VideoCLIP-XL: Advancing Long Description Understanding for Video CLIP Models

Jiapeng Wang[1]*, **Chengyu Wang**[2]†, **Kunzhe Huang**[2], **Jun Huang**[2], **Lianwen Jin**[1]†
[1]South China University of Technology, Guangzhou, China
[2]Alibaba Cloud Computing, Hangzhou, China
eejpwang@mail.scut.edu.cn, eelwjin@scut.edu.cn
{chengyu.wcy, huangkunzhe.hkz, huangjun.hj}@alibaba-inc.com

Abstract

Contrastive Language-Image Pre-training (CLIP) has been widely studied and applied in numerous applications. However, the emphasis on brief summary texts during pre-training prevents CLIP from understanding long descriptions. This issue is particularly acute regarding videos given that videos often contain abundant detailed contents. In this paper, we propose the **VideoCLIP-XL** (*eXtra Length*) model, which aims to unleash the long-description understanding capability of video CLIP models. Firstly, we establish an automatic data collection system and gather a large-scale **VILD** pre-training dataset[1] with *VIdeo* and *Long-Description* pairs. Then, we propose Text-similarity-guided Primary Component Matching (**TPCM**) to better learn the distribution of feature space while expanding the long description capability. We also introduce two new tasks namely Detail-aware Description Ranking (**DDR**) and Hallucination-aware Description Ranking (**HDR**) for further understanding improvement. Finally, we construct a Long Video Description Ranking (**LVDR**) benchmark[2] for evaluating the long-description capability more comprehensively. Extensive experimental results on widely-used text-video retrieval benchmarks with both short and long descriptions and our LVDR benchmark can fully demonstrate the effectiveness of our method.[3]

1 Introduction

The Contrastive Language-Image Pre-training (CLIP) model (Radford et al., 2021) represents a pivotal development in the field of vision-language pre-training. It integrates text and image encoders to align these two modalities through contrastive learning. This methodology has been effectively applied in various applications, such as zero-shot classification (Sun et al., 2023), text-image retrieval (Luo et al., 2023), and text-to-image generation (Rombach et al., 2022; Frans et al., 2022). However, one of the notable limitations of CLIP is its constrained capacity to process extensive textual descriptions, owing to its text encoder's reliance on maximum positional embeddings with length 77. This limitation greatly restricts the length of input text, and existing studies (Zhang et al., 2024) have also revealed a *de facto* effective token limit of just around 20.

Furthermore, the vanilla CLIP training procedure's emphasis on brief summary texts compels the text/vision encoder to focus predominantly on the main features of the text/visual input, often overlooking smaller, yet potentially critical details. This issue is *particularly acute in videos* as compared to images, given that videos encapsulate a wealth of details across successive frames, along with additional information such as the sequence and flow of activities, camera movements, etc. In this context, existing video CLIP models (Xu et al., 2021; Luo et al., 2022; Wang et al., 2023c) that employ vanilla CLIP training methodologies may struggle to accurately capture complex relationships and attributes, due to their reliance on a simplistic "bag of concepts" approach (Tang et al., 2023b). To overcome these limitations, enhancing the model's capability to comprehend long descriptions is crucial. Longer texts provide a rich tapestry of attributes and interconnections, offering a pathway to significantly improve the model's performance and applicability in more complex scenarios.

To this end, we propose **VideoCLIP-XL** (*eXtra Length*), to our knowledge, which is the first video CLIP model with long-description capability. (1) To be specific, recognizing the insufficiency of public datasets containing (*video*, *long*

*Contribution during internship at Alibaba Cloud Computing.
†Co-corresponding authors.
[1]https://huggingface.co/alibaba-pai/VILD.
[2]https://huggingface.co/alibaba-pai/LVDR.
[3]https://huggingface.co/alibaba-pai/VideoCLIP-XL.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16061–16075

description) pairs, we establish an automatic data collection system, designed to aggregate sufficient and high-quality pairs from multiple data sources. We have successfully collected over 2M (***VI**deo*, ***L**ong **D**escription*) pairs, denoted as our **VILD** pre-training dataset. (2) We have discovered that existing CLIP models for long texts (Zhang et al., 2024) lack the flexibility to dynamically adapt to the distribution changes within high-dimensional feature space. To address this issue, we introduce **T**ext-similarity-guided **P**rimary **C**omponent **M**atching (**TPCM**), a novel approach that enables the model to better learn cross-modal and cross-sample relative distances. (3) We claim that there are two attributes that CLIP models with the long-description understanding capability should naturally possess: for a given video and its associated descriptions, it should be able to assign a higher score when the description contains i) more rich and precise detailed contexts; or ii) fewer hallucinations with the same level of detail. To this end, we propose two new tasks to model these two attributes namely **D**etail-aware **D**escription **R**anking (**DDR**) and **H**allucination-aware **D**escription **R**anking (**HDR**). They make the video CLIP model to learn how to correctly rank multiple descriptions with different levels of details and hallucinations. (4) In order to better evaluate video CLIP models, we further release a **L**ong **V**ideo **D**escription **R**anking (**LVDR**) benchmark. Given each video and the corresponding ground-truth long description (after human correction) sampled from Shot2Story (Han et al., 2023), we iteratively modify a certain proportion of correct contents into hallucination in each step. The model is required to correctly rank these descriptions according to their faithfulness.

To evaluate the performance of VideoCLIP-XL, we conduct extensive experiments not only on the video & long-description dataset Shot2Story (Han et al., 2023), but also on traditional widely-used MSR-VTT (Xu et al., 2016), LSMDC (Rohrbach et al., 2015), DiDeMo (Anne Hendricks et al., 2017), MSVD (Chen and Dolan, 2011) and ActivityNet (Heilbron et al., 2015) benchmarks, for the text-video retrieval task. Moreover, we evaluate VideoCLIP-XL and other representative CLIP models on our proposed LVDR benchmark. Experimental results demonstrate that our method exhibits superior performance compared with state-of-the-art competitors.

Our main contributions are as follows:

- We propose the VideoCLIP-XL model to unleash the long-description understanding capability of video CLIP models. We also collect and release a new pre-training dataset VILD with over 2M video & long-description pairs using our automatic data collection system.
- In VideoCLIP-XL, we propose TPCM for dynamic feature learning while expanding the long description capability. We also propose two new tasks (i.e., DDR and HDR) to further model the effective attributes for better representation learning of long descriptions.
- To better evaluate video CLIP models' long description ability, we propose the LVDR benchmark for long description ranking.
- Extensive experiments show that VideoCLIP-XL clearly outperforms state-of-the-art models over various tasks and benchmarks.

2 Related Work

Image/Video CLIP models. CLIP (Radford et al., 2021) is a multimodal model based on contrastive learning. Its training data comprises a vast collection of text-image pairs, each image paired with a corresponding text description. Through contrastive learning, the model learns the matching relationship between text-image pairs. Owing to its robust zero-shot generalization capabilities, CLIP has been successfully deployed in numerous scenarios including detection (Gu et al., 2021; Li et al., 2022b), segmentation (Xu et al., 2022; Li et al., 2022a), image/video understanding (Luo et al., 2022; Xu et al., 2021; Tang et al., 2023a), retrieval (Wang et al., 2023a,b) and image generation (Ramesh et al., 2022; Frans et al., 2022; Crowson et al., 2022; Vinker et al., 2022). For video analysis, ViCLIP (Wang et al., 2023c) incorporates spatio-temporal attention within its video encoder and adopts partial random patch masking during training. Nonetheless, several subsequent studies (Kim et al., 2023; Zeng et al., 2021) have identified CLIP's inadequacy in extracting fine-grained information. These works implement contrastive methods similar to CLIP's to align complete sentence tokens with regions of the entire image. Furthermore, Long-CLIP (Zhang et al., 2024) proposes the use of primary component matching of CLIP features to improve the model's understanding of lengthy descriptions in images.

