This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Narrating the Video: Boosting Text-Video Retrieval via Comprehensive Utilization of Frame-Level Captions

Chan Hur* Jeong-hun Hong* Dong-hun Lee Dabin Kang Semin Myeong
Sang-hyo Park† Hyeyoung Park†
School of Computer Science and Engineering, Kyungpook National University
https://multimodal-understanding-group.github.io/NarVid/

Abstract

In recent text-video retrieval, the use of additional captions from vision-language models has shown promising effects on the performance. However, existing models using additional captions often have struggled to capture the rich semantics, including temporal changes, inherent in the video. In addition, incorrect information caused by generative models can lead to inaccurate retrieval. To address these issues, we propose a new framework, Narrating the Video (NarVid), which strategically leverages the comprehensive information available from frame-level captions, the narration. The proposed NarVid exploits narration in multiple ways: 1) feature enhancement through cross-modal interactions between narration and video, 2) query-aware adaptive filtering to suppress irrelevant or incorrect information, 3) dual-modal matching score by adding query-video similarity and query-narration similarity, and 4) hard-negative loss to learn discriminative features from multiple perspectives using the two similarities from different views. Experimental results demonstrate that NarVid achieves state-of-the-art performance on various benchmark datasets.

1. Introduction

With the increasing popularity of various video streaming platforms, text-to-video retrieval is emerging as an important topic in both academia and industry. In order to achieve accurate retrieval results, text-video retrieval encounters two challenging points: 1) a cross-modality gap from the heterogeneous data (vision and language) and 2) time-varying information in video. To address cross-modality challenges, many studies [9, 11, 20, 23, 25, 26, 35, 37, 39, 43, 45] have leveraged vision-language pre-trained models such as CLIP [30] or BLIP [18] as a text and video backbone to resolve cross-modality gap between text-to-video domain. Moreover, to capture the time-varying relationships for video data, several studies [9, 23, 25, 43] have employed temporal modules, such as transformer encoder [34] or 3D convolution [33], over video frames.

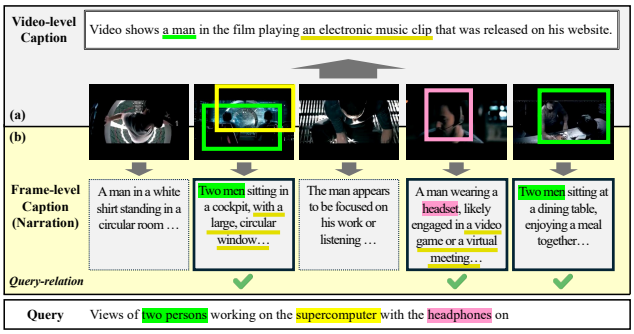


Figure 1. (a) Unlike video-level caption information that summarizes the entire video, (b) the proposed framework utilizes frame-level captions to capture time-varying rich information and efficiently leverage selective information based on specific queries.

Recently, some research [27, 40] has proposed to utilize generated captions from the video to bridge the gap between video and language. Cap4Video [40] derives video-level captions that summarize the entire video and then utilize the captions to mitigate the modality gap. However, as shown in Fig. 1(a), a video-level caption may not reflect the contextual variations across the entire video. Furthermore, the usage of a single caption can lead to performance degradation when the generated caption is incorrectly obtained. EA-VTR [27] have proposed an event-aware method to capture time-varying events in the video by introducing frame-level captions, but they focus on training video encoder for event detection to address event-related video tasks, and does not consider query-based selective usage of the captions which is essential in the retrieval task.

To overcome these challenges in caption-based approaches, we propose a novel framework named Narrating the Video (NarVid), which generates frame-level captions across the video and utilizes them comprehensively

*These authors contributed equally to this work.
†Corresponding authors, equal leading contribution.

throughout all stages of text-video retrieval. These frame-level captions obtained over time can be viewed as a story thus we call them a narration. As shown in Fig. 1(b), instead of the existing video-level caption, frame-level captions allow us to capture diverse objects and attributes throughout the entire video. Meanwhile, since not all captions in the narration are relevant to the query, appropriate handling of the generated information is required.

In our NarVid framework, we utilize the narration in a variety of ways to improve retrieval performance. First, we extract mutually enhanced features through co-attention between the pairs of video frames and the generated captions. The enhanced frame-level features are further refined by filtering out irrelevant frames and incorrectly generated captions based on similarities to the given query. The narration features are also used in the matching process, where query-video similarity and query-narration similarity are fused to provide better retrieval results. The query-video-narration relationship is also applied to define a novel hard negative loss for learning more discriminative features. The proposed comprehensive utilization of narration throughout the retrieval process successfully reduces the cross-modality gap, leading to state-of-the-art results in text-video retrieval.

To summarize, our contributions are as follows:

1. We propose a novel text-video retrieval framework, NarVid, which strategically leverages frame-level captions obtained from the generative model. These frame-level captions, called narration, could provide time-varying rich information to boost the retrieval performance.
2. Our NarVid framework comprehensively utilizes narration in the entire matching process through four modules: 1) enhancing cross-modal features, 2) filtering irrelevant or incorrect features, 3) complementing matching scores, and 4) learning with hard-negative sampling.
3. We validate the performance of NarVid through extensive experiments on four benchmark datasets. The proposed model outperforms state-of-the-art models in retrieval performances on MSR-VTT (52.7%), MSVD (53.1%), VATEX (68.4%), and DiDeMo (53.4%).

