Video-ColBERT: Contextualized Late Interaction for Text-to-Video Retrieval

Arun Reddy[1,2]* Alexander Martin[2]* Eugene Yang[2,3] Andrew Yates[2,3] Kate Sanders[2]
Kenton Murray[2,3] Reno Kriz[2,3] Celso M. de Melo[4] Benjamin Van Durme[2,3] Rama Chellappa[2]
[1]Johns Hopkins Applied Physics Laboratory [2]Johns Hopkins University
[3]Human Language Technology Center of Excellence [4]DEVCOM Army Research Laboratory
{areddy24,amart233}@jhu.edu

Abstract

In this work, we tackle the problem of text-to-video retrieval (T2VR). Inspired by the success of late interaction techniques in text-document, text-image, and text-video retrieval, our approach, Video-ColBERT, introduces a simple and efficient mechanism for fine-grained similarity assessment between queries and videos. Video-ColBERT is built upon three main components: a fine-grained spatial and temporal token-wise interaction, query and visual expansions, and a dual sigmoid loss during training. We find that this interaction and training paradigm leads to strong individual, yet compatible, representations for encoding video content. These representations lead to increases in performance on common text-to-video retrieval benchmarks compared to other bi-encoder methods.

1. Introduction

With an ever-increasing amount of video data being generated and stored daily, the need for effective and efficient retrieval methods has become more pressing than ever. Text-to-video retrieval (T2VR) aims to address this by ranking large collections of videos based on their relevance to natural language queries. However, the task remains challenging due to the inherent modality gap between text and video representations. While recent advances in cross-modal retrieval have started to bridge this gap [6, 31, 43, 52, 53], significant progress is still needed to achieve reliable and scalable performance in real-world settings.

A common approach to efficient retrieval is the use of a bi-encoder method [25, 41, 42], where the query and document are encoded separately. Bi-encoders offer efficiency advantages over cross encoders [19, 29, 38, 45] because they do not require expensive interactions at retrieval time and can instead operate on a pre-computed index of the target data. Some bi-encoder T2VR techniques [11, 34] use single vectors to represent the text query and the video (*e.g.* through mean pooling), then use a single dot product for similarity calculation at retrieval time. While this simple interaction may be sufficient for settings like language-image pre-training [41, 61, 62], it can be challenging to encode video concent and query concepts in a single feature vector [24, 46]. Other works have applied more expressive interactions, like token-wise interaction, to the video retrieval problem [35, 49] using CLIP [41] adapted for videos. We argue that these works are limited in their exploration of token-wise interaction. Specifically, these approaches employ overly complicated interaction mechanisms or have limited final representation(s) used in their interactions.

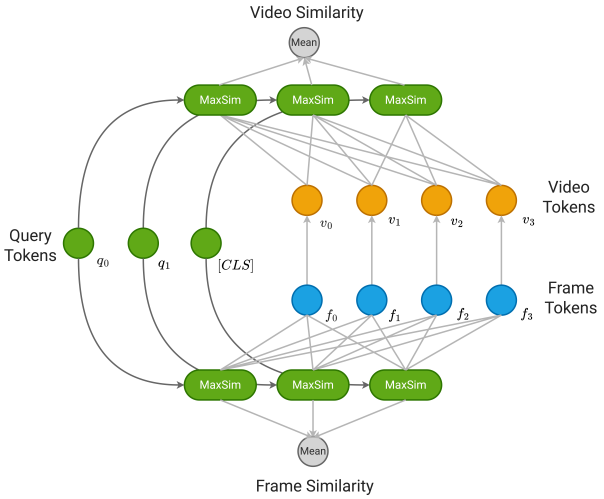

Figure 1. VIDEO-COLBERT architecture, combining token-wise interaction on both static frame features (blue) and temporally contextualized video features (orange).

ColBERT [26] is a text-to-text retrieval model that uses a multi-vector bi-encoder retrieval method. This approach achieves an effective middle ground between expensive cross-encoders and single-vector bi-encoders by enabling late interaction between individual query and document tokens, which maintains the efficiency of simple dot product-

*Equal contribution.

based interactions but captures more of the relevant context. Specifically, ColBERT's *MaxSim* operator employs a summation over maximum similarity operations, which allows different aspects of the text query to individually detect relevant content in documents. Additionally, ColBERT introduced unique interactions between query padding tokens and document tokens for *soft query augmentation.*

In this work, we introduce VIDEO-COLBERT, a bi-encoder approach to T2VR that leverages token-wise interaction at both the spatial and spatio-temporal levels. VIDEO-COLBERT incorporates a modification to the *MaxSim* operation, *MeanMaxSim* (MMS), which replaces the summation with a mean to better accommodate variable length queries and to control the magnitude of the overall similarity. On top of this modified interaction, VIDEO-COLBERT uses two MMS operations over both independent visual frame features and contextualized frame features to strengthen the fine-grained spatial and temporal interaction (Fig. 1). This dual MMS interaction is trained with a specialized sigmoid-based loss objective to strengthen the independence and compatibility of the spatial and spatio-temporal representations. We find that the combination of the interactions and the loss function enhances the robustness of each representation resulting in a more resilient fusion during evaluation. Our contributions and method can be summarized as follows

1. We introduce a novel T2VR method, VIDEO-COLBERT, a multi-level late-interaction retrieval model that provides better retrieval effectiveness with comparable model size.
2. We introduce a stronger contextualized token-wise interaction by performing MMS at the spatial and spatio-temporal levels.
3. We introduce a sigmoid-based loss for effectively training our version of token-wise interaction with both spatial and spatio-temporal visual features.
4. We present a series of experiments that analyze key aspects of our method for interactions, training objectives and query augmentations.

2. Related Works

Text-to-Video Retrieval. Early works in text-to-video retrieval made use of pre-trained experts [17, 32, 51] to represent videos. Later attempts to perform end-to-end video-language pre-training [57, 65] (on datasets like HowTo100M [36]) saw limited success due to lack of scale and the poor quality of paired text. Frozen in Time [1] showed that both image-text and video-text pairs could be used to train an enhanced dual encoder model for video retrieval.

Since the advent of the groundbreaking image-text contrastive model CLIP [41], several works have sought to adapt it to the video retrieval problem [2, 14, 18, 19, 28, 33–35, 37, 49, 59, 64**?**]. Many of these methods use CLIP to extract frame- or even patch-level representations, with some employing additional transformer layers for temporal modeling. Other work has looked to create stronger temporal representations of the video. Liu et al. [33] adopt token shift operations and selection modules to further improve video encoding and increase interaction amongst meaningful frames. Deng et al. [11] continue to build upon this approach using a "prompt cube" to force interaction between all pairs of frames in the video which are then aggregated into a final representation.