Vision-Language Datasets. As the capabilities of multimodal models advance, the need transcends traditional fixed-category image datasets such as ImageNet (Deng et al., 2009) and CIFAR10 (Krizhevsky et al., 2009). Contemporary open-world applications require datasets that encompass both images/videos and their associated text descriptions. Common open-world image-language datasets include Visual Genome (Krishna et al., 2017), Conceptual-12M (Changpinyo et al., 2021), SBU (Ordonez et al., 2011), COCO (Lin et al., 2014), and LAION-5B (Schuhmann et al., 2022). Typical video-language datasets comprise MSR-VTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), LSMDC (Rohrbach et al., 2015), WebVid (Bain et al., 2021), InternVid (Wang et al., 2023c), and Panda-70M (Chen et al., 2024). However, these datasets generally contain only short captions. On the other hand, a few datasets focus on long descriptions. ShareGPT4V (Chen et al., 2023) is a large-scale dataset with 1.2M long captions for images. Shot2Story (Han et al., 2023) includes 20K video clips, each with detailed shot-level captions and comprehensive video summaries. MiraData (Ju et al., 2024) deals with uncut video segments and features structural long captions. It contains 57.8K video clips across two scenarios: gaming and city/scenic exploration. The average description length in these collections is often orders of magnitude longer than those in previous datasets (Zhang et al., 2024).

3 Methodology

In this section, we introduce our automatic data collection system and the resulting *VIdeo & Long-Description* (VILD) pre-training dataset (Sect. 3.1), the text-similarity-guided primary component matching (TPCM) technique (Sect. 3.2), two new description ranking tasks (Sect. 3.3), and the new Long Video Description Ranking (LVDR) benchmark dataset (Sect. 3.4).

3.1 VIdeo & Long-Description (VILD) Dataset

Training CLIP models often necessitates a substantial corpus of vision-text pairs. In image processing, the advent of open-source large multimodal models (LMMs) and the availability of APIs such as GPT-4V (Achiam et al., 2023) have spurred efforts to annotate images with detailed long descriptions. For example, ShareGPT4V (Chen et al., 2023) is a large dataset, originating from a high-quality curated set of 100K captions gathered using GPT-4V and expanded to 1.2M through a caption model.

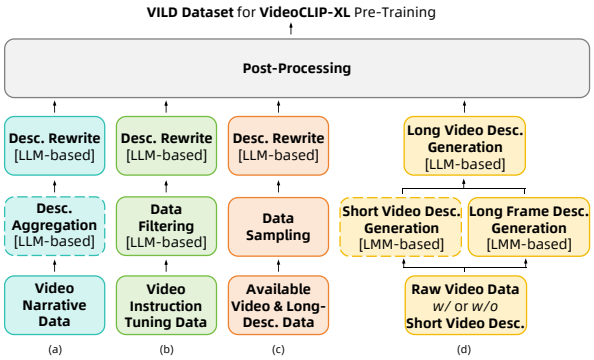


Figure 1: The automatic data collection system for our VILD dataset. Desc. is short for description.

However, video datasets with extensive long descriptions, especially in the open domain, remain markedly scarce. For instance, Shot2Story (Han et al., 2023) offers 20K video clips, each accompanied by shot-level captions and video summaries. After annotating with LMMs, further manual corrections ensure the reliability of these long descriptions, *thereby qualifying it as a trustworthy evaluation set, which is excluded from our training data*. MiraData (Ju et al., 2024) leverages GPT4V to produce long captions for 57.8K video clips, restricted to gaming and city/scenic exploration scenarios. Open-Sora-Dataset (PKU-YuanGroup, 2024) utilizes LMMs to generate descriptive narratives for 40.2K videos, predominantly in natural landscapes.

In light of the scarcity of open-domain video & long-description pairs, we engineer an automatic data collection system, as depicted in Fig. 1. Our approach harnesses multiple sources, chiefly encompassing video narrative data, video instruction tuning data, raw videos, and available video & long-description pairs.

(a) Video Narrative Data. Video narrative data often contains human-concerned descriptions produced by human annotators, which can describe the whole scene, the main activities, and events involving multiple actors and objects. We adopt the VidLN (Voigtlaender et al., 2023) dataset, consisting of human-annotated individual-level descriptions of each single main people/animal/objective in the video along with the background. To make the dataset serve our purpose, we employ large language models (LLMs) to aggregate individual-level narratives into whole-level descriptions through prompt engineering (i.e., the *Desc. Aggregation*

step). Finally, in consideration of training efficacy and robustness, we further utilize LLMs to rewrite the whole-level description (i.e., the *Desc. Rewrite* step). This process involves generating varied textual descriptions of the same meaning, while preserving both main contents and detail attributes unchanged. The details of utilized LLMs and prompts used in the two steps are shown in Appendix A.1.

(b) Video Instruction Tuning Data. Alongside with the emergence of LMMs, extensive video instruction tuning datasets have also been publicly available. For example, VideoInstruct100K (Maaz et al., 2023) contains question-answer pairs related to video summarization, description-based question-answers, and creative/generative question-answers. VideoChat (Li et al., 2023b) provides a rich dataset featuring elaborate video descriptions and dialogues, enhancing data variety by embracing temporal and causal aspects within video instructions. These datasets were originally crafted to train a genre-independent video understanding model, rather than to curate video descriptions. Consequently, our method includes employing LLMs for *Data Filtering* to exclude samples extraneous to video descriptions. We employ prompt engineering and also provide some demonstration examples to aid LLMs in achieving better affects. Finally, the *Desc. Rewrite* step is also conducted. The details of utilized LLMs and prompts are shown in Appendix A.1.

(c) Available Video & Long-Description Data. As previously mentioned, existing datasets pairing videos with long text descriptions are often limited by either the quantity or the domains/genres of videos. In this regard, we perform the *Data Sampling* operation over these datasets. Specifically, 57.8K video clips of gaming and city/scenic exploration scenarios in MiraData (Ju et al., 2024) are all included in VILD. 50K long captions describing natural landscape are randomly sampled from Open-Sora-Dataset (PKU-YuanGroup, 2024). The *Desc. Rewrite* step is also involved at the end.

(d) Raw Video Data. In order to further expand the amount of training data, we leverage LMMs and LLMs to generate long descriptions given raw videos (some combined with corresponding short captions). An optional *Short Video Desc. Generation* step is required using off-the-shelf models (Li et al., 2023a; Huang et al., 2023; Zhang et al., 2023; Yu et al., 2023) if there are no short captions available. For computation efficiency, we randomly sample over 2M video clips with high-quality short captions generated by a number of teacher models and a fine-tuned caption selection model from Panda-70M (Chen et al., 2024). Then, we sample k (k=3 in our setting) frames from each video clip at equal intervals as key-frames and use LMMs to annotate them with long descriptions. We do not conduct this for each frame, as it would be extremely time-consuming and laborious. Next, given short description of the whole video and long descriptions of its key-frames, we ask LLMs to integrate them into long description of the whole video. The assistance of short video description can alleviate the hallucinations present in frame descriptions. Our findings have also reached a consensus with existing studies (Wang et al., 2023c, 2024) that directly using video LMMs (Li et al., 2023b; Maaz et al., 2023) to describe videos for long captions can lead to sub-optimal results. The details of utilized LLMs/LMMs and prompts are shown in Appendix A.1.