2. Related works

2.1. Text-Video Retrieval

Vision-language pre-trained models such as CLIP [30] have demonstrated the feasibility of learning a multimodal embedding space through contrastive learning on large-scale data, exhibiting strong generalization performance. Building on this success, recent research has focused on extending CLIP to video domain tasks, such as text-video retrieval. CLIP4Clip [25] successfully adapted CLIP for text-video retrieval by extracting features from individual video frames and aggregating them. Following CLIP4Clip, X-CLIP [26] enhanced the features of each modality by integrating both coarse-grained and fine-grained features. Additionally, X-Pool [11] utilized an attention mechanism to highlight frames that hold the most relevance to the text. Recently, probabilistic approaches have emerged to capture variability in the data. UAVTR [9] introduced a technique to represent the pairs of text-video features as probability distributions. T-MASS [35] presents an approach by modeling text as a stochastic embedding, enriching the semantic representation. Although these CLIP-based retrieval models have shown significant performance improvements, they have limitations in bridging the substantial domain gap between video and text. Our paper mitigates this limitation by employing auxiliary information from the generative model and utilizing the generated information in diverse ways.

2.2. Vision-Language Model with Generated Information

Recent work on vision-language models has shown that captions generated using large language models can improve feature representation. In the image domain, ALIP [42] introduced triplet similarities among images, text, and generated captions to improve text-image retrieval performance. LaCLIP [8] used rewritten textual descriptions to augment CLIP learning for obtaining enhanced multi-modal features. In the video domain, recent studies have enhanced cross-modal interactions by utilizing additional data, such as captions derived from videos. CoVR [32] improved the composed video retrieval performance by using cross-attention between videos and generated captions, and Cap4Video [40] proposed a co-attention framework using single-sentence captions summarizing entire videos. Nonetheless, these studies primarily focus on global video content, often overlooking local details. In this context, EA-VTR [27] uses frame-level captions in defining the loss function of the encoder by comparing every possible pair of frames and captions from all videos, which may cause undesirable effects by unrefined information. Our approach strategically utilizes frame-level captions, including query-aware filtering methods, to emphasize the temporal aspects of the videos and capture local details, reducing the domain gap between video and text.

3. Proposed Methods

3.1. Overall Process

Problem formulation. Given a query text, the text-to-video retrieval task is to identify the most semantically relevant video from a set of candidate videos V. The task is formally defined as the optimization problem of finding the optimal video $\boldsymbol{v}^*$ that maximizes the similarity with the given query $\boldsymbol{q}$, which can be written as

$$\boldsymbol{v}^* = \arg\max_{\boldsymbol{v} \in V} s(\boldsymbol{q}, \boldsymbol{v}). \tag{1}$$

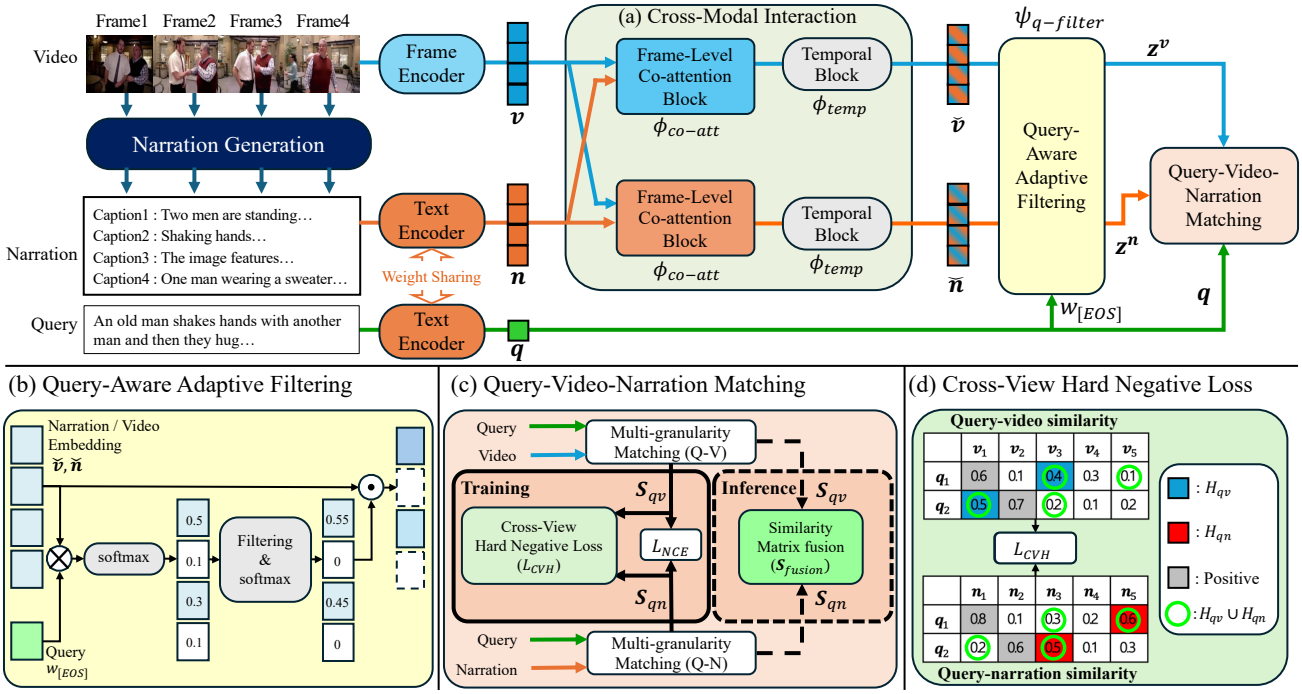


Figure 2. An overview of the proposed framework, NarVid. The method first generates frame-level captions (narration) for each video. (a) Using the frame-level features of the video and narration, enhanced features are obtained through cross-modal interaction with co-attention and temporal block. (b) These enhanced features are further refined using query-aware adaptive filtering. (c) Then, the query-video and query-narration similarity matrices obtained through the multi-granularity matching are utilized for training and inference. (d) To enhance the discriminative ability of the model, we additionally use a cross-view hard negative loss during training.