Tokenwise Late Interaction. In the domain of text document retrieval, ColBERT [26] showed the promise of contextualized late interaction techniques, which enable fine-grained interaction between queries and documents while maintaining efficiency at retrieval time. FILIP [60] applies a similar idea to image-text matching, where individual query token features interact with image patch features. Recently, ColPali [15] applied this to visual document retrieval with vision-language models for retrieving PDF files with textual queries. Likewise, several works also explored token-wise interaction for video retrieval [13, 20, 23, 35, 49, 54, 56]. Among them, X-CLIP [35] uses multi-grained interactions on both the query (word and sentence) and document (frame and video) sides. Another variant, DRL [49], uses weighted token-wise interaction to take into account the importance of query words and video frames. UATVR [13] adds extra learnable tokens as input to the query and video encoders to enrich the token-wise interaction. Unlike VIDEO-COLBERT, none of the aforementioned techniques perform interaction on both spatial and spatio-temporal visual features.

3. Preliminaries

Problem Formulation & Notation. Text-to-video retrieval is the task of ranking a collection of videos based on relevance to a given natural language text query. Formally, given a sequence of text query tokens $Q = \{q_i, \ldots, q_M\}$ and a set of videos $\{V_1, \ldots, V_C\}$, the objective of a T2VR method is to produce a video ranking that aligns with true relevance judgments. In this work, we focus on the T2VR setting where only visual information (*i.e.* video frames) can be used to assess query-video relevance.

We denote each video as a temporally ordered sequence of sampled frames: $V = \{f_1, \ldots f_N\}$, where each f_i has an independent spatial representation $\mathbf{f}_i$ (extracted by an image encoder) and temporally contextualized representations $\mathbf{v}_i$ (produced by a temporal transformer operating on $\{\mathbf{f}_1, \ldots, \mathbf{f}_N\}$).

Interaction Mechanisms. When considering how query tokens should interact with video frames to compute a query-video relevance score, several options exist. Among the simplest of these are single-vector techniques, like CLIP4Clip [34], which perform a pooling operation over frame features to arrive at a single unified video representation. Similarly, the query can also be represented using a single vector $\mathbf{q}$, typically using a special aggregation token in a text transformer model. Then, the similarity score can be computed via a dot product (equivalent to cosine similarity, assuming L2 normalization) of the query and video vectors. Formally, the similarity computation using mean pooling (MP) is defined as:

$$MP(Q,V) = \mathbf{q} \cdot \frac{1}{N}\sum_{i=1}^{N} \mathbf{f}_i \tag{1}$$

While simple and efficient, the MP approach assumes that single vectors are sufficient to adequately represent both the query and video content. Such an assumption may not hold as the complexity of text queries and videos increases. An alternative approach is to employ a more fine-grained interaction between query and video by computing cosine similarity at the individual token level. For example, a ColBERT-style [26] summation over *MaxSim* operations (SMS) could be applied to video retrieval as:

$$SMS(Q,V) = \sum_{j=1}^{M} \max_i(\mathbf{q}_j \cdot \mathbf{f}_i) \tag{2}$$

The SMS formulation for similarity calculation allows each query token representation to effectively "scan" the video frames for relevant content, then contribute to the summation the maximum similarity found among all frames. We employ a similar fine-grained interaction approach in VIDEO-COLBERT.

4. Method

We now describe VIDEO-COLBERT, a fine-grained approach for adapting image-text dual encoder models (like CLIP [41] and SigLIP [62]) for T2VR. VIDEO-COLBERT (depicted in Fig. 2) has 3 main aspects: (i) fine-grained spatial and temporal interaction, performing MMS on both independent frames and their contextualized representations, (ii) query and visual expansion tokens which allow for additional information to be encoded for abstract queries and for additional high-level temporal information from the video, and (iii) a dual sigmoid loss for training strong independent, yet compatible, spatial and spatio-temporal representations.

4.1. Fine-Grained Spatial & Temporal Interaction

The first component of our query-video interaction mechanism is a modified form of SMS (Eq. (2)), which replaces the summation with a mean to better accommodate variable length queries. This interaction, which we denote as MMS_F, operates on static frame features extracted by an image encoder (specifically, using the [CLS] token of a vision transformer [12]):

$$MMS_F(Q,V) = \frac{1}{M}\sum_{j=1}^{M} \max_i(\mathbf{q}_j \cdot \mathbf{f}_i) \tag{3}$$

where $\mathbf{q}_j$ denotes the output of the j-th query token from the query encoder (*i.e.* a contextualized query token feature). Because a set of individual image features is unable to capture relationships across time, we also perform temporal modeling by processing the frame [CLS] tokens with additional transformer layers. The output of these temporal transformer layers is a sequence of temporally contextualized "video" features $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$. Our method performs token-wise query interaction with these features as well:

$$MMS_V(Q,V) = \frac{1}{M}\sum_{j=1}^{M} \max_i(\mathbf{q}_j \cdot \mathbf{v}_i) \tag{4}$$

In contrast to [49, 60], we intentionally calculate both MMS_F and MMS_V in only one direction. In other words, only query token features are used to select relevant visual features and not the other way around. We argue that there exists an inherent asymmetry between queries and videos, and that the overall similarity score should not be diminished by the presence of video frames that do not correspond with any query tokens. The final query-video similarity score in VIDEO-COLBERT is a sum of the frame-level and video-level MMS scores:

$$MMS_{FV}(Q,V) = MMS_F(Q,V) + MMS_V(Q,V) \tag{5}$$

Such summation can also be considered as the Borda score of the set of frame scores [10].

The aim of MMS_{FV} is to better capture the interaction between purely spatial information and spatio-temporal video features from the query by incorporating two levels of interaction. The MMS_F operation locates relevant static information, while the MMS_V operation matches dynamic concepts. Unlike previous works that only use features after temporal modeling, the contextualized video representations in VIDEO-COLBERT can encode more temporal information because the temporal layers have less need to preserve purely spatial concepts and can instead focus on capturing higher-level cross-frame and global interactions.