Finally, the *Post-Processing* step is performed. NSFW examples are filtered out. Next, we utilize ViCLIP (Wang et al., 2023c) and LongCLIP (Zhang et al., 2024) to filter out low-quality examples with average video-text similarity smaller than 0.20. We finally collect over 2M video & long-description data pairs as our VILD dataset for model pre-training. More detailed comparisons of data statistics information are shown in Appendix A.2.

3.2 Text-similarity-guided Primary Component Matching (TCPM)

The vanilla pre-training of CLIP models consumes vision-text pairs (v, t) as inputs. v can be images or videos. It makes no assumptions on specific single-modal encoder architectures. Given a vision encoder E_v and a text encoder E_t, single-modal features are first extracted as $f_v = E_v(v)$, $f_t = E_t(t)$. Then, contrastive learning with the InfoNCE (Oord et al., 2018) loss typically is employed to learn the correspondence between vision and text. In particular, this can be formulated as:

$$\mathcal{L}_{\text{CL}}(f_t, f_v) = \frac{1}{2N} \sum\nolimits_{N} \mathcal{L}_{\text{InfoNCE}}^{f_t \to f_v} + \mathcal{L}_{\text{InfoNCE}}^{f_v \to f_t}, \quad (1)$$

where N is the batch size and

$$\mathcal{L}_{\text{InfoNCE}}^{f_t \to f_v} = -\log \frac{\exp(sim(f_t, f_v^+)/\tau)}{\sum_{f_v \in \{f_v^+, f_v^-\}} \exp(sim(f_t, f_v)/\tau)}, \quad (2)$$

and vise versa. Here, τ is the temperature hyper-parameter, sim is the cosine similarity calculation,

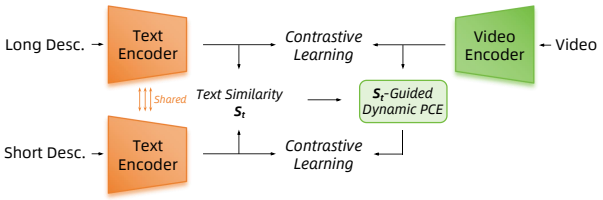


Figure 2: The proposed text-similarity-guided primary component matching (TPCM) technique.

f_v^+ is the *positive* vision feature which is paired with the text feature f_t, and f_v^- are *negative* vision features that are formed by other unpaired images/videos in the current training batch.

To extend the long-description understanding capacity of CLIP models, Long-CLIP (Zhang et al., 2024) is proposed to use primary component matching for image CLIPs. Given the short description, the long description, and the vision input (*st*, *lt*, *v*), the loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{CL}}(f_{lt}, f_v) + \alpha_1 \mathcal{L}_{\text{CL}}(f_{st}, f'_v), \quad (3)$$

where α_1 is the ratio hyper-parameter and $f'_v = \text{PCE}(f_v,\ 32)$. Here, PCE is short for primary component extraction that consists of the component-decomposition function $\mathcal{F}$ (which decomposes the feature into vectors of different attributes and their importance), the component-filtration function $\mathcal{E}$ (which filters out less important attributes), and the component reconstruction function $\mathcal{F}^{-1}$ (which reconstructs the feature). In the implementation of $\mathcal{E}$, Long-CLIP selects the most important 32 attributes as the retained ones.

However, when extending this technique for video pre-training, we have found that since videos usually contain richer contents and more details than images, this fixed strategy cannot dynamically adapt to the severe distribution changes of high-dimensional feature spaces of video CLIPs during learning (shown in Fig. 5). In this regard, we propose to use the cosine text similarity between lt and st as a signal to guide the PCE process, as shown in Fig. 2. Therefore, we re-write $\hat{f}_v$ as follows:

$$\hat{f}_v = \text{PCE}(f_v,\ \mathcal{G}(sim(f_{lt},\ f_{st}))), \quad (4)$$

where $\mathcal{G}$ represents that we preserve the attributes in descending order of importance until the similarity between $\hat{f}_v$ and f_v reaches the similarity between lt and st.

3.3 Two Description Ranking Tasks

We posit that video CLIP models designed to comprehend long descriptions should inherently exhibit two characteristics: given a video and its associated descriptions, the model should assign a higher score to descriptions (1) with richer and more precise contexts and (2) that are more accurate and less prone to hallucinations, given an equivalent level of detail. To realize these principles, we introduce two novel tasks: Detail-aware Description Ranking (DDR) and Hallucination-aware Description Ranking (HDR) to address the respective attributes. Our preparatory steps involve employing syntactic analysis tools, such as NLTK (Bird et al., 2009) and spaCy (Honnibal et al., 2020), to execute part-of-speech tagging and parse syntactic structures within the long-description ground truths.

Subsequently, we synthesize multiple description candidates for each video to facilitate DDR and HDR training. As illustrated in Fig. 3(a), in each step, we selectively replace specific words (nouns, numerals, colors, or terms related to direction, verbs) with their semantically disparate counterparts within the same syntactic category (e.g., *boys* to *girls*, *white* to *blue*, *throwing* to *lifting*), and perform this replacement $m-1$ times. This method yields a series of progressively hallucinated descriptions, denoted as $\mathbf{t^H} = \{t_1^H, t_2^H, \ldots, t_m^H\}$. Analogously, as depicted in Fig. 3(b), each step involves randomly excising sub-sentences, adjectives, numerals, or dependency parse sub-trees from the current description. This process recursively gener-

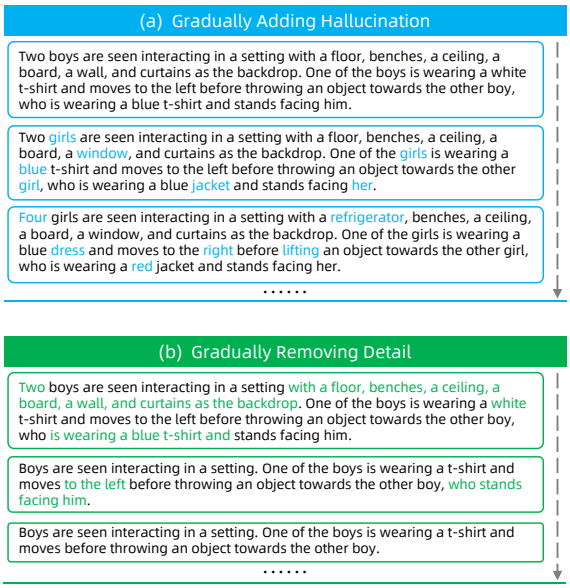


Figure 3: Examples of text samples generated for (a) hallucination-aware and (b) detail-aware description ranking tasks. *Blue* and *green* words refer to replaced hallucination content and detailed content to be deleted, respectively. Best viewed in color.

