✓	35.4	63.1	73.5	172.0
TVMM-Mask	2	✓	**36.4**	64.1	75.1	175.6
TVMM-Mask	4	✓	34.8	63.5	73.4	173.1
TVMM-Mask	10	✓	33.7	62.7	72.5	168.9
TVMM-Mask	1569=#tokens	✓	32.3	59.8	71.7	163.8

Table 6: Ablation study on MSR-VTT to investigate the effect of the number of input video frames.

Number of Video Frames	Method	Text-to-Video			
		R@1	R@5	R@10	SumR
1	Baseline	24.6	48.5	60.2	133.3
1	Ours	25.3	50.8	61.2	137.3
4	Baseline	33.6	59.1	71.0	163.7
4	Ours	34.6	61.9	73.9	170.4
8	Baseline	34.1	61.0	72.8	167.9
8	Ours	36.4	64.2	75.7	176.1

Production of Multiple Visual Prototypes. As mentioned above, we use mask-based method to aggregate token features from the final layer. In this section, we design two other variants to study whether this production is good enough.

- TVMM-Part. We divide the token features into several parts according to the frame order. The part's feature centers are then treated as prototypes.
- TVMM-Decoder. We use several embeddings as queries, and the token features as memory, then we throw query and memory into two-layer decoder to process the queries as visual prototypes.

TVMM-Mask is our method which is introduced in Section 2. As shown in Table 5 ,TVMM-Mask outperforms TVMM-Part and TVMM-Decoder with a clear margin. TVMM-Part improved little compared with the baseline, only 1.3 on SumR. It shows that TVMM-Part is very limited in its ability to model multiple prototypes.

The Number of Visual Prototypes. As shown in Table 5, multiple prototypes is better than single prototype. With $K = 3$ and its variance loss, TVMM-Mask outperforms the baseline with 5.3 on RSum. This reflects a video can contain rich instance and event information and the positive text only describes a part of it. TVMM find correspondence in the video with the help of multiple visual prototypes matching. The choice of number of prototypes is critical. K=3 is the best among 1,2,3,4 and 1569. When K is greater than or equal to 3, as K gets larger, the performance continues to degrade. A appropriate number of prototypes can assist the model in learning representative prototypes. When there are too many prototypes, the learned prototypes become noise rather than representative.

The Number of Input Visual Frames. In end-to-end training framework, the number of video frames is a parameter that has a big impact on performance. As shown in Table 6, as the number of video frames increases, so does the performance. It's worth noting that even on single-frame input, our method also improves the baseline with 4.0 on SumR. It proves the importance of multiple prototype matching.

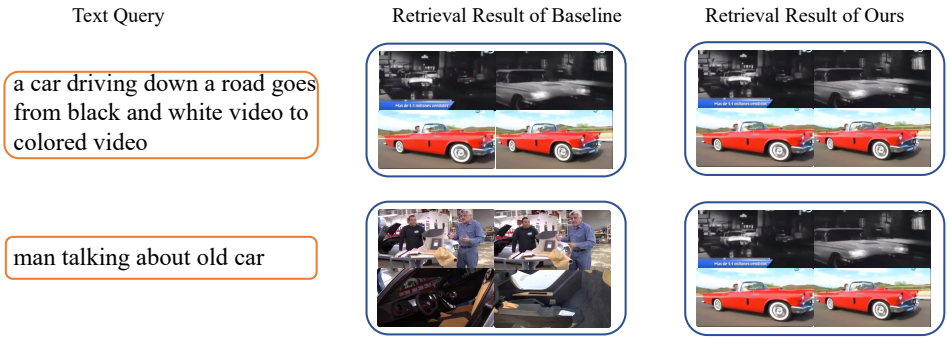

Figure 4: Some examples of text-video retrieval to compare the baseline with our method. The first video in the middle shows a guy talking about the transition from an old car to a new one. The second video in the middle shows a man describing various parts of a car to another person.

Figure 5: Visualization of multi prototypes. We select a video and its three corresponding captions from MSR-VTT. The captions are matched with the prototype that was the most similar. Then, to visualize the prototypes, we plot the heatmaps based on the mask values on the four sampled frames. The four frames depict a serious man, a sad woman, a serious woman, and a pair talking to each other.

Some Retrieval Examples. Each test video in the original MSR-VTT split has only one text description. In this experiment, we take some samples from the training data set and add them to the test set to demonstrate the superiority of the method. When compared to the baseline, our method provides correct retrieval results for the two test text queries, as shown in Figure 4.4. It's because our TVMM model can alleviate the ambiguity problem in video-text correspondence.

Visualization of multi prototypes. We choose a video and its three captions from MSR-VTT. The captions are matched with the most similar prototype. The heatmaps based on the mask values on the four sampled frames are then plotted to visualize the prototypes. The four frames show a serious man, a sad woman, a serious woman, and a pair conversing. As you see, the three captions correspond well with the prototypes, which attend to different contents of the video.

5 Conclusion

In this paper, we propose a Text-Adaptive Multiple Visual Prototype Matching Model. It generates multiple prototypes for video automatically to account for its rich information, and we present a text-adaptive matching strategy to find correspondence between texts and videos. Then, we propose a variance loss to encourage different prototypes to pay attention to different video contents. Experimental results on four public video retrieval datasets demonstrate the advantages of our model. Ablation studies are also carried out to verify the effectiveness of each part of our model.