4.2. Query & Visual Expansion

In addition to our interaction mechanism, we again take inspiration from ColBERT [26] in our use of *soft query augmentation*. ColBERT finds that including extra padding token features in the token-wise interaction enhances retrieval

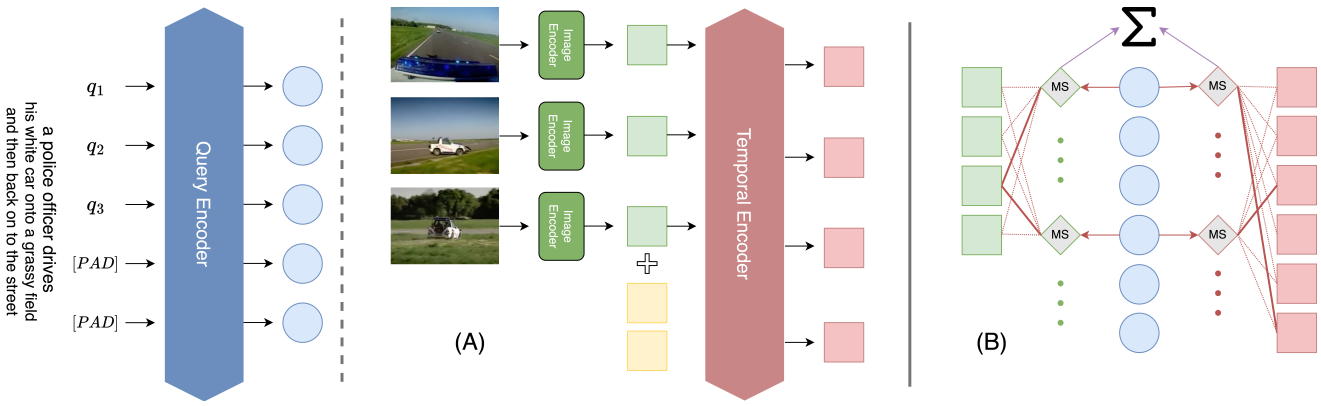

Figure 2. Overview of VIDEO-COLBERT. (A) Shows the VIDEO-COLBERT bi-encoder process. The query is encoded by a text encoder ( ). The images are encoded independently with an image-encoder ( ) and produce [CLS] tokes for frames ( ). The frames and additional visual expansion tokens ( ) are passed through the temporal transformer( ). (B) Shows the dual MMS_{FV} interaction, where encoded query tokens ( ) perform *MaxSim* operations ( ) with the frames ( ) and video features ( ). A summation is then performed over each *MaxSim* operation for the frames and video features respectively and the two summations are combined for the final relevance score.

performance. The authors hypothesize that augmenting the query with these additional tokens enables a learnable form of query expansion [16, 26] , whereby additional "search terms" implied by the base query can be computed to enhance retrieval. Following this intuition, we incorporate pad tokens into both the MMS_F and MMS_V interactions.

The intuition behind query expansion tokens also extends to the video side. Following Fang et al. [13], we incorporate visual expansion tokens, alongside the frame [CLS] tokens, as input to the temporal transformer. During training, we encourage $\mathbf{v}_i$ to deviate from the corresponding $\mathbf{f}_i$. This allows the video features to have strong contextualized representations, while the expansion tokens capture high-level global video features which span across multiple frames. With the composite MMS_{FV} similarity score, we can then ensure that our interaction incorporates both spatial information as well as stronger temporal information.

4.3. Dual Sigmoid Loss

Prior methods that train bi-encoder models for T2VR primarily utilize a bi-directional, softmax-based InfoNCE loss [39], defined as follows:

$$-\frac{1}{2|B|}\sum_{i=1}^{|B|}\left(\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|B|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}+\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|B|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}\right)$$

In the context of text-video retrieval, $\mathbf{x}$ and $\mathbf{y}$ would be text and video representations. The overall loss incorporates both text-to-video and video-to-text InfoNCE losses by performing softmax normalization across both dimensions of the batch similarity matrix. However, InfoNCE is sensitive to the negative example selection [7, 25, 44]. Such examples may be hard to obtain without the help of an effective matching or selection model to start with, which is the kind of model that we aim to train in the first place. Furthermore, recent advancements in image-text contrastive learning have demonstrated the advantages of sigmoid-based losses over their softmax counterparts [62]. The sigmoid loss turns the original contrastive objective into a series of independent binary classification tasks, thereby eliminating the need for computing global normalization factors. The sigmoid loss has also been shown to be more robust to noisy data [62], which is prevalent in T2VR datasets in both the quality of annotations [5] and the ambiguity of abstract text queries and descriptions [63].

For these reasons, we adopt the sigmoid loss when training VIDEO-COLBERT. Specifically, the loss is defined as:

$$-\frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{j=1}^{|B|}\log\frac{1}{1+e^{z_{ij}(-t\cdot MMS(Q_i,V_j)+b)}} \quad (6)$$

where z_{ij} is a label indicating positive $(+1)$ and negative (-1) pairings, t denotes a learnable logit scaling factor and b is a learnable logit bias.

Building upon the sigmoid loss and the fine-grained spatial and spatio-temporal interactions in MMS_{FV}, we propose a dual loss function that fuses MMS_F and MMS_V at the ranking level. Since the information in MMS_F and MMS_V propagates from different levels, their magnitudes will naturally differ. Given this, it is preferable to compute separate sigmoid losses on the MMS_F and MMS_V similarity matrices and avoid the multi-loss scaling issue [8, 30]. Furthermore, separate losses encourage stronger independent representations from each level (frame and video). Thus, we employ a dual loss formulation (Eq. (7)) that computes the global loss as the linear combination of the losses of the MMS_F and MMS_V interactions.

$$\mathcal{L}_F = \frac{1}{1 + e^{z_{ij}(-t(MMS_F)+b)}}$$

$$\mathcal{L}_V = \frac{1}{1 + e^{z_{ij}(-t(MMS_V)+b)}}$$

$$\mathcal{L}_D = \lambda_F \mathcal{L}_F + \lambda_V \mathcal{L}_V \tag{7}$$

λ_F and λ_V act as additional hyperparameters, allowing more importance to be placed on spatial or temporal features.

5. Experiments

5.1. Datasets

We evaluate VIDEO-COLBERT on several T2VR datasets, in which video captions serve as proxies for user queries. We use only English captions for all datasets.