(a)

[A]: In the video, two women are standing at a wooden table in a kitchen filled with various cooking tools and food materials. The woman in dark green is speaking to the camera while the woman in gray laughs and cracks an egg into a black bowl, placing the eggshell in a green bowl. The kitchen is well-equipped with a refrigerator and an oven, and there is a beautiful wreath hanging on the cupboard door. As the scene continues, the woman in gray lowers her head and stirs something in a [5]***bowl*** while talking, with the oven nearby. In the next part of the video, a person is seen stirring egg liquid in a purple bowl with an egg beater. Another person pours [6]***white*** paste from a red container into the [3]***egg*** liquid and then adds white powder from a [1]***red*** bowl. The woman stirring the mixture puts down the egg beater and opens the lid of a white bottle on the table. The table also has six eggs on a blue [4]***egg*** rack, along with other cooking tools and food materials. Throughout the video, a female voice narrates the steps, which include whisking up eggs, adding fresh ricotta, and incorporating some Parmesan [2]***cheese***. The entire video takes place in a cozy kitchen setting, with the participants engaging in a cooking activity.
[B]: [1]red→cyan, [2]cheese→tomato
[C]: [1]red→cyan, [2]cheese→tomato, [3]egg→potato, [4]egg→bread
[D]: [1]red→cyan, [2]cheese→tomato, [3]egg→potato, [4]egg→bread, [5]bowl→plate, [6]white→gold

VideoCLIP-XL: A>B>C>D Long-CLIP: D>C>B>A
ViCLIP: A=B=C=D CLIP: A=B=C=D

(b)

[A]: The video begins with a woman, dressed in a [4]***white*** suspender, [1]***shorts***, and boots, sitting alone on an expansive sandy field. Her knees bear the marks of a recent fall, as indicated by her scraped knees and her own admission of having fallen. An unseen individual extends a hand to help her up, offering assistance after her tumble. The scene then transitions to the same [2]***woman***, still in her [5]***white*** [6]***boots***, now skateboarding across the sandy terrain. The skateboard is unique, featuring a single wheel in the center with the [7]***woman***'s feet positioned on either side of it. She skillfully maneuvers the [8]***board*** across the sandy field, a stark contrast to her previous fall. The final scene introduces a new character, a man wearing goggles and an eye mask. He [9]***stands*** on the same sandy field, with the woman now rollerblading in the background, her [3]***white*** clothing billowing in the wind. A sandstorm begins to form in the distance behind them, adding an element of drama to the scene. Despite the impending storm, the pair continue their activities, seemingly unfazed by the harsh conditions.
[B]: [1]shorts→jackets, [2]woman→elder, [3]white→brown
[C]: [1]shorts→jackets, [2]woman→elder, [3]white→brown, [4]white→grey, [5]white→peach, [6]boots→sneakers
[D]: [1]shorts→jackets, [2]woman→elder, [3]white→brown, [4]white→grey, [5]white→peach, [6]boots→sneakers, [7]woman→man, [8]board→clock, [9]stands→pushes

VideoCLIP-XL: A>B>C>D Long-CLIP: A>B>D>C
ViCLIP: D>A=B>C CLIP: A=B>D>C

Figure 4: Qualitative examples on our LVDR benchmark. We calculate the cosine similarities between different long descriptions and the same video using video CLIP models. Descriptions are sorted based on these similarities in descending order after retaining 8 decimal places.

ates $m-1$ sequentially less detailed descriptions for each video, expressed as $\mathbf{t^D} = \{t_1^D, t_2^D, \ldots, t_m^D\}$.

For $\mathbf{t^H}$ or $\mathbf{t^D}$, given the same corresponding video, we hope that the model can generate a higher similarity score for the description appearing earlier in the sequence. For example, for the DDR task, we formulate the loss function as follows:

$$\mathcal{L}_{\text{DDR}} = \frac{1}{\frac{m(m-1)}{2}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \text{ReLU}(-(\Delta_{i,j}^D - \alpha_D)), \tag{5}$$

where α_D is the similarity difference gap and

$$\Delta_{i,j}^D = sim(f_{t_i^D}, f_v) - sim(f_{t_j^D}, f_v). \tag{6}$$

The intuition behind this learning objective comes from the requirement for the model to be able to differentiate between various descriptions with the minimum distinction degree α_D. Similarly, for HDR, we have the loss function:

$$\mathcal{L}_{\text{HDR}} = \frac{1}{\frac{m(m-1)}{2}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \text{ReLU}(-(\Delta_{i,j}^H - \alpha_H)). \tag{7}$$

The total loss of our pre-training process is:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\text{CL}}(f_{lt}, f_v) + \alpha_1 \mathcal{L}_{\text{CL}}(f_{st}, f_v') + \\ & \alpha_2 \mathcal{L}_{\text{DDR}} + \alpha_3 \mathcal{L}_{\text{HDR}},\end{aligned} \tag{8}$$

where α_2 and α_3 are balancing hyper-parameters.

3.4 The New LVDR Benchmark

Hallucination is ubiquitous in contemporary LLMs and LMMs (Liu et al., 2024a). Given a video, the video CLIP model with the ability to understand long texts should naturally possess the discernment to distinguish between correct and erroneous texts in long descriptions. To better evaluate such ability, we propose the Long Video Description Ranking (LVDR) benchmark. We first randomly sample 2K video & long-description pairs from Shot2Story (Han et al., 2023). Then, we perform a synthesis process similar to Fig. 3(a), iterating $p-1$ times and altering q words during each iteration, and resulting in totally p descriptions with gradually increasing degrees of hallucination. We denote such a subset as $p \times q$ and construct five subsets as $\{4\times1, 4\times2, 4\times3, 4\times4, 4\times5\}$. Each distinct subset is manually reviewed to avoid inappropriate replacement. Representative examples are provided in Fig. 4. Based on our analysis,

Method	MSR-VTT		LSMDC		DiDeMo		MSVD		ActivityNet		Avg.	
	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
CLIP (Radford et al., 2021)	30.4	24.2	13.9	11.9	12.7	18.7	40.5	57.2	9.1	13.2	21.3	25.0
VideoCLIP (Xu et al., 2021)	10.4	-	-	-	16.6	-	-	-	-	-	-	-
CLIP4Clip (Luo et al., 2022)	32.0	-	15.1	-	-	-	38.5	-	-	-	-	-
ViCLIP (Wang et al., 2023c)	42.4	41.3	20.1	16.9	38.7	39.1	49.1	75.1	32.1	31.4	36.5	40.8
VideoCLIP-XL (Ours)	**50.1**	**49.9**	**22.8**	**24.6**	**47.7**	**47.9**	**51.9**	**76.7**	**46.4**‡	**48.1**‡	**43.8**	**49.5**

Table 1: R@1 scores of zero-shot text-video retrieval on MSR-VTT (Xu et al., 2016), LSMDC (Rohrbach et al., 2015), DiDeMo (Anne Hendricks et al., 2017), MSVD (Chen and Dolan, 2011), and ActivityNet (Heilbron et al., 2015). T2V and V2T are short for text-to-video and video-to-text, hereinafter the same. ‡Due to the high overlap between the videos in ActivityNet and VideoInstruct100K (Maaz et al., 2023), the latter is excluded from the pre-training data of our model tested on the former.