Here, $s(\boldsymbol{q}, \boldsymbol{v})$ is the similarity function that needs to be modeled appropriately to achieve good retrieval performance.

Pipeline. The proposed framework for calculating similarity follows the process outlined in Fig. 2. A query and a video are represented as $\boldsymbol{q} = (w_1, \ldots, w_L, w_{[EOS]})$ and $\boldsymbol{v} = (f_1, \ldots, f_K)$ using clip encoders, where w_L is the last word of the query, $w_{[EOS]}$ denotes a special token representing the query's global meaning, and f_K is the last frame of the video. Additionally, we generate a sequence of captions for individual frames using the narration generation module, represented as $\boldsymbol{n} = (c_1, \ldots, c_K)$ (Sec. 3.2). The encoded features of the narration and video are enhanced through mutual co-attention and temporal transformation (Sec. 3.3). Query-aware adaptive filtering refines these features based on the query (Sec. 3.4), then query-video and query-narration matching is employed to obtain the final similarity score (Sec. 3.5). A cross-view hard negative loss is also defined by combining query-video and query-narration similarities to enhance retrieval performance (Sec. 3.6.1).

3.2. Narration Generation from Video

In this paper, we try to solve some limitations of the previous work [40] that uses a video-level caption to describe a video. Since a video-level caption is usually generated by focusing on the key scenes, it is likely to miss other narrative flows in the video clip. In addition, when a single generated caption is used to represent the entire video frames, any errors (*e.g.*, hallucinations) in the generation model can severely affect the retrieval performance. To address these issues, we propose the concept of narration. For each image frame f_k in a video $\boldsymbol{v} = (f_1, \ldots, f_K)$, we apply a large multimodal model (LMM) [22] to independently interpret it and generate a caption c_k, so that we have a sequence of captions, $\boldsymbol{n} = (c_1, \ldots, c_K)$. Note that our framework does not focus on any specific caption generation model, the choice of captioning model and prompts can be flexible, considering various environments. The generated captions arranged in chronological order can linguistically express the semantic changes in the video over time. The generated narration is subsequently used in various ways in our proposed NarVid to improve the performance of text-to-video retrieval.

3.3. Video-Narration Cross-Modal Interaction

Each caption c_k in the generated narration can be paired with a video frame f_k to interact and enhance mutual information. Specifically, we use the co-attention structure [24] for interactions, a method commonly employed for cross-modal information enrichment. Fig. 2 (a) shows the cross-modal interaction between each video $\boldsymbol{v}$ and its corresponding narration $\boldsymbol{n}$. In this context, the transformer blocks per-

form co-attention operations with cross-modal inputs to obtain the enhanced features $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{n}}$, defined as

$$(\hat{\boldsymbol{v}}, \hat{\boldsymbol{n}}) = \phi_{co-att}(\boldsymbol{v}, \boldsymbol{n}). \tag{2}$$

Through co-attention between video frames and their corresponding captions instead of a video-level caption, we can naturally capture intricate relationships between specific frames and their narrative context. This approach enhances features by leveraging rich information from all frames through individual frame-caption pairs.

The obtained sequential features $\hat{\boldsymbol{v}} = (\hat{f}_1, \ldots, \hat{f}_K)$ and $\hat{\boldsymbol{n}} = (\hat{c}_1, \ldots, \hat{c}_K)$ are then passed through a temporal block [25] to capture temporal relationships in the sequence of each modality, producing temporally enhanced features as follows:

$$\check{\boldsymbol{v}} = \phi_{temp}(\hat{\boldsymbol{v}}), \quad \check{\boldsymbol{n}} = \phi_{temp}(\hat{\boldsymbol{n}}). \tag{3}$$

Since the enhanced features $\check{\boldsymbol{v}}$ and $\check{\boldsymbol{n}}$ are still in sequential forms with frame-level features, we can apply a filtering process to select query-relevant features.

3.4. Query-Aware Adaptive Filtering

To filter out query-irrelevant frames or erroneous captions in the narration, we propose a query-aware filtering module. Conventional video frame filtering methods employ a top-k approach [15, 39], selecting the top-k frames with the highest similarity to the query. However, using a fixed for filtering may cause a loss of important frames or the inclusion of irrelevant ones. To overcome this limitation, we adopt the nucleus sampling method [13] to implement nucleus filtering, allowing the adaptive determination of the number of keyframes.

Fig. 2 (b) illustrates our nucleus filtering process. For a given query $\boldsymbol{q} = (w_1, \ldots, w_L, w_{[EOS]})$ and the enhanced features $\check{\boldsymbol{v}} = (\check{f}_1, \ldots, \check{f}_K)$ and $\check{\boldsymbol{n}} = (\check{c}_1, \ldots, \check{c}_K)$, we first compute the frame-level similarities between $w_{[EOS]}$ and the features $\check{f}_k$ and $\check{c}_k$ $(k = 1, \ldots, K)$. We then apply the softmax function to normalize these similarity values to the range $[0, 1]$. In the feature selection phase, we start with the feature having the highest similarity and add features one by one until the cumulative similarity exceeds a predefined threshold p. Through this process, we obtain a pair of final sequences of features, $\boldsymbol{z}^v$ for video and $\boldsymbol{z}^n$ for narration, defined as

$$(\boldsymbol{z}^v, \boldsymbol{z}^n) = \psi_{q-filter}(\check{\boldsymbol{v}}, \check{\boldsymbol{n}}; \boldsymbol{q}). \tag{4}$$

This adaptive filtering process effectively selects key features for video and narration that are highly relevant to the specific query.