- *MSR-VTT* [58] contains 10,000 total videos, each paired with 20 captions. We use the `Training-9K` and `1K-A` splits for training and testing respectively.
- *MSVD* [4] contains 1,200 training videos and 670 test videos, each paired with roughly 40 captions.
- *VATEX* [50] contains 25,991 training videos and 1,500 test videos, each with 10 captions.
- *DiDeMo* [22] contains 8,391 training videos and 1,004 test videos. Each video is paired with approximately four temporally localized captions, which we concatenate to create a paragraph-video retrieval task.
- *ActivityNet* [21] contains 10,009 training videos and 4,917 test videos. We again concatenate the temporally localized captions to create a paragraph-video retrieval task for this dataset.

5.2. Implementation Details

Network Architecture & Training. Both the query and video encoders in VIDEO-COLBERT are initialized from CLIP ViT-B/16 or CLIP ViT-B/32. We additionally introduce a variant of the ViT-B/16 initialized from SigLIP [62] which was trained on the WebLI dataset using a sigmoid loss. For temporal modeling, we use 4 transformer layers based on the text encoder of the underlying dual encoder model. We set the number of visual expansion tokens to 2, and set both λ_F and λ_V to 1 during training. Our models are fine-tuned using the Adam [27] optimizer, with learning rate of 1×10^{-7} for pre-trained image and text encoder parameters and 1×10^{-4} for temporal transformer parameters. We freeze the positional encodings, patch embeddings and token embeddings during fine-tuning. All other transformer parameters are trained. More details about our training setting and dataset specific configurations can be found in the Appendix.

Text Pre-Processing. We adopt the same tokenizer and special token mappings used in the original CLIP and SigLIP models. For CLIP [41], we prepend the token sequence with a `<|startoftext|>` token and add an `<|endoftext|>` token at the end. For query augmentation using CLIP, we enable self-attention and MMS interaction with additional pad tokens (ID #0, which corresponds to an exclamation point `!`) used to fix the token sequence length to 32 for MSR-VTT, MSVD and VATEX, and 64 for DiDeMo and ActivityNet. Because SigLIP uses a "last token" aggregation strategy and always performs self-attention across a length-64 padded token sequence, we use 64 text tokens when using SigLIP as a backbone.

Video Pre-Processing. We perform TSN-style [48] uniform sampling to select frames from videos. Each frame is resized to 224×224 without maintaining aspect ratio in order to avoid information loss resulting from center cropping. In line with previous work, we sample 12 frames for MSR-VTT, MSVD and VATEX, while using 64 frames for DiDeMo and ActivityNet.

5.3. Baselines & Evaluation Metrics

Our comparisons to other work focus on alternative bi-encoder approaches for text-video retrieval. We compare against alternative interaction mechanisms and context aggregation strategies rather than orthogonal approaches like captioning [56], large-scale video-text pre-training [43], and expensive attention interactions that use cross-modal transformers or multimodal large language models [3, 6, 19]. In our comparisons, we include the best bi-encoder approaches for text-video retrieval [18, 33–35, 49, 64] built upon CLIP-B/32, CLIP-B/16 and SigLIP-B/16. For a fairer comparison with our SigLIP encoder variant, we also upgrade the backbone in CLIP4CLIP [34] to create SigLIP4Clip. We implement the mean pooling (meanP) and sequence transformer (seqTransf) variants of the CLIP4Clip method and perform fine-tuning using an InfoNCE loss.

For evaluation, we report recall at 1 (R@1), 5 (R@5) and 10 (R@10). As advocated for by [55], we also include normalized discounted cumulative gain (nDCG), a commonly used metric in text-based information retrieval. We use nDCG@10 and abbreviate it as nDCG in our tables. For all metrics, higher number indicate better performance.

5.4. Benchmark Results

In Tab. 1, we show results on three sentence-to-video retrieval datasets. We see that VIDEO-COLBERT achieves competitive or state-of-the-art results in several settings. When using CLIP-B/32 as a backbone, we find that VIDEO-COLBERT outperforms other approaches that utilize token-wise interaction (even those that use more granular patch-level information [33]). We find that DRL [49], when us-

Method	MSR-VTT				MSVD				VATEX			
	R@1	R@5	R@10	nDCG	R@1	R@5	R@10	nDCG	R@1	R@5	R@10	nDCG
ClipBERT [29]	22.0	46.8	59.9	–	–	–	–	–	–	–	–	–
Support Set [40]	30.1	58.5	69.3	–	28.4	60.0	72.9	–	45.9	82.4	90.4	–
Frozen [1]	32.5	61.5	71.2	–	33.7	64.7	76.3	–	–	–	–	–
ViT-B/32												
CLIP4Clip-meanP [34]	43.1	70.4	80.8	–	46.2	76.1	84.6	–	–	–	–	–
CLIP4Clip-seqTransf [34]	44.5	71.4	81.6	–	45.2	75.5	84.3	–	–	–	–	–
CenterCLIP [64]	44.2	71.6	82.1	–	47.6	77.5	86.0	–	–	–	–	–
CLIP2TV [18]	46.1	72.5	82.9	–	47.0	76.5	85.1	–	–	–	–	–
TS2-Net [33]	47.0	74.5	83.8	–	44.6	75.8	–	–	59.1	90.0	95.2	–
X-CLIP [35]	46.1	73.0	–	–	47.1	77.8	–	–	–	–	–	–
DRL [49]	47.4	74.6	83.8	–	**48.3**	**79.1**	**87.3**	–	**63.5**	**91.7**	**96.5**	–
VIDEO-COLBERT (CLIP-B/32)	**48.1**	**74.9**	**83.9**	**0.652**	46.0	75.0	84.0	**0.645**	61.8	90.8	95.7	**0.794**
ViT-B/16												
CLIP2TV [18]	49.3	74.7	83.6	–	–	–	–	–	–	–	–	–
TS2-Net [33]	49.4	75.6	85.3	–	–	–	–	–	–	–	–	–
X-CLIP [35]	49.3	75.8	84.8	–	50.4	80.6	–	–	–	–	–	–
DRL [49]	50.2	76.5	84.7	–	50.0	81.5	**89.5**	–	65.7	92.6	96.7	–
SigLIP4CLIP-meanP [34]	46.2	71.9	81.6	0.633	50.1	78.3	86.7	0.680	64.2	92.2	96.4	0.812
SigLIP4CLIP-seqTransf [34]	45.7	71.0	80.0	0.623	47.4	76.6	85.0	0.659	66.1	92.9	96.8	0.824
VIDEO-COLBERT (CLIP-B/16)	51.0	**77.1**	**85.5**	**0.677**	50.2	79.6	87.8	0.683	66.8	92.9	96.8	0.826
VIDEO-COLBERT (SigLIP-B/16)	**51.5**	76.3	**85.5**	**0.677**	**55.2**	**82.9**	89.4	**0.724**	**68.0**	**93.4**	**96.9**	**0.833**

Table 1. Results on sentence-to-video retrieval tasks using MSR-VTT, MSVD and VATEX datasets. **Bold** indicates best performance for a particular model size, and underline indicates second best.