Method	MSR-VTT		LSMDC		DiDeMo		MSVD		ActivityNet		Avg.	
	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
CLIP (Radford et al., 2021)	38.2	38.7	22.5	22.6	32.2	33.9	52.3	69.9	26.1	26.9	34.3	38.4
VideoCLIP (Xu et al., 2021)	30.9	-	-	-	-	-	-	-	-	-	-	-
CLIP4Clip (Luo et al., 2022)	45.6	45.9	24.3	23.8	43.0	43.6	45.2	48.4	40.3	41.6	39.7	40.7
ViCLIP (Wang et al., 2023c)	52.5	51.8	33.0	32.5	49.4	50.2	53.1	79.0	49.8	48.1	47.6	52.3
VideoCLIP-XL (Ours)	**57.0**	**56.6**	**34.2**	**32.6**	**62.3**	**62.7**	**55.6**	**81.4**	**58.4**‡	**59.2**‡	**53.5**	**58.5**

Table 2: R@1 scores of fine-tuned text-video retrieval on MSR-VTT, LSMDC, DiDeMo, MSVD, and ActivityNet. ‡Due to the high overlap between the videos in ActivityNet and VideoInstruct100K (Maaz et al., 2023), the latter is excluded from the pre-training data of our model tested on the former.

a better model needs to be able to correctly rank these descriptions in descending order of similarity given the video. Thus, we also design the evaluation criterion named ranking score (RS) which can be formulated as:

$$\text{RS} = \frac{100}{\frac{m(m-1)}{2}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathbb{1}(sim(f_{t_i}, f_v) > sim(f_{t_j}, f_v)). \tag{9}$$

Here, $\mathbb{1}$ is the indicator function.

4 Experiments

4.1 Implementation Detail

We adopt the model structure of CLIP (Radford et al., 2021) with ViT-L/14 and leverage spatio-temporal attention in the video encoder with the weight initialization from ViCLIP (Wang et al., 2023c). We further pre-train VideoCLIP-XL on our VILD dataset for 2 epochs. All experiments are implemented in PyTorch and run on NVIDIA Tesla A100-80G GPUs. More experimental details are given in Appendix A.3.

4.2 Performance Comparison

We compare VideoCLIP-XL with strong competitors in three different downstream tasks: text-video retrieval on traditional benchmarks, text-video retrieval on long-description benchmarks, and description ranking on our LVDR benchmark.

Results on traditional benchmarks for text-video retrieval are shown in Tab. 1 and 2. We can find that, VideoCLIP-XL exhibits superior performance on all benchmarks compared with other video CLIP models under both zero-shot and fine-tuning settings. For example, VideoCLIP-XL outperforms the previous state-of-the-art ViCLIP, with an improvement of +7.7/+8.6 T2V/V2T zero-shot R@1 scores and +4.5/+4.8 T2V/V2T fine-tuning R@1 scores on MSR-VTT. It is worth noting that, although our method mainly focuses on learning fine-grained features in videos and texts, its effective training strategy can also result in significant improvements on all benchmarks, regardless of whether the texts are detailed or not.

As in Tab. 4, VideoCLIP-XL also surpasses other competitors significantly on Shot2Story under the long description setting. In Shot2Story, each video clip consists of multiple video shots which switch between different scenes to express the same main event. This requires the model to have the ability to fully understand mainline activity from multiple complex scenarios. Performances demonstrate that our method exhibits significant advantages whether

Method	LVDR Benchmark				
	4×1	4×2	4×3	4×4	4×5
CLIP (Radford et al., 2021)	30.12/27.83/18.76	47.97/38.46/31.61	57.00/43.61/39.31	65.31/50.53/48.78	69.58/53.17/53.31
ViCLIP (Wang et al., 2023c)	18.95/24.93/18.40	33.63/41.57/32.43	44.47/52.06/43.10	52.43/58.81/50.46	58.61/62.10/55.55
Long-CLIP (Zhang et al., 2024)	70.07/49.67/51.11	81.66/65.45/68.78	86.49/73.91/77.54	89.84/79.90/83.32	91.70/83.60/86.50
VideoCLIP-XL (Ours)	**80.32/70.93/72.06**	**90.72/83.70/86.11**	**94.41/89.87/91.92**	**95.87/91.93/93.49**	**96.99/94.17/95.47**

Table 3: Ranking score (RS)/Kendall's tau (KT)/Spearman rank-order correlation coefficient (SC) of long video description ranking on the proposed LVDR benchmark.

Method	Shot2Story-W		Shot2Story-S	
	T2V	V2T	T2V	V2T
CLIP (Radford et al., 2021)	65.80	66.00	45.40	45.35
ViCLIP (Wang et al., 2023c)	37.53	37.71	48.44	46.17
Long-CLIP (Zhang et al., 2024)	74.74	80.59	47.70	43.14
VideoCLIP-XL (Ours)	**95.28**	**94.73**	**70.30**	**67.79**

Table 4: R@1 of text-video retrieval on Shot2Story (Han et al., 2023) with long video descriptions.

using the whole video clip (Shot2Story-W) or each shot (Shot2Story-S) as an individual for the text-video retrieval task.

The results on our LVDR benchmark are shown in Tab. 3. VideoCLIP-XL has a stronger identification ability compared with competitors to perceive inaccurate content in long video descriptions and assign them lower similarity scores. For example, under the 4×1 setting where *only 1 original word is randomly replaced with a wrong one* between adjacent generated descriptions, our model can surpass Long-CLIP (which focuses on long text understanding for images) with +10.25 ranking score. We can also observe that as the level of single-step hallucination increases from shallow to deep (4×1 to 4×5), the video CLIP models can naturally distinguish different long video descriptions better.

4.3 Ablation Study

In this subsection, we aim to explore the effectiveness of each component in our method.

As shown in Fig. 1, our VILD pre-training dataset is formed by the aggregation of four parts from different data sources. For parts (a)(b)(c), the data resource often utilizes the powerful GPT-4V (Achiam et al., 2023) or human efforts to generate the text information before our LLM-based steps. While for part (d), we use open-source LLMs for generating long descriptions from raw videos. The results in Tab. 5(a) show the data effectiveness. Although the effect of using open-source LLMs for automated data synthesis can naturally lag behind GPT-4V/human efforts by a margin, it can still achieve state-of-the-art performance compared with existing competitors. In addition, adding (d) on top of (a)(b)(c) can further result in obvious improvements. This also demonstrates the effectiveness of our data synthesis pipeline.

#	Pre-Training Data	MLDMA (R@1)	S2S (R@1)	LVDR (RS)
1	Part (a)(b)(c) of VILD	45.52	80.37	90.97
2	Part (d) of VILD	44.54	78.77	89.23
3	Full VILD	**46.61**	**82.03**	**91.67**

(a)

#	TPCM	DDR	HDR	MLDMA (R@1)	S2S (R@1)	LVDR (RS)
1	Baseline (Zhang et al., 2024)			45.62	81.47	84.87
2	✓			46.06	**82.03**	84.87
3	✓	✓		46.58	**82.03**	86.07
4	✓		✓	46.07	**82.03**	91.42
5	✓	✓	✓	**46.61**	**82.03**	**91.67**

(b)

Table 5: Ablation study for components of our method. MLDMA indicates the averaged zero-shot text-video retrieval R@1 score of benchmarks in Tab. 1. S2S is short for Shot2Story.

As shown in Tab. 5(b) #2 v.s. #1, TPCM can achieve +0.44 R@1 gain on traditional text-video retrieval datasets and +0.56 R@1 gain on Shot2Story. In addition, it can dynamically modify the feature space distribution during pre-training, which is reflected in the increase of PCA dimension, as shown in Fig. 5.

The effectiveness of DDR and HDR can also be found in Tab. 5(b). Compared #3 with #2, DDR achieves a +0.52 R@1 gain on traditional benchmarks and +1.20 RS gain on LVDR. As for HDR, compared #4 with #2, it obtains +6.55 RS gain on LVDR. Furthermore, conducting both tasks together is more effective than using either one alone on MLDMA and LVDR, as shown in #5 v.s. #2.