3.5. Query-Video-Narration Matching

The filtered feature vectors are used in a detailed multi-granularity matching process using coarse and fine granularities [12, 38]. In addition to traditional inter-modality matching between query and video, we propose an additional intra-modality matching between query and narration to improve retrieval performance.

Coarse-grained matching is a common method for matching overall semantics between query and video. In addition to query-video matching, a similar query-narration matching process is conducted in NarVid. Specifically, we first apply weighted pooling to the sequence of features for the video $\boldsymbol{z}^v$ and narration $\boldsymbol{z}^n$ to obtain their coarse representations, where the weights are determined by the similarity scores from query-aware filtering module (Sec. 3.4). The $w_{[EOS]}$ token is used for the coarse representation of the query $\boldsymbol{q}$. Then, the coarse-grained matching scores $s_{coarse}(\boldsymbol{q}, \boldsymbol{z}^v)$ and $s_{coarse}(\boldsymbol{q}, \boldsymbol{z}^n)$ are calculated using cosine similarity between their coarse representations, effectively leveraging the filtering scores as weights in the pooling process.

Fine-grained matching is a detailed matching strategy for capturing fine-level relationships between query and video. We extend the conventional method to include query-narration matching. Specifically, the query-video matching score $s_{fine}(\boldsymbol{q}, \boldsymbol{z}^v)$ is defined as the maximum cosine similarity between frame and word embeddings of the query, following the approach in Cap4Video [40]. In NarVid, we also define the query-narration matching score $s_{fine}(\boldsymbol{q}, \boldsymbol{z}^n)$ as the maximum similarity between the embeddings of frame-level captions and words.

By averaging the coarse and fine scores, we obtain two multi-grained similarity scores, s_{qv} for query-video matching and s_{qn} for query-narration matching, which are used for training and inference. For more detailed information is qiven in supplementary Sec. A

3.6. Training and Inference

3.6.1. Training Objective

Contrastive loss. For a given batch set $\{(\boldsymbol{q}_i, \boldsymbol{v}_i, \boldsymbol{n}_i)\}_{i=1}^{B}$, we calculate the query-video similarity matrix $\boldsymbol{S}_{qv}$ and the query-narration similarity matrix $\boldsymbol{S}_{qn}$, where each element is defined as $\boldsymbol{S}_{qv}(i, j) = s_{qv}(\boldsymbol{q}_i, \boldsymbol{v}_j)$, $\boldsymbol{S}_{qn}(i, j) = s_{qn}(\boldsymbol{q}_i, \boldsymbol{n}_j)$. We utilize these similarity matrices to enhance retrieval performance by considering both inter-modality (query-video) and intra-modality (query-narration) aspects. This approach modifies the InfoNCE loss [29] commonly used in text-video retrieval to magnify the discrepancy between positive and negative samples from both perspec-

tives. The modified loss functions are defined as

$$L_{qv} = -\frac{1}{2B}\sum_{i=1}^{B}\left\{\log\frac{e^{\boldsymbol{S}_{qv}(i,i)}}{\sum_{j=1}^{B}e^{\boldsymbol{S}_{qv}(i,j)}} + \log\frac{e^{\boldsymbol{S}_{qv}(i,i)}}{\sum_{j=1}^{B}e^{\boldsymbol{S}_{qv}(j,i)}}\right\},$$

$$L_{qn} = -\frac{1}{2B}\sum_{i=1}^{B}\left\{\log\frac{e^{\boldsymbol{S}_{qn}(i,i)}}{\sum_{j=1}^{B}e^{\boldsymbol{S}_{qn}(i,j)}} + \log\frac{e^{\boldsymbol{S}_{qn}(i,i)}}{\sum_{j=1}^{B}e^{\boldsymbol{S}_{qn}(j,i)}}\right\},$$

$$L_{NCE} = \frac{1}{2}(L_{qv} + L_{qn}). \tag{5}$$

Cross-view hard negative loss. Hard negative samples in contrastive learning mean negative samples that closely resemble positive samples, challenging the model to distinguish between them. Although several studies [4, 31] showed that more discriminative features can be obtained if such hard negative samples are well utilized for training, they exclusively focused on either inter-modality or intra-modality when selecting hard negatives. In this paper, we propose a novel hard negative loss that leverages both different views: inter-modality and intra-modality. As shown in Fig. 2 (d) using the similarity matrices $\boldsymbol{S}_{qv}$ and $\boldsymbol{S}_{qn}$, we define two index sets of hard negatives for a given query $\boldsymbol{q}_i$ from the views of inter-modality and intra-modality. these sets are defined as

$$\begin{aligned} H_{qv}(i) &= \{j \mid \boldsymbol{S}_{qv}(i,i) - \boldsymbol{S}_{qv}(i,j) < \lambda\sigma_i^{qv}, j \neq i\}, \\ H_{qn}(i) &= \{j \mid \boldsymbol{S}_{qn}(i,i) - \boldsymbol{S}_{qn}(i,j) < \lambda\sigma_i^{qn}, j \neq i\}. \end{aligned} \tag{6}$$