Method	DiDeMo				ActivityNet			
	R@1	R@5	R@10	nDCG	R@1	R@5	R@10	nDCG
ClipBERT [29]	20.4	48.0	60.8	–	21.3	49.0	63.5	–
All-in-One [47]	32.7	61.4	73.5	–	22.4	53.7	67.7	–
Frozen [1]	34.6	65.0	74.7	–	–	–	–	–
ViT-B/32								
CLIP4Clip-meanP [34]	43.4	70.2	80.6	–	40.5	72.4	–	–
CLIP4Clip-seqTransf [34]	43.4	70.2	80.6	–	40.5	72.4	–	–
CenterCLIP [64]	–	–	–	–	43.9	74.6	85.8	–
CLIP2TV [18]	45.5	69.7	80.6	–	44.1	**75.2**	–	–
X-CLIP [35]	45.2	74.0	–	–	44.3	74.1	–	–
DRL [49]	47.9	73.8	82.7	–	44.2	74.5	**86.1**	–
VIDEO-COLBERT (CLIP-B/32)	**48.2**	**75.1**	**83.7**	**0.654**	**45.5**	74.6	85.5	**0.645**
ViT-B/16								
CLIP4Clip-meanP [34]	44.8	75.1	–	–	44.0	73.9	–	–
CLIP4Clip-seqTransf [34]	44.8	73.4	–	–	44.5	75.2	–	–
X-CLIP [35]	47.8	**79.3**	–	–	46.2	75.5	–	–
DRL [49]	49.0	76.5	84.5	–	46.2	77.3	**88.2**	–
VIDEO-COLBERT (CLIP-B/16)	**51.9**	78.3	**85.6**	**0.682**	**50.6**	**78.0**	87.9	**0.685**
VIDEO-COLBERT (SigLIP-B/16)	51.7	76.1	84.8	0.675	45.8	76.3	86.7	0.656

Table 2. Results on paragraph-to-video retrieval tasks using DiDeMo and ActivityNet datasets. **Bold** indicates best performance for a particular model size, and underline indicates second best.

ing the cheaper CLIP-B/32 model, outperforms VIDEO-COLBERT on MSVD and VATEX, likely due to its use of channel decorrelation regularization and extra learnable token weightings. When using ViT-B/16 backbones, VIDEO-COLBERT exhibits even more impressive retrieval performance compared to alternative methods. For example, with SigLIP-B/16, VIDEO-COLBERT sets a new state-of-the-art on MSRVTT, MSVD and VATEX. The results using CLIP4Clip with an upgraded SigLIP model indicate that our performance gains are not solely attributed to the improved backbone, but rather that our two-level token-wise interaction strategy provides an effective way to match fine-grained text and video features.

In Tab. 2 we report results on two paragraph-to-video retrieval benchmarks. We find again that VIDEO-COLBERT achieves impressive results, with particularly strong performance on DiDeMo using all three backbones. While our CLIP-B/16-based model outperforms other methods on ActivityNet by a large margin, we find that the SigLIP-based variant of VIDEO-COLBERT performs relatively poorly on ActivityNet. This is likely because SigLIP was pre-trained on text with a maximum length of 16 tokens, while the cap-

Type	Name	Frame	Video	R@1	R@5	R@10	nDCG
MP	–	✓	✗	42.1	69.6	79.3	0.601
	–	✗	✓	43.4	70.4	80.1	0.610
MMS	MMS_F	✓	✗	44.3	71.5	82.4	0.626
	MMS_V	✗	✓	47.0	74.1	82.4	0.643
	MMS_{FV}	✓	✓	**48.1**	**74.9**	**83.9**	**0.652**
	RRF	✓	✓	46.8	74.6	83.8	0.646

Table 3. Effect of interaction type (coarse- and fine-grained) and interaction involvement (frame and video features) on MSR-VTT retrieval. All results use the CLIP-B/32 backbone.

tions in ActivityNet are of much longer length.

6. Additional Analysis & Discussion

We perform several additional experiments to analyze each modeling choice made in VIDEO-COLBERT using the MSR-VTT [58] dataset.

Interaction Mechanisms. In Tab. 3, we compare different strategies for interacting query token features with visual features. We observe that mean pool (MP), a coarse-grained similarity computation between the query [CLS] token and a mean pooling across frame features, results in significantly lower retrieval performance than our fine-grained token-wise interaction methods. In fact, we find that MMS operating only on static frame features (MMS_F) outperforms MP with additional temporal modeling. Employing MMS on temporally contextualized visual features (MMS_V) leads to large performance gains over the frame-only approach. Finally, we see the best overall performance when combining both frame-level and video-level token-wise similarities in MMS_{FV}, suggesting that the two metrics complement one another by enabling specialization on different concepts.

Given that MMS_F and MMS_V produce independent, yet complementary, similarity scores for each video, we consider an alternative strategy for combining their rankings rather than adding them together as in MMS_{FV}. Specifically, we explore the use of reciprocal rank fusion (RRF) [9], a widely used method for combining the outputs of multiple ranking methods in document retrieval. RRF is particularly advantageous when the similarity scores produced by different methods exist on different scales, which is possible when computing similarity scores before and after a temporal transformer. Interestingly, we find that simply summing the frame and video MMS scores (MMS_{FV}) results in superior retrieval performance compared to using RRF in evaluation. This observation suggests that the actual score differences in MMS_F and MMS_V are meaningful, and that fusing only with the reciprocal of ranks disposes of useful information, indicating potential opportunities to improve the performance even further through more score alignment between the two scoring functions.

Loss Type	Loss Function	R@1	R@5	R@10	nDCG
Combined	InfoNCE	45.1	71.4	81.4	0.625
	Sigmoid	47.1	73.1	83.9	0.647
Dual	InfoNCE	45.3	73.0	83.8	0.640
	Sigmoid	**48.1**	**74.9**	**83.9**	**0.652**

Table 4. Effect of loss type and loss function on MSR-VTT retrieval. All results use the a CLIP-B/32 backbone.