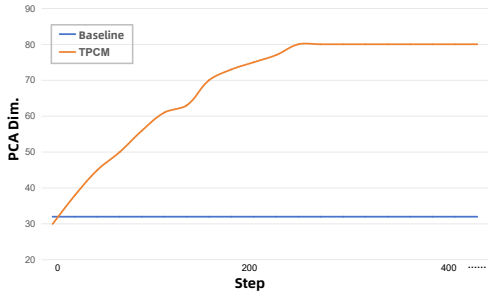


Figure 5: TPCM can dynamically adjust the dimension of the attribute vectors that need to be retained during pre-training. Dim. is short for dimension.

5 Conclusion

In this paper, we propose **VideoCLIP-XL**, a video CLIP model with long-description capability. We establish an automatic data collection system to gather our VILD dataset, and propose TPCM to better learn the distribution of feature space during pre-training while expanding the long-description capability. We also introduce two new tasks namely DDR and HDR for further understanding improvement. Our LVDR benchmark is helpful for evaluating the long-description capability more comprehensively. Extensive experimental results have demonstrated the effectiveness of our method.

For future research, we plan to refine the pre-training methodology and increase the amount of data and model size for further improvement. We will also attempt to integrate the architecture of cross-encoders and LLMs into our method.

Limitations

Although VideoCLIP-XL is trained to enable long-description understanding capacity, limited by the amount of pre-training data and the feature extraction capability of single-modal encoders, there is still room for improvement. The scale, quality, and variety of data can be further extended, and the model structure and model size of feature extractors can also be scaled up. The application of our method in the structures of cross-encoders and LLMs is also worth exploring. These improvements are left to our subsequent work.

Ethical Considerations

The techniques for training the VideoCLIP-XL model presented in this work are fully methodological, thereby there are no direct negative social impacts of our method. Additionally, we have filtered out NSFW examples from our pre-training data to ensure that the seen contents are suitable for public distribution.

Acknowledgements

This research is supported in part by National Natural Science Foundation of China (Grant No.: 62441604, 62476093). It is also partially supported by Alibaba Cloud Computing, through Research Talent Program with South China University of Technology.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL-HLT*, pages 190–200.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. ShareGPT4V: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. 2024. Panda-70M: Captioning 70M videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. VQGAN-CLIP: Open domain image generation and editing with natural language guidance. In *ECCV*, pages 88–105.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

Kevin Frans, Lisa Soros, and Olaf Witkowski. 2022. CLIPDraw: Exploring text-to-drawing synthesis through language-image encoders. *NeurIPS*, 35:5207–5218.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*.

Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. 2023. Shot2Story20K: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2311.17043*.

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2023. Tag2Text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE TPAMI*, 33(1):117–128.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024. Miradata: A large-scale video dataset with long durations and structured captions. https://github.com/mira-space/MiraData.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. 2023. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *CVPR*, pages 11144–11154.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. 2022a. Language-driven semantic segmentation. In *ICLR*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202, pages 19730–19742.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023c. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022b. Grounded language-image pre-training. In *CVPR*, pages 10965–10975.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*, volume 8693, pages 740–755.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:1807.03748*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Ziyang Luo, Pu Zhao, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. LexLIP: Lexicon-bottlenecked language-image pre-training for large-scale image-text sparse retrieval. In *ICCV*, pages 11172–11183.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640.

Marius Muja and David G Lowe. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP*, 2(331-340):2.

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

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. In *NeruIPS*, pages 1143–1151.

PKU-YuanGroup. 2024. Open-sora-dataset. https://github.com/PKU-YuanGroup/Open-Sora-Dataset.

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. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *CVPR*, pages 3202–3212.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2023. Alpha-CLIP: A CLIP model focusing on wherever you want. *arXiv preprint arXiv:2312.03818*.

Moming Tang, Chengyu Wang, Jianing Wang, Chuanqi Tan, Songfang Huang, Cen Chen, and Weining Qian. 2023a. XtremeCLIP: Extremely parameter-efficient tuning for low-resource vision language understanding. In *ACL (Findings)*, pages 6368–6376.

Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. 2023b. When are lemons purple? the concept association bias of vision-language models. In *EMNLP*, pages 14333–14348.

Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. 2022. Clipasso: Semantically-aware object sketching. *TOG*, 41(4):1–11.

Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. 2023. Connecting vision and language with video localized narratives. In *CVPR*, pages 2461–2471.

Jiapeng Wang, Chengyu Wang, Xiaodan Wang, Jun Huang, and Lianwen Jin. 2023a. CocaCLIP: Exploring distillation of fully-connected knowledge interaction graph for lightweight text-image retrieval. In *ACL*, pages 71–80.

Xiaodan Wang, Chengyu Wang, Lei Li, Zhixu Li, Ben Chen, Linbo Jin, Jun Huang, Yanghua Xiao, and Ming Gao. 2023b. FashionKLIP: Enhancing e-commerce image-text retrieval with fashion multi-modal conceptual knowledge graph. In *ACL*, pages 149–158.

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

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023c. InternVid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. 2024. InternVideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, pages 6787–6800.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. 2022. GroupViT: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18113–18123.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, pages 5036–5045.

Keunwoo Peter Yu, Zheyuan Zhang, Fengyuan Hu, and Joyce Chai. 2023. Efficient in-context learning in vision-language models for egocentric videos. *arXiv preprint arXiv:2311.17041*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-CLIP: Unlocking the long-text capability of CLIP. *arXiv preprint arXiv:2403.15378*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI*, volume 32.

A Appendix

A.1 Details of VILD Data Generation

During the data generation for VILD, we leverage Qwen1.5-72B-Chat (Bai et al., 2023) in LLM-based steps and LLaVA-v1.6-34B (Liu et al., 2024b) in LMM-based steps. All prompts we used are listed as follows:

[*Desc. Aggregation*]

"The following are descriptions of the subjects or background in a video. Please organize them together into a single description of the entire video. Do not omit any content, nor add any new content that is not included or uncertain.

{examples}

Descriptions: {individual-level descriptions}

Output:"

[*Desc. Rewrite*]

"The following is a video description. Please output a rewritten version. Do not omit any content, nor add any new content that is not included or uncertain.

{examples}

Description: {input description}

Output:"

[*Data Filtering*]

"Determine if the following conversation is talking about the overall/comprehensive-level description/content of a video. If yes, output Yes; otherwise, output No.

{examples}

Conversation: {input conversation}

Output:"

[*Long Frame Desc. Generation*]

"Precisely describe this image."

[*Long Video Desc. Generation*]

"We will provide a description of a video and some frame descriptions of it. Directly output an enriched video description according to them. Remove repetitive contents. Do not describe any content that is uncertain or not included. Do not describe individual frames. Do not describe specific subjects, use generic words instead.

{examples}

Video Description: {short video description}

Frame Descriptions: {long frame descriptions}

Output:"

A.2 Details of Data Statistics

More detailed comparisons of data statistics information are shown in Tab. 6.

A.3 Details of Experimental Settings

We sample 8 frames for each video during pre-training. Stretching of the vanilla absolute positional embedding from 77 to 248 is also applied following (Zhang et al., 2024). During pre-training, we set the batch size 1664, warm-up steps 200, weight decay 0.02, and max learning rate 4e-6. The learning rate decreases in a cosine schedule after warm-up. α_1, α_2, α_3, α_D, and α_H are empirically set as 0.1, 1.0, 10.0, 0.0, and 0.0 respectively. m in the DDR and HDR tasks is set as 5.