Here, σ_i^{qv} and σ_i^{qn} are the standard deviation of the values in i-th row of $\boldsymbol{S}_{qv}$ and $\boldsymbol{S}_{qn}$, and λ is a hyperparameter. The unified hard negative index for i-th query, H_i, is obtained by combining both sets:

$$H_i = H_{qv}(i) \cup H_{qn}(i). \tag{7}$$

We can apply the same process for the transposed matrices, $\boldsymbol{S}_{qv}^T$ and $\boldsymbol{S}_{qn}^T$ to obtain hard negatives for i-th video $\boldsymbol{v}_i$ and narration $\boldsymbol{n}_i$, denoting these as H_i^T. Subsequently, we implement a hard negative rank loss, built on the concept of hinge loss [6], using the index sets of unified hard negatives. Specifically, using the threshold value $\lambda\sigma_i^{qv}$ scaled with η, the hard negative loss for query-video pairs is defined as

$$L_{qv}^{hard} = \frac{1}{2B}\sum_{i=1}^{B}\left(\sum_{j\in H_i}\Delta_{qv}(i,j) + \sum_{j\in H_i^T}\Delta_{vq}(i,j)\right),$$

where

$$\begin{aligned} \Delta_{qv}(i,j) &= \max(0, \boldsymbol{S}_{qv}(i,j) - \boldsymbol{S}_{qv}(i,i) + \eta\lambda\sigma_i^{qv}), \\ \Delta_{vq}(i,j) &= \max(0, \boldsymbol{S}_{qv}(j,i) - \boldsymbol{S}_{qv}(i,i) + \eta\lambda\sigma_i^{vq}). \end{aligned} \tag{8}$$

Similarly, We can define the hard negative loss L_{qn}^{hard} for query-narration pairs using the similarity matrices $\boldsymbol{S}_{qn}$. The overall cross-view hard negative loss is then defined as

$$L_{CVH} = L_{qv}^{hard} + L_{qn}^{hard}, \tag{9}$$

and the final training loss is defined as

$$L = L_{NCE} + \alpha L_{CVH}, \tag{10}$$

where α is a hyperparameter.

3.6.2. Inference Pipeline

Since the elements in the two similarity matrices $\boldsymbol{S}_{qv}$ and $\boldsymbol{S}_{qn}$ often exist in different ranges of values, the direct summation of the two matrices can disproportionately emphasize the narration aspect. To mitigate this problem, the element values in each matrix are separately standardized before summation, using their means μ^{qv} and μ^{qn} and the standard deviations σ^{qv} and σ^{qn}. For inference, the two standardized matrices are then fused to form the final score matrix $\boldsymbol{S}_{fusion}$, ensuring balanced consideration of both aspects, which can be defined as

$$\boldsymbol{S}_{fusion}(i,j) = \frac{\boldsymbol{S}_{qv}(i,j) - \mu^{qv}}{\sigma^{qv}} + \frac{\boldsymbol{S}_{qn}(i,j) - \mu^{qn}}{\sigma^{qn}}. \tag{11}$$

4. Experimental Results

4.1. Experimental Settings

Datasets and metrics. We validate our framework on four benchmark datasets using the same protocol as in Cap4Video [40]. MSR-VTT [41] consists of 10K videos, each containing approximately 20 captions, and we utilize the 1k-A test split [44] for evaluation. MSVD [3] includes 1,970 videos, with each video accompanied by an average of 40 captions. VATEX [36] comprises 35K video clips, each featuring multilingual captions. DiDeMo [1] consists of 10K videos annotated with around 40K captions, where we evaluate video-paragraph retrieval by concatenating all sentence descriptions for each video into a single query.

For evaluation, we employ standard retrieval metrics: Recall at Rank K (R@K), Median Rank (MdR), and Mean Rank (MnR). Higher values for R@K indicate better performance, while lower values for both MdR and MnR are preferred.

Implementation details. We use the baseline settings of CLIP4Clip [25] with ViT-B/16 [7] as the visual backbone and the CLIP text encoder for query and narration encoding. For a detailed analysis of MSR-VTT, we additionally use ViT-B/32 as a visual encoder. To obtain the

Method	Venue	Text-to-Video					Video-to-Text				
		R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
CLIP-ViT-B/32											
CLIP4Clip [25]	Neurocomputing'22	44.5	71.4	81.6	2.0	15.3	42.7	70.9	80.6	2.0	11.6
X-CLIP [26]	MM'22	46.1	73.0	83.1	2.0	13.2	46.8	73.3	84.0	2.0	9.1
X-Pool [11]	CVPR'22	46.9	72.8	82.2	2.0	14.3	44.4	73.3	84.0	2.0	9.0
UATVR [9]	ICCV'23	47.5	73.9	83.5	2.0	12.3	46.9	73.8	83.8	2.0	8.6
TEFAL [16]	ICCV'23	49.4	75.9	83.9	2.0	-	-	-	-	-	-
PAU [17]	NeurIPS'23	48.5	72.7	82.5	2.0	14.0	-	-	-	-	-
Cap4Video [40]	CVPR'23	49.3	74.3	83.8	2.0	12.0	47.1	73.7	84.3	2.0	8.7
MPT [45]	MM'24	46.3	70.9	80.7	-	15.6	45.0	70.9	80.6	-	12.7
ProTA [10]	ICME'24	48.1	75.4	84.3	2.0	12.5	45.9	75.5	84.6	2.0	9.0
T-MASS [35]	CVPR'24	50.2	75.3	85.1	**1.0**	11.9	47.7	**78.0**	**86.3**	2.0	8.0
NarVid		**51.0**	**76.4**	**85.2**	**1.0**	**11.6**	**50.0**	75.4	83.8	**1.5**	**7.9**
CLIP-ViT-B/16											
X-CLIP [26]	MM'22	49.3	75.8	84.8	2.0	12.2	48.9	76.8	84.5	2.0	8.1
UATVR [9]	ICCV'23	50.8	76.3	85.5	**1.0**	12.4	48.1	76.3	85.4	2.0	8.0
TEFAL [16]	ICCV'23	49.9	76.2	84.4	2.0	11.4	-	-	-	-	-
Cap4Video [40]	CVPR'23	51.4	75.7	83.9	**1.0**	12.4	49.0	75.2	85.0	2.0	8.0
MPT [45]	MM'24	49.2	72.9	82.4	-	15.5	47.4	73.9	83.4	-	10.9
ProTA [10]	ICME'24	50.9	77.0	85.4	**1.0**	11.1	48.5	77.0	87.0	2.0	7.9
T-MASS [35]	CVPR'24	**52.7**	77.1	**85.6**	**1.0**	**10.5**	50.9	**80.2**	**88.0**	**1.0**	**7.4**
NarVid		**52.7**	**77.7**	**85.6**	**1.0**	12.3	**51.1**	76.8	85.2	**1.0**	9.0