Choice of Loss Function. In Tab. 4, we analyze the effect of different loss functions on the retrieval performance of VIDEO-COLBERT. We find that the sigmoid loss [62] significantly outperforms the standard InfoNCE [39] contrastive loss typically used in T2VR. Despite not being pre-trained with sigmoid loss, CLIP-B/32 still benefits from its use during fine-tuning. We find that the choice of logit scale and bias parameters is critical for proper convergence using a sigmoid loss. We fix these values to those obtained from SigLIP pre-training ($t = 4.77$ and $b = -12.93$).

We also compare our dual sigmoid loss formulation, in which two similarity matrices are created using MMS_F and MMS_V scores, with a "combined" alternative where only a single similarity matrix is created during training by summing the frame-level and video-level matrices:

$$\mathcal{L}_{\text{combined}} = \frac{1}{1 + e^{z_{ij}(-t \cdot MMS_{FV} + b)}} \tag{8}$$

We observe a small improvement when using the dual loss over the combined method, possibly due to stronger learning objectives on the individual MMS interactions.

Query Augmentation. In Tab. 5 we analyze the effect of applying ColBERT's [26] soft query augmentation technique in the video retrieval setting. For both CLIP and SigLIP we observe improvements in retrieval performance when query padding tokens participate in the token-wise MMS interactions. The increase is more substantial when using CLIP, likely due to the attention mechanism used in the CLIP text encoder. CLIP's text encoder employs a causal attention mask, which means that earlier tokens in the sequence are not permitted to attend to future ones, effectively limiting their contextualization and thus their utility in token-wise interaction. Query augmentation can mitigate this effect by adding additional tokens to the interaction that are able to attend to all of the original query tokens. SigLIP, on the other hand, uses full (*i.e.* bidirectional) self-attention in its text encoder. Since its text tokens are already fully contextualized, the benefit of the additional tokens is more limited. However, the positive improvement gives credence to Khattab and Zaharia [26]'s hypothesis that query augmentation can introduce additional soft search terms that enhance retrieval.

In Fig. 3 we perform a more in-depth analysis of what types of queries benefit most from soft augmentation. We

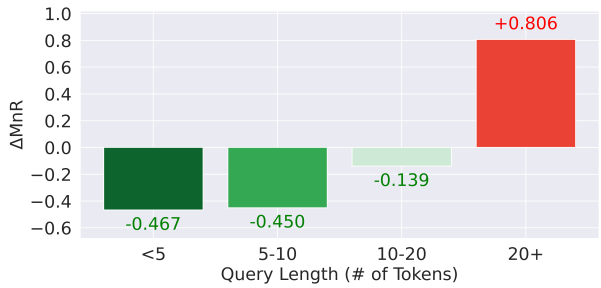

Figure 3. Effect of soft query augmentation on MSR-VTT video ranks for queries of different lengths. The plot depicts average change in rank (lower is better) using a CLIP-B/32 backbone.

Backbone	Query Aug.	R@1	R@5	R@10	nDCG
CLIP-B/32	✗	45.3	72.5	82.0	0.631
	✓	48.1	74.9	83.9	0.652
SigLIP-B/16	✗	51.0	75.9	85.1	0.675
	✓	51.5	76.3	85.5	0.677

Table 5. Effect of including pad tokens for soft query augmentation in MMS token-wise interaction. Results on MSR-VTT.

Query	Before ↓	After ↓	Δ ↓
a lady talks into a megaphone	100	5	−95
anchor talking about a shows	97	46	−51
a woman is talking about movies	50	12	−38
a man is dodging bombs	100	66	−34
a kid unwrapping his presents	2	1	−1
fox newscasters discuss chris christie and his poll numbers	3	3	0
three woman doing a fashion show to music	1	2	+1
fast moving time is shown here	2	54	+52
a person is explaining something	38	100	+62
explainin about the scene in the net	15	84	+69

Table 6. The most improved and degraded queries after using soft query augmentation (for queries <20 tokens in length). **Before:** rank of the query's target video with no query augmentation. **After:** rank of the target video when performing query augmentation. Δ: change in video rank after applying query augmentation. For all metrics, lower is better. The maximum rank is capped at 100.

observe a negative correlation between the query length and the use of expansions. Short queries of less than 20 tokens see the most benefit (as indicated by the mean rank of the target video), while queries longer than 20 tokens seem to be negatively impacted by query augmentation. One possible explanation for this phenomenon is that shorter, more abstract queries leave room for possible expansions while longer queries are either more descriptive or noisy.

In Tab. 6, we highlight some of the most improved and most degraded queries as a result of query augmentation. We observe that the queries that benefit most seem to be more intuitive to conceptualize. For example a justifiable inference about an "anchor talking about a shows" might be the location of a news room or a news logo in the video frames. Queries with no or little change in their performance seem to be sufficiently descriptive to begin with. These queries target key features of the videos that can be

# of Frames	R@1	R@5	R@10	nDCG
4	45.2	70.7	80.7	0.622
12	48.1	74.9	83.9	0.652
20	48.4	74.8	83.4	0.652

Table 7. Effect of number of sampled video frames on MSR-VTT retrieval performance using CLIP-B/32 backbone.

Temporal Transformer Depth	R@1	R@5	R@10	nDCG
2	47.0	74.0	82.3	0.642
4	48.1	74.9	83.9	0.652
8	48.5	74.1	82.6	0.650

Table 8. Effect of # of temporal transformer layers on MSR-VTT retrieval performance using CLIP-B/32 backbone.

matched with a variety of content without the need for expansion, and often perform well even without the expansions. However, we observe that queries with the most negative change from query expansions have the opposite properties to those that experience the largest gain. With query content like "fast moving time" and "a person is explaining something," it is less obvious what additional search terms could be introduced to enhance retrieval of the target video.

Number of Sampled Frames. In Tab. 7 we assess the impact of the number of sampled video frames on retrieval performance. On the MSR-VTT dataset, we see only marginal improvements from increasing the number of frames beyond 12. However, it is important to keep in mind that this analysis is highly dataset-dependent. Retrieval tasks that depend on fine-grained motion will likely benefit from a higher sampling rate, while more spatially-heavy ones will see little improvements from denser frame selection.

