Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large-scale video-language pre-training has shown significant improvement on video-language understanding tasks. Previous studies of video-language pre-training mainly focus on short-form videos (i.e., within 30 seconds) and sentences, leaving long-form video-language pre-training rarely explored. Directly learning representation from long-form videos and language is challenging due to the difficulty of modeling long-range relationship and heavy computation burden caused by more frames. In this paper, we introduce a Long-Form Video-Language Pre-training (LF-VLP) model, and train it on a large-scale long-form video and paragraph dataset constructed from an existing public dataset. To effectively capture the rich temporal dynamics and to better align video and language in an efficient end-to-end manner, we introduce two novel designs in our LF-VLP model. We first propose a multimodal temporal contrastive loss (MTC) to learn the temporal relation across different modalities by encouraging fine-grained alignment between long-form videos and paragraphs. Second, we propose a hierarchical temporal window attention (HTWA) mechanism to effectively capture long-range dependency while reducing computational cost in Transformer. We fine-tune the pre-trained LF-VLP model on seven downstream long-form video-language understanding tasks of paragraph-to-video retrieval or long-form video QA, and achieve the new state-of-the-art performances. Specifically, our model achieves 16.1% relative improvement on ActivityNet paragraph-to-video retrieval task and 2.4% on How2QA task, respectively.

1 Introduction

In recent years, research on video understanding has attracted extensive attention due to the huge amount of videos available everywhere in our daily life. Previous research works on video understanding [13, 14, 36, 38, 49] mainly focus on short-form video (i.e., $<$ 30 seconds) analysis and the semantics are limited to certain types (e.g., actions, scenes) using one-hot vectors as supervision. However, videos are usually long-form (i.e., $>$ 30 seconds) [41] in real scenarios. Human annotated labels for video understanding (e.g., actions) are difficult to cover the rich semantic and dynamic information contained in those videos. On the other hand, video-language pre-training paradigm provides a way to learn cross-modal representation from video and language pairs [3, 22, 43, 47]. While long-form video and language pre-training has not been well studied yet. In this paper, we explore to directly exploit long-form video and language pairs for pre-training.

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

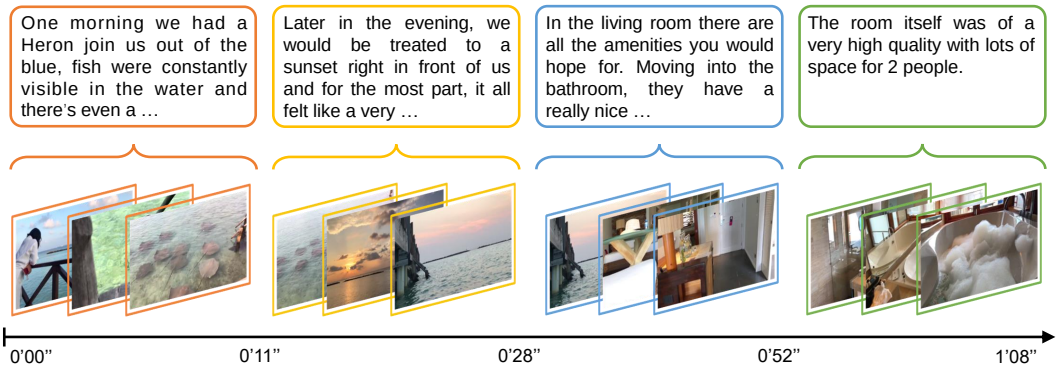

Figure 1: An example of long-form video-paragraph pair with several clips and sentences. It contains complicated story line and rich temporal dynamic. Each sentence can only describe a short clip, and understanding the whole video needs the ability of long-range spacial temporal reasoning.

Although long-form video-language joint learning has been explored in downstream tasks [15, 18, 21, 48, 50], they either use pre-extracted video features which lead to the sub-optimal understanding, or utilize image encoder to extract frame features which fail to model the long-range dependency in long-form videos. Recent works [3, 5, 27] have shown that a video Transformer backbone helps to capture long-range dependency in an end-to-end fashion. An intuitive way for long-form video-language pre-training is to adopt a video-Transformer based short-form video-language pre-training model with long-form data. However, there are two main challenges in such a design. Firstly, long-form videos contain more complicated story lines and richer temporal dynamics as shown in Fig. 1. Simply aligning video and paragraph as a pre-training task like previous models [3, 43] will ignore the temporal relation between clips and sentences, thus hinder the quality of learned representation. Secondly, to feed more frames in Transformer-based video encoder will largely increase computational cost considering the self-attention operation.