Table 1. Retrieval results comparisons on MSR-VTT. Bold denotes the best performance. Results are reported without any post-processing such as QB-Norm [2] or DSL [5].

Method	Venue	MSVD			VATEX			DiDeMo		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP4Clip [25]	Neurocomputing'22	45.2	75.5	84.3	55.9	89.2	95.0	42.8	68.5	79.2
UATVR [9]	ICCV'23	49.7	79.0	87.3	64.5	92.6	96.8	45.8	73.7	83.3
TEFAL [16]	ICCV'23	-	-	-	61.0	90.4	95.3	-	-	-
PAU [17]	NeurIPS'23	47.3	77.4	85.5	-	-	-	48.6	76.0	84.5
Cap4Video [40]	CVPR'23	51.8	80.8	88.3	66.6	93.1	97.0	52.0	79.4	87.5
MPT [45]	MM'24	-	-	-	-	-	-	46.4	72.2	81.4
ProTA [10]	ICME'24	-	-	-	-	-	-	47.2	74.6	83.0
T-MASS [35]	CVPR'24	-	-	-	65.6	93.9	**97.2**	53.3	**80.1**	**87.7**
NarVid		**53.1**	**81.4**	**88.8**	**68.4**	**94.0**	97.1	**53.4**	79.1	86.3

Table 2. Text-to-video retrieval results on MSVD, VATEX, and DiDeMo. Bold denotes the best performance. Results are reported without any post-processing such as QB-Norm [2] or DSL [5].

narration, LLaVa 1.5 7B [22] is used as the baseline generative model. We use a batch size of 64 and train for 5 epochs on three datasets, except for DiDeMo, where a batch size of 24 and 10 learning epochs are used. The threshold p used in query-aware filtering is set to 0.4 for ViT-B/32 and 0.5 for ViT-B/16. All experiments are conducted on two NVIDIA RTX A6000 GPUs. Additional implementation details can be found in Supplementary Sec. B.2

4.2. Comparison with State-of-the-Arts

We compare NarVid with state-of-the-art models across the four benchmark datasets.

MSR-VTT dataset. Tab. 1 shows the superior retrieval performance of NarVid compared to other methods across the overall recall metrics. Compared with the baseline CLIP4Clip, without using any generation information, NarVid achieves a dramatic improvement across all retrieval metrics. The result shows the impressive effect of the utilization of narration. Furthermore, NarVid also outperforms Cap4Video, which uses video-level captions, on both ViT-B/32 and ViT-B/16 backbones across most evaluation metrics, including video-to-text retrieval. These results highlight the benefits of the proposed comprehensive utilization of frame-level captions. We also compare NarVid to

Proposed modules				Text-to-Video		
NM	CMI	QAF	CVH	R@1	R@5	R@10
				44.5	71.4	81.6
✓				48.6	76.2	84.6
✓	✓			49.5	76.3	84.9
✓	✓	✓		50.4	76.1	84.6
✓	✓	✓	✓	**51.0**	**76.4**	**85.2**

Table 3. Ablation study results on MSR-VTT. NM, CMI, QAF, and CVH denote the modules for narration matching, cross-modal interaction, query-aware filtering, and cross-view hard negative loss.

the recent state-of-the-art model T-MASS [35], which uses stochastic text embedding as a training strategy. Although NarVid outperforms T-MASS on R@1 across all datasets, it shows decreased performances on R@5 and R@10. We conclude that these performance gaps can be attributed to the stochastic embedding methods proposed by T-MASS, which augment the number of matched text-video pairs. Such stochastic embedding methods in the matching phase could be partially integrated into NarVid as an additional training strategy to improve performance in future work.

Tab. 2 presents the comparative results for the MSVD, VATEX, and DiDeMo datasets.

MSVD dataset. We demonstrate clear performance improvements across all recall metrics. However, when compared to other datasets, the performance improvement in R@1 over the baseline model (CLIP4Clip) is relatively modest (MSVD +7.9%, VATEX +12.5%, and DiDeMo +10.2%). It may be attributed to the simplicity and brevity of MSVD queries. We also observed that some queries could be matched to multiple videos in a test set, which may limit the effectiveness of narration information. This tendency is observed in the results of the MSR-VTT dataset with similar forms of queries.

VATEX dataset. NarVid exhibits a significant improvement in R@1 compared to the baseline. Due to the properties of the VATEX dataset, the query sentences are relatively long and contain words describing various attributes. Such sentences can be well matched with the diverse information included in the narration, which can lead to a substantial improvement in retrieval performance.

DiDeMo dataset. NarVid demonstrates superior performance in R@1 compared to other models; however, the performance gap is smaller than observed in other datasets. This reduced margin is likely due to the static nature of DiDeMo videos, which diminishes the effectiveness of the rich temporal information that narration provides.