Temporal Transformer Depth. In Tab. 8 we experiment with different temporal transformer depths for forming the video features used in our token-wise MMS_V interaction. We find there to be a modest benefit from adding additional layers, suggesting that more powerful temporal modeling may be beneficial for retrieval. This finding, again, depends heavily on the types of videos being retrieved.

7. Conclusions

In this work, we introduced VIDEO-COLBERT, a novel approach for text-to-video retrieval that uses fine-grained interactions with both spatial and spatio-temporal visual features. Additionally, VIDEO-COLBERT is the first method to employ a sigmoid-based loss in T2VR, with our dual sigmoid loss formulation. We find that this interaction and training paradigm leads to strong representations for encoding spatial and temporal information while still being compatible when combined during retrieval. In the augmenta-

tions of our token-wise interaction, we find that augmenting the query with padding tokens is beneficial for short queries, improving on the retrieval performance from pure query-to-video interaction. Additionally, we find that using the reciprocal rank fusion, an effective ranked retrieval fusion method, hurts retrieval performance, highlighting the potential for future exploration of deeper alignment between our scoring functions.

Acknowledgments. This research was partially sponsored by the Army Research Laboratory under Cooperative Agreement W911NF-21-2-0211. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *ICCV*, pages 1708–1718, Montreal, QC, Canada, 2021. IEEE. 2, 6

[2] Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. A CLIP-Hitchhiker's Guide to Long Video Retrieval, 2022. 2

[3] Meng Cao, Haoran Tang, Jinfa Huang, Peng Jin, Can Zhang, Ruyang Liu, Long Chen, Xiaodan Liang, Li Yuan, and Ge Li. RAP: Efficient Text-Video Retrieval with Sparse-and-Correlated Adapter. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7160–7174, Bangkok, Thailand and virtual meeting, 2024. Association for Computational Linguistics. 5

[4] David Chen and William 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, Portland, Oregon, USA, 2011. ACL. 5

[5] Haoran Chen, Jianmin Li, Simone Frintrop, and Xiaolin Hu. The MSR-Video to Text Dataset with Clean Annotations. *Computer Vision and Image Understanding*, 2024. 4

[6] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset. In *NeurIPS*, 2023. 1, 5

[7] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased Contrastive Learning. In *NeurIPS*, 2020. 4

[8] Roberto Cipolla, Yarin Gal, and Alex Kendall. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *CVPR*, pages 7482–7491, Salt Lake City, UT, USA, 2018. IEEE. 4

[9] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759, 2009. 7

[10] Andreas Darmann and Christian Klamler. Using the borda rule for ranking sets of objects. *Social Choice and Welfare*, 53(3):399–414, 2019. 3

[11] Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, and Qi Wu. Prompt Switch: Efficient CLIP Adaptation for Text-Video Retrieval. In *ICCV*, pages 15602–15612, Paris, France, 2023. IEEE. 1, 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 3

[13] 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 *ICCV*, pages 13677–13687, Paris, France, 2023. IEEE. 2, 4

[14] Han Fang, Pengfei Xiong, Luhui Xu, and Wenhan Luo. Transferring Image-CLIP to Video-Text Retrieval via Temporal Relations. *IEEE Transactions on Multimedia*, 25: 7772–7785, 2023. 2

[15] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient Document Retrieval with Vision Language Models, 2024. 2

[16] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. A white box analysis of colbert. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 257–263. Springer, 2021. 4

[17] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal Transformer for Video Retrieval. In *ECCV*, 2020. 2

[18] Zijian Gao, Jingyu Liu, Weiqi Sun, Sheng Chen, Dedan Chang, and Lili Zhao. CLIP2TV: Align, Match and Distill for Video-Text Retrieval, 2022. 2, 5, 6

[19] Satya Krishna Gorti, Noel Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval. In *CVPR*, pages 4996–5005, New Orleans, LA, USA, 2022. IEEE. 1, 2, 5

[20] 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 *ICCV*, pages 11130–11139, Paris, France, 2023. IEEE. 2

[21] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, Boston, MA, USA, 2015. IEEE. 5

[22] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing Moments in Video with Natural Language. In *ICCV*, pages 5804–5813, Venice, 2017. IEEE. 5

[23] Jie Jiang, Shaobo Min, Weijie Kong, Dihong Gong, Hongfa Wang, Zhifeng Li, and Wei Liu. Tencent Text-Video Re-

trieval: Hierarchical Cross-Modal Interactions with Multi-Level Representations. *IEEE Access*, 2022. 2

[24] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A Clifton, and Jie Chen. Expectation-Maximization Contrastive Learning for Compact Video-and-Language Representations. In *NeurIPS*, 2022. 1

[25] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020. 1, 4

[26] Omar Khattab and Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *SIGIR*, pages 39–48, 2020. 1, 2, 3, 4, 7

[27] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 5

[28] Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, Nolan King, Eugene Yang, and Benjamin Van Durme. MultiVENT 2.0: A Massive Multilingual Benchmark for Event-Centric Video Retrieval, 2024. 2

[29] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling. In *CVPR*, pages 7327–7337, Nashville, TN, USA, 2021. IEEE. 1, 6

[30] Jian-Yu Li, Zhi-Hui Zhan, Yun Li, and Jun Zhang. Multiple Tasks for Multiple Objectives: A New Multiobjective Optimization Method via Multitask Optimization. *IEEE Transactions on Evolutionary Computation*, pages 1–1, 2023. 4

[31] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked Teacher: Towards Training-Efficient Video Foundation Models. In *ICCV*, 2024. 1

[32] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use What You Have: Video Retrieval Using Representations From Collaborative Experts, 2020. 2

[33] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. TS2-Net: Token Shift and Selection Transformer for Text-Video Retrieval. In *ECCV*, pages 319–335, Cham, 2022. Springer Nature Switzerland. 2, 5, 6

[34] 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, 3, 5, 6

[35] 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 *MM*, pages 638–647, New York, NY, USA, 2022. Association for Computing Machinery. 1, 2, 5, 6

[36] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640, Seoul, Korea (South), 2019. IEEE. 2

[37] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding Language-Image Pretrained Models for General Video Recognition. In *ECCV*, 2022. 2

[38] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document Ranking with a Pretrained Sequence-to-Sequence Model, 2020. 1