During pre-training, as shown in Eq. 8, we use both long descriptions to enable VideoCLIP-XL to learn the semantics of long texts, and short descriptions to maintain the original short text ability. For videos in our VILD dataset that do not have paired short descriptions from the origin resource, we use Qwen1.5-72B-Chat to generate them based on long descriptions. The prompt we used is:

"The following is a detailed video description. Please extract its core content and summarize it into a really short sentence. Do not exceed 10 words.

{examples}

Description: {long video description}

Dataset	Year	Caption Source	Domain	Video Num.	Avg. Video Len.	Avg. Text Len.
HowTo100M (Miech et al., 2019)	2019	ASR	Open	136M	3.6s	4.0 words
HD-VILA-100M (Xue et al., 2022)	2022	ASR	Open	103M	13.4s	32.5 words
MSVD (Chen and Dolan, 2011)	2011	Manual	Open	1970	9.7s	8.7 words
LSMDC (Rohrbach et al., 2015)	2015	Manual	Movie	118K	4.8s	7.0 words
MSR-VTT (Xu et al., 2016)	2016	Manual	Open	10K	15.0s	9.3 words
DiDeMo (Anne Hendricks et al., 2017)	2017	Manual	Flickr	27K	6.9s	8.0 words
ActivityNet (Heilbron et al., 2015)	2017	Manual	Action	100K	36.0s	13.5 words
YouCook2 (Zhou et al., 2018)	2018	Manual	Cooking	14K	19.6s	8.8 words
VATEX (Wang et al., 2019)	2019	Manual	Open	41K	10.0s	15.2 words
Panda-70M (Chen et al., 2024)	2024	Automatic	Open	70.8M	8.5s	13.2 words
VILD (Ours)	2024	Automatic	Open	2.1M	15.4s	74.2 words
LVDR (Ours)	2024	Automatic+Manual	Open	2K	17.5s	230.7 words

Table 6: Comparison of data statistics information.

A woman in a red shirt, with wavy blonde hair, is standing in a garden, pointing at a pink-flowered plant. She appears engaged in explaining or demonstrating gardening techniques, surrounded by greenery and trees, creating an atmosphere of nature and learning.

A white wreath adorned with pine cones hangs against a wooden wall, showcasing a blend of dried plant materials like glossy green pine branches. The semi-circular arrangement features neatly organized branches and varied-sized pine cones, exhibiting natural shades of green and brown. The rustic wooden wall with horizontal plank design adds a complementary backdrop, evoking a seasonal or cozy atmosphere.

Figure 6: Examples of synthetic long video captions in our VILD dataset.

Output:"

For fine-tuned setting of text-video retrieval on traditional benchmarks, we tune our pre-trained VideoCLIP-XL with the vanilla text-video contrastive learning loss on each training set of the evaluated benchmarks. During both training and testing, we sample 12 frames. Detailed hyper-parameters are the same as ViCLIP (Wang et al., 2023c). While in the zero-shot setting, along with the evaluations for Shot2Story and LVDR, we sample only 8 frames.

For the image CLIP models such as Long-CLIP, we calculate the similarity between the averaged image feature of frames and the text feature.

A.4 Performance Comparison with More Models

As shown in Tab. 7, we involve more recent powerful and large *cross-encoder* models (Li et al., 2023c; Wang et al., 2024) for comprehensive comparisons. *Cross-encoder* models, especially large multi-modal models (LMMs), typically add additional Transformer layers to model the deep interaction between vision and text representations. The model can generally boost the retrieval performance, while resulting in an unbearably slow retrieval speed when applied to the entire image/video collection since the cross-modal costs are required for each image/video sample whenever a new text query is given. In contrast, our VideoCLIP-XL which has the *dual-encoder* structure has obviously fewer parameters and retrieval time cost. *Dual-encoder* encodes the visual and textual inputs in a wholly decoupled manner. The vision representation is allowed to be pre-computed and re-used independent of the text queries. Such approaches can utilize fast approximate nearest neighbor (ANN) search (Muja and Lowe, 2009; Jegou et al., 2010; Johnson et al., 2019) at runtime to ensure high efficiency. For example, VideoCLIP-XL generally surpasses UMT-L (Li et al., 2023c) on zero-shot text-video retrieval and has $\sim$4.14$\times$ faster retrieval speed on MSR-VTT without any bells and whistles, which can also indicate the effectiveness of our pre-training stage. It is also $\sim$8.69$\times$ faster than InternVideo2$_{s2}$-1B. For fine-tuning, large *cross-encoder* models naturally surpass *dual-encoder* models owing to the cross-modal feature interaction. Yet, these models still suffer from the low inference speed issue, and hence can hardly be deployed in real-time applications.

Method	MSR-VTT		LSMDC		DiDeMo		MSVD		ActivityNet		Avg.	
	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
Dual-Encoder												
CLIP (Radford et al., 2021)	30.4	24.2	13.9	11.9	12.7	18.7	40.5	57.2	9.1	13.2	21.3	25.0
VideoCLIP (Xu et al., 2021)	10.4	-	-	-	16.6	-	-	-	-	-	-	-
CLIP4Clip (Luo et al., 2022)	32.0	-	15.1	-	-	-	38.5	-	-	-	-	-
ViCLIP (Wang et al., 2023c)	42.4	41.3	20.1	16.9	38.7	39.1	49.1	75.1	32.1	31.4	36.5	40.8
VideoCLIP-XL (Ours) [V:304M/T:124M/C:0M/47.53s]	**50.1**	**49.9**	**22.8**	**24.6**	**47.7**	**47.9**	**51.9**	**76.7**	**46.4**‡	**48.1**‡	**43.8**	**49.5**
Cross-Encoder												
UMT-L (Li et al., 2023c) **[V:304M/T:271M/C:84M/196.68s]**	40.7	37.1	24.9	21.9	48.6	49.9	49.0	74.5	41.9	39.4	41.0	44.6
InternVideo2$_{s2}$-1B (Wang et al., 2024) **[V:1049M/T:271M/C:88M/413.09s]**	51.9	50.9	32.0	27.3	57.0	54.3	58.1	83.3	60.4	54.8	51.9	54.1
InternVideo2$_{s2}$-6B (Wang et al., 2024) **[NOT publicly available]**	55.9	53.7	33.8	30.1	57.9	57.1	59.3	83.1	63.2	56.5	54.0	56.1

(a)