In summary, our method achieves state-of-the-art results across all four benchmarks, demonstrating the effectiveness of the narration-based approach in text-video retrieval.

Method	Hyperparameter	R@1	R@5	R@10
None	-	48.1	74.7	82.7
Top-k	k = 3	50.3	76.4	84.9
	k = 4	50.2	76.4	85.0
Nucleus filtering	p = 0.3 (2.55)	50.3	**76.6**	84.5
	p = 0.4 (3.57)	**51.0**	76.4	**85.2**

Table 4. Change of accuracies depending on the filtering methods for MSR-VTT. The values in parentheses are the average numbers of frames selected during inference.

α	R@1	R@5	R@10	MnR	MdR
0	50.4	76.1	84.6	1.0	**11.3**
0.5	50.8	76.4	84.8	1.0	11.4
1.0	**51.0**	**76.4**	**85.2**	**1.0**	11.6
1.5	50.1	76.3	85.2	1.0	11.4

Table 5. Performance changes depending on the hyperparameter of hard negative loss on MSR-VTT.

4.3. Ablation Study

In this section, we conduct an ablation study on MSR-VTT to thoroughly evaluate the effectiveness of the four modules proposed for narration utilization.

Narration matching. We first validate the impact of incorporating query-narration matching into the model. As shown in Tab. 3, the R@1 increased by 4.1%, demonstrating that the rich information in the narration serves as complementary data during the matching process. However, it is slightly lower than that of Cap4Video, indicating that the simple incorporation of frame-level captions is insufficient to fully leverage its potential. This also supports the need for comprehensive utilization through additional modules.

Video-narration cross-modal interaction. Beyond the complementary information provided by narration during the matching stage, we analyze the impact of feature enhancement through co-attention and temporal blocks. As shown in Tab. 3, the use of enhanced features improves retrieval performance across all recall metrics. It indicates the importance of cross-modal interaction and the temporal information in video and narration features.

Query-aware adaptive filtering. Although the enhanced features contain rich information, not all features are relevant to the given query, and some erroneous information may be included. Therefore, proper filtering is necessary to ensure relevance and accuracy. The results in Tab. 3 demonstrate the importance of filtering, showing improved R@1 as irrelevant or inaccurate information is effectively removed. Furthermore, Tab. 4 highlights that unfiltered information can lead to a decline in performance, emphasizing the necessity of filtering methods, as significant performance gains are observed regardless of the filtering method employed. However, the optimal number of frames may vary depending on the samples. Notably, adaptive nucleus

filtering with $p = 0.4$ achieves higher R@1 accuracy while maintaining a similar average number of selected frames (3.57), compared to fixed selection with k = 3 or k = 4 used in previous studies. This supports the conclusion that adaptive nucleus filtering is more efficient than the fixed top-k selection approach.

Cross-view hard negative loss. The narration is also used to define a hard negative loss, including intra-modal similarity. The bottom row of Tab. 3 shows an improvement across all recall metrics, indicating that the proposed hard negative loss helps the model learn more discriminative features. Tab. 5 shows the dependence on the hyperparameter α, which controls the proportion of the hard negative loss term. The best performance is achieved with $\alpha = 1$, implying the necessity of a substantial proportion of hard negatives for optimal performance.

Impact of different narration generation models. We analyze how different caption generation methods affect retrieval performance on NarVid. Specifically, as shown in Fig. 3, we compared narrations from the output of several types of captioners: a traditional image-captioning model [14], popular vision-language pre-trained models [18, 19], and powerful large multimodal models [21, 22]. The results show that the use of narration improves performance on all metrics regardless of the captioner, demonstrating that the strategy we propose works for a variety of caption generation models. Interestingly, we observed that more powerful VLMs generate detailed captions containing many query-irrelevant words. These well-generated but query-misaligned captions could negatively impact retrieval performance. Detailed experimental results about the caption generation can be found in Supplementary Sec. C

4.4. Qualitative Results

We also compare the actual retrieval result (R@1) from our NarVid with the Cap4Video [40] using a video-level caption. The video-level caption summarizing the entire video struggles to reflect the whole content of the video, especially in videos with frequent scene changes. In contrast, the frame-level captions not only better capture the whole video contents but also help focus on the segments most relevant to the query. As shown in Fig. 4, the video-level caption prioritizes the theme of outfits, leading to incorrect retrieval results linked to fashion shows. Conversely, our narration provides rich information, including outfits, diverse styles, and smiling expressions. Utilizing this information properly, NarVid could obtain more accurate results.

5. Conclusions

We propose the NarVid framework which focuses on the comprehensive utilization of frame-level generated captions, named narration, for the text-to-video retrieval task. The narration naturally provides rich semantic information

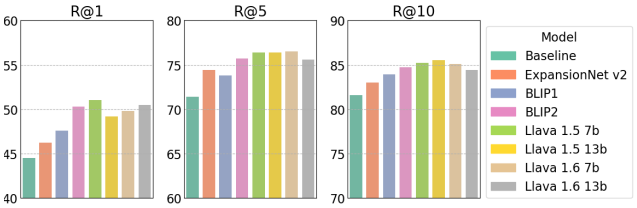


Figure 3. Effectiveness of various captioners on MSR-VTT dataset. Note that except for the changes in the captioners, all other architecture is the same.