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, 2019. 4, 7

[40] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *ICLR*, 2021. 6

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 1, 2, 3, 5

[42] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 1

[43] Mamshad Nayeem Rizve, Fan Fei, Jayakrishnan Unnikrishnan, Son Tran, Benjamin Z. Yao, Belinda Zeng, Mubarak Shah, and Trishul Chilimbi. VidLA: Video-Language Alignment at Scale. In *CVPR*, pages 14043–14055, Seattle, WA, USA, 2024. IEEE. 1, 5

[44] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation, 2020. 4

[45] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. In *ICCV*, 2019. 1

[46] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. In *NeurIPS*, pages 38032–38045, Red Hook, NY, USA, 2022. 1

[47] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 6

[48] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV*, 2016. 5

[49] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled Representation Learning for Text-Video Retrieval, 2022. 1, 2, 3, 5, 6

[50] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang

Wang, and William Yang Wang. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *ICCV*, pages 4580–4590, Seoul, Korea (South), 2019. IEEE. 5

[51] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2VLAD: Global-Local Sequence Alignment for Text-Video Retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5075–5084, Nashville, TN, USA, 2021. IEEE. 2

[52] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. In *ICLR*, 2024. 1

[53] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. In *ECCV*, 2024. 1

[54] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified Coarse-to-Fine Alignment for Video-Text Retrieval. In *ICCV*, pages 2804–2815, Paris, France, 2023. IEEE. 2

[55] Michael Wray, Hazel Doughty, and Dima Damen. On Semantic Similarity in Video Retrieval. In *CVPR*, pages 3649–3659, Nashville, TN, USA, 2021. IEEE. 5

[56] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval? In *CVPR*, pages 10704–10713, Vancouver, BC, Canada, 2023. IEEE. 2, 5

[57] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *EMNLP*, 2021. 2

[58] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*, pages 5288–5296, Las Vegas, NV, USA, 2016. IEEE. 5, 7

[59] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Alignment. In *ICLR*, 2023. 2

[60] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *ICLR*, 2022. 2, 3

[61] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-Shot Transfer with Locked-image text Tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18102–18112, New Orleans, LA, USA, 2022. IEEE. 1

[62] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *ICCV*, pages 11941–11952, Paris, France, 2023. IEEE. 1, 3, 4, 5, 7

[63] Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. Video-aided Unsupervised Grammar Induction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1513–1524, Online, 2021. Association for Computational Linguistics. 4

[64] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. CenterCLIP: Token Clustering for Efficient Text-Video Retrieval. In *SIGIR*, pages 970–981, New York, NY, USA, 2022. Association for Computing Machinery. 2, 5, 6

[65] Linchao Zhu and Yi Yang. ActBERT: Learning Global-Local Video-Text Representations. In *CVPR*, pages 8743–8752, Seattle, WA, USA, 2020. IEEE. 2

Video-ColBERT: Contextualized Late Interaction for Text-to-Video Retrieval

Supplementary Material

8. Additional Training Details

We provide additional details about our training configurations in Tab. 9. We also indicate dataset-specific settings, like batch size and number of epochs, in Tab. 10.

Setting	Value
Learning Rate Schedule	Linear
Warmup Proportion (Linear)	10%
CLIP Param. Learning Rate	1e-7
Temporal Layer Learning Rate	1e-4
Optimizer	Adam
Adam Betas	$\beta_1 = 0.9, \beta_2 = 0.98$
Adam ϵ	1e-6
Weight Decay	0.01
Max. Grad. Norm	1

Table 9. Training settings for VIDEO-COLBERT.

Dataset	Backbone Type	Batch Size	Epochs
MSR-VTT	ViT-B/32	256	5
	ViT-B/16	128	5
MSVD	ViT-B/32	256	5
	ViT-B/16	128	5
VATEX	ViT-B/32	256	10
	ViT-B/16	128	10
DiDeMo	ViT-B/32	64	20
	ViT-B/16	64	20
ActivityNet	ViT-B/32	64	20
	ViT-B/16	64	20

Table 10. Dataset-specific training settings.

9. Effect of Query Pad Token Choice

In Tab. 11, we show how video retrieval results are affected by different choices of padding token when using soft query augmentation in VIDEO-COLBERT with a CLIP-B/32 backbone. Ordinarily (*e.g.* when using only the special aggregation token to represent the query), the choice of padding token does not have any influence on retrieval outcomes. However, when performing soft query augmentation, all self-attention operations involve padding tokens, and the outputs of these extra tokens are used for interaction with visual features. As a result, the choice of pad token *does* have an impact on retrieval results when using query augmentation and token-wise interaction. Because we freeze the token embeddings in the text encoder, we find that the choice of padding token has a noticeable effect on retrieval metrics. This is due to the fact that certain tokens will have pre-existing semantics that are better aligned with the query augmentation task than others. We found that the exclamation mark leads to the best performance out of the options we considered.

Token ID	Token Text	R@1	R@5	R@10	nDCG
31	`@`	46.0	74.6	83.3	0.644
49407	`<|endoftext|>`	46.0	73.3	82.3	0.638
3002	`...`	47.8	74.6	83.6	0.652
13530	`___</w>`	47.9	72.8	83.5	0.646
49406	`<|startoftext|>`	48.0	74.3	**84.0**	**0.653**
0	`!`	**48.1**	**74.9**	83.9	0.652

Table 11. Effect of choice of padding token for soft query augmentation. Results on MSR-VTT using CLIP-B/32 backbone.

10. Visualization

In Fig. 4, we explore how interactions between text tokens and frame representations change before and after the temporal transformer layers by visualizing the maximally similar frame to certain query tokens. To enhance the interpretability of this exploration, we do not use query or visual expansion during encoding. Generally, we find that the frame representations before and after the temporal encoder behave differently during interaction with the text tokens. In Fig. 4, the most obvious shift is in the similarities of "field" and "street." Prior to the temporal encoding, "street" and "field" correspond to frames that clearly represent the singular visual concept: a large grassy field with the car in the distance, and the street from the first person view of the car. After the temporal encoder, they then become associated with new frames: one with the car slightly on the grass field and another when the car is driving back onto the street. We interpret these results as a sign of stronger temporal contextualization in the frame representations after the encoding. Specifically, the associated frames seem to shift from depicting static concepts to more dynamic ones when temporally contextualized features are used.

Figure 4. Visualization of the interactions between query tokens and video frames before and after the temporal encoder of VIDEO-COLBERT, trained on MSR-VTT. The green arrow (|) represents the interaction between query tokens and frames **before** temporal encoding. The red arrow (|) represents the interaction between query tokens and frames **after** the temporal encoding.