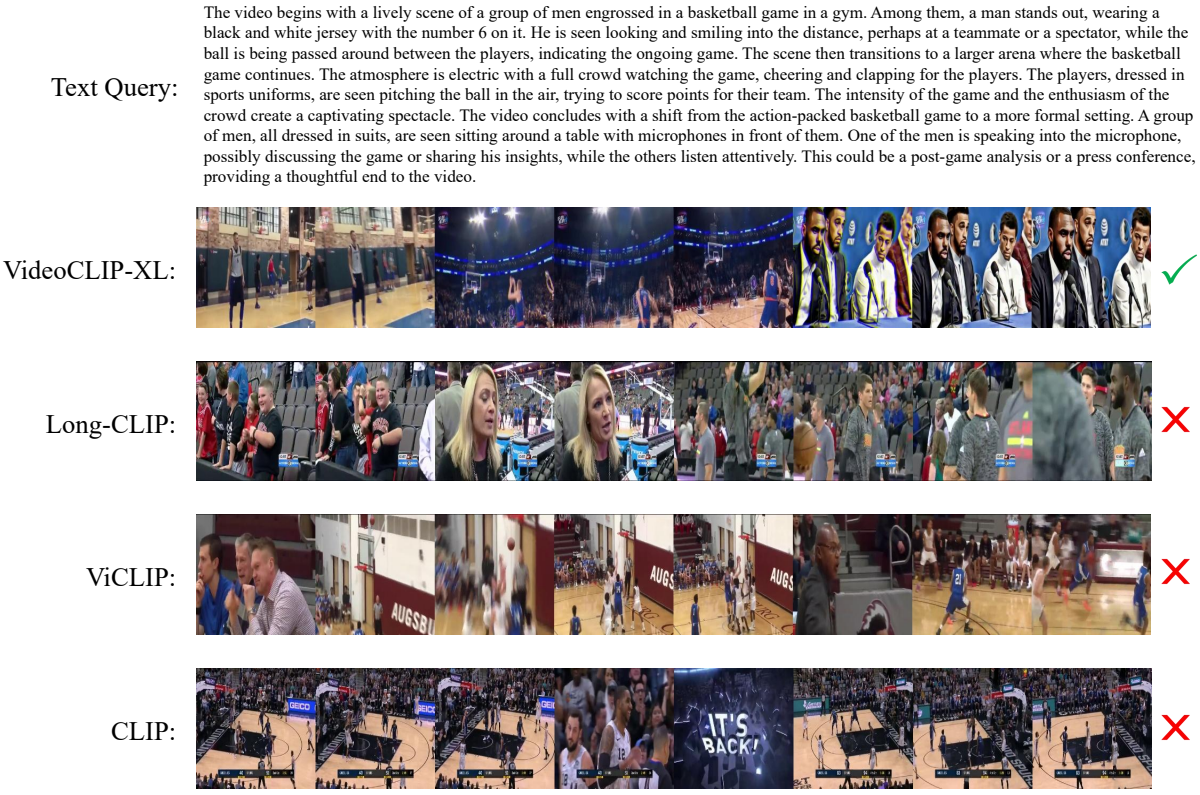
Method	MSR-VTT		LSMDC		DiDeMo		MSVD		ActivityNet		Avg.	
	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T	T2V	V2T
Dual-Encoder												
CLIP (Radford et al., 2021)	38.2	38.7	22.5	22.6	32.2	33.9	52.3	69.9	26.1	26.9	34.3	38.4
VideoCLIP (Xu et al., 2021)	30.9	-	-	-	-	-	-	-	-	-	-	-
CLIP4Clip (Luo et al., 2022)	45.6	45.9	24.3	23.8	43.0	43.6	45.2	48.4	40.3	41.6	39.7	40.7
ViCLIP (Wang et al., 2023c)	52.5	51.8	33.0	32.5	49.4	50.2	53.1	79.0	49.8	48.1	47.6	52.3
VideoCLIP-XL (Ours)	**57.0**	**56.6**	**34.2**	**32.6**	**62.3**	**62.7**	**55.6**	**81.4**	**58.4**‡	**59.2**‡	**53.5**	**58.5**
Cross-Encoder												
UMT-L (Li et al., 2023c)	58.8	58.6	43.0	41.4	70.4	65.7	58.2	82.4	66.8	64.4	59.4	62.5
InternVideo2$_{s2}$-6B (Wang et al., 2024)	62.8	60.2	46.4	46.7	74.2	71.9	61.4	85.2	74.1	69.7	63.8	66.7

(b)

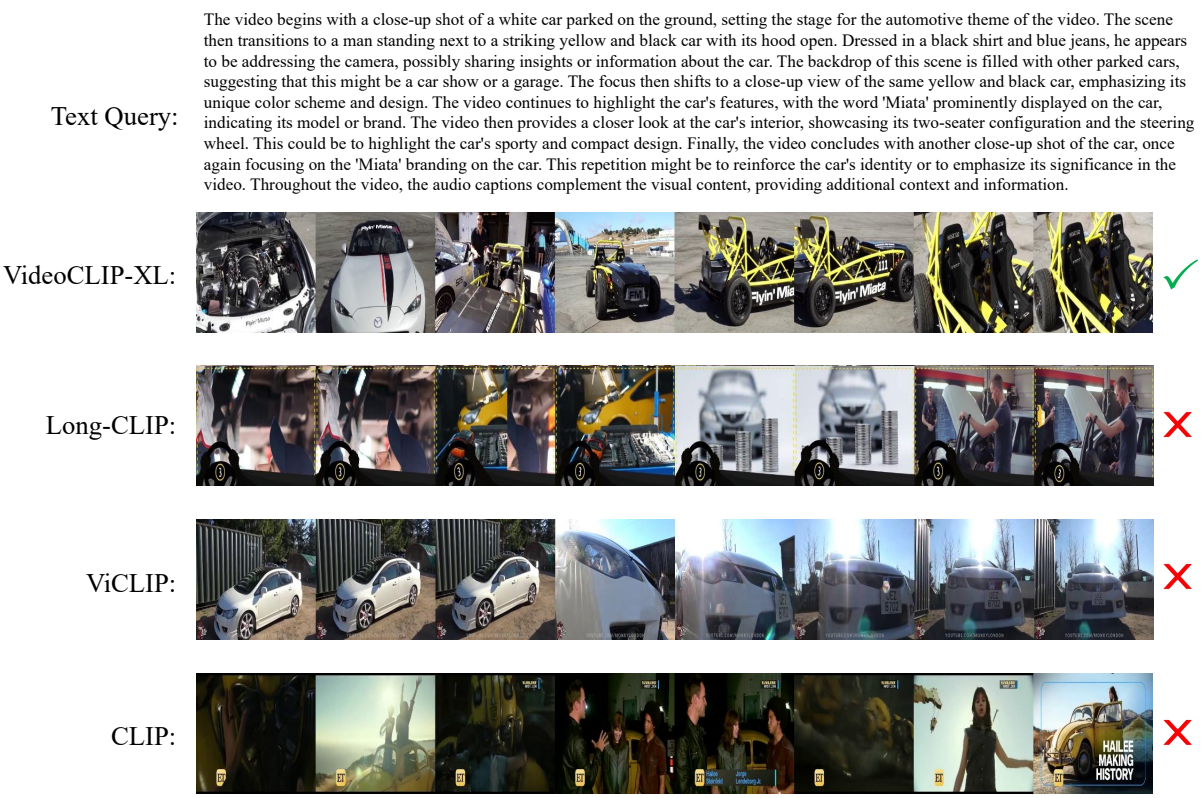
Table 7: R@1 scores of (a) zero-shot and (b) fine-tuned text-video retrieval on MSR-VTT (Xu et al., 2016), LSMDC (Rohrbach et al., 2015), DiDeMo (Anne Hendricks et al., 2017), MSVD (Chen and Dolan, 2011), and ActivityNet (Heilbron et al., 2015). ‡Due to the high overlap between the videos in ActivityNet and VideoInstruct100K (Maaz et al., 2023), the latter is excluded from the pre-training data of our model tested on the former. **[V:304M/T:124M/C:0M/47.53s]** indicates that the vision encoder has 304M parameters, the text encoder has 124M parameters, the cross-encoder has 0M parameters, and this model needs 47.53s for text-video retrieval on MSR-VTT test set (1000 text-video pairs, tested on a single A100-80G GPU). The same goes for others.

A.5 More Qualitative Results

We give some examples of our synthetic long video descriptions acquired by Fig. 1(d) in Fig. 6. And qualitative examples of text-to-video retrieval results on the Shot2Story benchmark are shown in Fig. 7. We can find that compared to competitors, our VideoCLIP-XL can achieve more accurate and matching video retrieval results.



(a)



(b)

Figure 7: Qualitative examples of text-to-video retrieval on the Shot2Story benchmark.