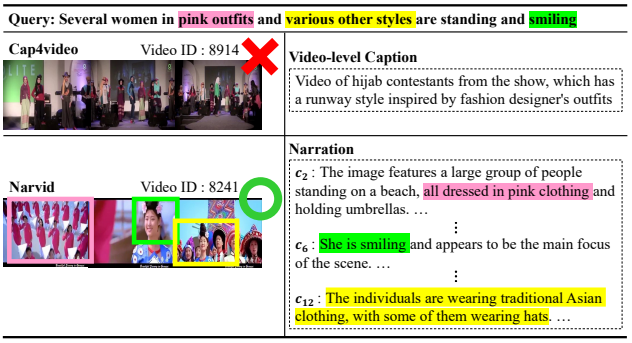


Figure 4. Text-to-video retrieval results (R@1) on MSR-VTT test set. For the Cap4Video, the video-level caption does not match the words in the query. For NarVid, narration contains richer information over frames, and it leads to the correct retrieval results.

arranged in chronological order. We utilize narrations in strategic ways, such as feature enhancement through cross-modal interaction, filtering out inappropriate information, and defining a novel similarity and loss function for matching and training. Extensive experiments on benchmark datasets validate the superiority of NarVid and reveal its potential effectiveness for retrieval tasks.

Limitations. To improve real-world applicability, a pre-computation and storage process for narration generation is necessary for the NarVid framework. Efficient narration generation techniques could be employed to reduce the pre-computation burden. In addition, although we demonstrated that our framework works over different caption generators in Fig. 3, the improvement in retrieval performance is closely tied to the capability of the caption generator. A more thorough investigation into the caption generation process, incorporating prompt engineering and advanced generation techniques, could further enhance the retrieval performance.

Acknowledgments

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2021-II212068, Artificial Intelligence Innovation Hub) in part by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research

Center) support program(IITP-2024-RS-2024-00437718) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), and in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00227431, Development of 3D space digital media standard technology).

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 5

[2] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205, 2022. 6

[3] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 5

[4] Yuxin Chen, Zongyang Ma, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Ying Shan, Xiaojuan Qi, and Weiming Hu. How to make cross encoder a good teacher for efficient image-text retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26994–27003, 2024. 5

[5] 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. 6

[6] Corinna Cortes. Support-vector networks. *Machine Learning*, 1995. 5

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[8] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *Advances in Neural Information Processing Systems*, pages 35544–35575, 2023. 2

[9] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13723–13733, 2023. 1, 2, 6

[10] Han Fang, Xianghao Zang, Chao Ban, Zerun Feng, Lanxiang Zhou, Zhongjiang He, Yongxiang Li, and Hao Sun. Prota: Probabilistic token aggregation for text-video retrieval. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE Computer Society, 2024. 6

[11] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015, 2022. 1, 2, 6

[12] Peiyan Guan, Renjing Pei, Bin Shao, Jianzhuang Liu, Weimian Li, Jiaxi Gu, Hang Xu, Songcen Xu, Youliang Yan, and Edmund Y Lam. Pidro: Parallel isomeric attention with dynamic routing for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11164–11173, 2023. 4

[13] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. 4

[14] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. Exploiting multiple sequence lengths in fast end to end training for image captioning. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2173–2182. IEEE Computer Society, 2023. 8

[15] Zhiming Hu, Angela Ning Ye, Salar Hosseini Khorasgani, and Iqbal Mohomed. Adaclip: Towards pragmatic multimodal video retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5623–5633, 2023. 4

[16] Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12054–12064, 2023. 6

[17] Hao Li, Jingkuan Song, Lianli Gao, Xiaosu Zhu, and Hengtao Shen. Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. In *Advances in Neural Information Processing Systems*, pages 24564–24585, 2023. 6

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 8

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 8

[20] Qian Li, Yucheng Zhou, Cheng Ji, Feihong Lu, Jianian Gong, Shangguang Wang, and Jianxin Li. Multi-modal inductive framework for text-video retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2389–2398, 2024. 1

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023. 8

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916, 2023. 3, 6, 8

[23] Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, and Thomas H. Li. Revisiting temporal modeling for

clip-based image-to-video knowledge transferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6555–6564, 2023. 1

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 2019. 3

[25] 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 and captioning. *Neurocomputing*, 508:293–304, 2022. 1, 2, 4, 5, 6

[26] 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 *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 1, 2, 6

[27] Zongyang Ma, Ziqi Zhang, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Yingmin Luo, Xu Li, Xiaojuan Qi, Ying Shan, et al. Ea-vtr: Event-aware video-text retrieval. In *European Conference on Computer Vision*, pages 76–94, 2024. 1, 2

[28] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[30] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[31] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021. 5

[32] Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. Composed video retrieval via enriched context and discriminative embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26896–26906, 2024. 2

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 1

[35] Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16551–16560, 2024. 1, 2, 6, 7

[36] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591, 2019. 5

[37] Xiao Wang, Yaoyu Li, Tian Gan, Zheng Zhang, Jingjing Lv, and Liqiang Nie. Rtq: Rethinking video-language understanding based on image-text model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 557–566, 2023. 1

[38] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2816–2827, 2023. 4

[39] Mengxia Wu, Min Cao, Yang Bai, Ziyin Zeng, Chen Chen, Liqiang Nie, and Min Zhang. An empirical study of frame selection for text-to-video retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6821–6832, 2023. 1, 4

[40] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10704–10713, 2023. 1, 2, 3, 4, 5, 6, 8

[41] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5

[42] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931, 2023. 2

[43] Huanjin Yao, Wenhao Wu, and Zhiheng Li. Side4video: Spatial-temporal side network for memory-efficient image-to-video transfer learning. *arXiv preprint arXiv:2311.15769*, 2023. 1

[44] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European conference on computer vision*, pages 471–487, 2018. 5

[45] Haonan Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Mpt: Multi-grained prompt tuning for text-video retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1206–1214, 2024. 1, 6