References

[1] E. Amrani, R. Ben-Ari, D. Rotman, and A. M. Bronstein, "Noise estimation using density estimation for self-supervised multimodal learning," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 6644–6652. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16822

[2] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 1708–1718. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.00175

[3] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *CoRR*, vol. abs/2105.04906, 2021. [Online]. Available: https://arxiv.org/abs/2105.04906

[4] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 813–824. [Online]. Available: http://proceedings.mlr.press/v139/bertasius21a.html

[5] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. The Association for Computer Linguistics, 2011, pp. 190–200. [Online]. Available: https://aclanthology.org/P11-1020/

[6] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 10 635–10 644. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_Fine-Grained_Video-Text_Retrieval_With_Hierarchical_Graph_Reasoning_CVPR_2020_paper.html

[7] I. Croitoru, S. Bogolin, M. Leordeanu, H. Jin, A. Zisserman, S. Albanie, and Y. Liu, "Teachtext: Crossmodal generalized distillation for text-video retrieval," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 11 563–11 573. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01138

[8] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, and M. Wang, "Dual encoding for video retrieval by text," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[9] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: improving visual-semantic embeddings with hard negatives," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 12. [Online]. Available: http://bmvc2018.org/contents/papers/0344.pdf

[10] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12349. Springer, 2020, pp. 214–229. [Online]. Available: https://doi.org/10.1007/978-3-030-58548-8_13

[11] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. C. Russell, "Localizing moments in video with natural language," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 5804–5813. [Online]. Available: https://doi.org/10.1109/ICCV.2017.618

[12] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *CoRR*, vol. abs/1411.2539, 2014. [Online]. Available: http://arxiv.org/abs/1411.2539

[13] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less is more: Clipbert for video-and-language learning via sparse sampling," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 7331–7341. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Lei_Less_Is_More_ClipBERT_for_Video-and-Language_Learning_via_Sparse_Sampling_CVPR_2021_paper.html

[14] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," *CoRR*, vol. abs/2201.12086, 2022. [Online]. Available: https://arxiv.org/abs/2201.12086

[15] S. Liu, H. Fan, S. Qian, Y. Chen, W. Ding, and Z. Wang, "Hit: Hierarchical transformer with momentum contrast for video-text retrieval," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 11 895–11 905. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01170

[16] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 2019, p. 279. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/0363-paper.pdf

[17] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 13–23. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html

[18] A. Miech, I. Laptev, and J. Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," *CoRR*, vol. abs/1804.02516, 2018. [Online]. Available: http://arxiv.org/abs/1804.02516

[19] A. Miech, D. Zhukov, J. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 2630–2640. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00272

[20] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR 2018, Yokohama, Japan, June 11-14, 2018*, K. Aizawa, M. S. Lew, and S. Satoh, Eds. ACM, 2018, pp. 19–27. [Online]. Available: https://doi.org/10.1145/3206025.3206064

[21] M. Patrick, P. Huang, Y. M. Asano, F. Metze, A. G. Hauptmann, J. F. Henriques, and A. Vedaldi, "Support-set bottlenecks for video-text representation learning," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=EqoXe2zmhrh

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[23] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[24] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3202–3212. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298940

[25] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. J. Pal, H. Larochelle, A. C. Courville, and B. Schiele, "Movie description," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 94–120, 2017. [Online]. Available: https://doi.org/10.1007/s11263-016-0987-1

[26] A. Rouditchenko, A. W. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. S. Feris, B. Kingsbury, M. Picheny, A. Torralba, and J. R. Glass, "Avlnet: Learning audio-visual language representations from instructional videos," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 1584–1588. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-1312

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019. [Online]. Available: http://arxiv.org/abs/1910.01108

[28] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, R. Mihalcea, J. Y. Chai, and A. Sarkar, Eds. The Association for Computational Linguistics, 2015, pp. 1494–1504. [Online]. Available: https://doi.org/10.3115/v1/n15-1173

[29] W. Wang, M. Zhang, R. Chen, G. Cai, P. Zhou, P. Peng, X. Guo, J. Wu, and X. Sun, "Dig into multi-modal cues for video retrieval with hierarchical alignment," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Z. Zhou, Ed. ijcai.org, 2021, pp. 1113–1121. [Online]. Available: https://doi.org/10.24963/ijcai.2021/154

[30] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 5288–5296. [Online]. Available: https://doi.org/10.1109/CVPR.2016.571

[31] J. Yang, Y. Bisk, and J. Gao, "Taco: Token-aware cascade contrastive learning for video-text alignment," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 11 542–11 552. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01136

[32] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11211. Springer, 2018, pp. 487–503. [Online]. Available: https://doi.org/10.1007/978-3-030-01234-2_29

[33] B. Zhang, H. Hu, and F. Sha, "Cross-modal and hierarchical modeling of video and text," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11217. Springer, 2018, pp. 385–401. [Online]. Available: https://doi.org/10.1007/978-3-030-01261-8_23

[34] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 8743–8752. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Zhu_ActBERT_Learning_Global-Local_Video-Text_Representations_CVPR_2020_paper.html

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No]
   (b) Did you include complete g xproofs of all theoretical results? [No]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The requirement method is public.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The data is publicly available.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] We didn't use crowdsourcing and conducted research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] We didn't use crowdsourcing and conducted research with human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] We didn't use crowdsourcing and conducted research with human subjects.