To overcome the above challenges, we propose a Long-Form Video-Language Pre-training model (LF-VLP) with two novel designs. First, to better align long-form video-language pairs and learn temporal relationship between visual and language modalities, we propose a multimodal temporal contrastive loss (MTC) that learns temporal alignment between video clips and single sentences. MTC encourages similarity between two modalities to be consistent with their temporal relationship. In other words, the embedding distance between a video clip and a sentence closer in time should be smaller than its distance with sentences that are far in time. Combining with global alignment between video and paragraph, MTC ensures the model to capture temporal relation between video clips and single sentences, and further helps to improve the quality of joint representation.

Second, to utilize the advantage of Transformer for capturing long-range dependency while efficiently processing more frames for end-to-end training, we propose a hierarchical temporal window attention (HTWA) mechanism. As shown in Fig. 1, the frames sparsely sampled from a long-form video have large spatial and motion gap, thus directly computing self-attention on all frames in all layers of Transformer is inefficient and unnecessary. Instead, we only learn the attention between adjacent frames in the first few layers that focus more on details of spatial and temporal information. Then we gradually expanding the window size in the following layers, where the higher level representation enables the model to better capture relation between frames far apart. The computational cost is largely reduced with the hierarchical temporal window attention mechanism.

We conduct experiments and evaluate LF-VLP on seven downstream long-form video-language understanding tasks of paragraph-video retrieval and long-form video question answering. We surpass the state-of-the-art models pre-trained on short videos by a large margin. Our results demonstrate the benefit of modeling long-range dependency for long-form videos. We also verify the effectiveness

of our proposed multimodal temporal contrastive loss and hierarchical temporal window attention mechanism through ablation studies.

Our contributions are summarized as follows: (1) We are the first to study end-to-end long-form video-language pre-training with large-scale video-paragraph data. We propose a multimodal temporal contrastive loss to capture the temporal relationship between clips and sentences while improving the joint representation of long-form video and language. (2) We design a temporal window attention mechanism for video Transformer backbone, which can learn long-form video representation effectively and efficiently. (3) We verify the effectiveness of our model on a wide range of downstream long-form video-language tasks. Our model achieves the state-of-the-art performance on four paragraph-video retrieval and two long-form video question answering tasks.

2 Related Work

2.1 Video Representation

Most previous video encoders utilize 3D-CNN based backbones [7, 38, 42]. These models show promising performance on short-form video understanding tasks, such as action classification and detection [7, 6, 16]. However, CNN has limited receptive field and cannot effectively capture long-range dependency. Recent works have extended Vision Transformer [12] for video representation and demonstrate the benefit of long-range temporal learning [5, 27]. To reduce the computational cost, TimeSformer [5] introduces a factorized spacetime attention, while Video Swin-Transformer [26] restricts self-attention in a local 3D window. However, TimeSformer [5] is still computationally expensive when the number of input frames becomes large. Video Swin-Transformer [27] adopts a fix-sized temporal window which is not suitable for videos with large time spans. We propose a hierarchical temporal window attention to effectively learn the long-range dependency in long-form videos while reduce the computational cost.

2.2 Long-form Video Understanding

Long-form video understanding is less explored in previous studies. Some works explore using long-term context for improving recognition performance [35, 40]. Typical long-form video understanding tasks contain shot or event boundary detection [4] and temporal action detection [6], but these tasks cannot reveal ability of high level understanding of the model. Jointly understanding long-form videos with language is a way to discovering the rich semantics contained in videos and many benchmarks are proposed recently, such as paragraph-video retrieval [1, 2, 19, 31] and long-form video question answering [22, 24, 46]. Previous works that explore these tasks mostly use pre-extracted features, which hinder the performance because of sub-optimal features [15, 18, 48, 50]. We study end-to-end long-form video-language pre-training and transfer to long-form video-language understanding tasks.

2.3 Video-Language Pretraining

Inspired by the success of image-language pre-training [9, 17, 33], video-language pre-training is also explored recently; however, these works mainly focus on short-form videos [3, 28, 30, 43]. Some works use 3D-CNN as video backbone [28, 30]. To utilize the advancement of Transformer, some works use sparsely sampled frames to reduce the computation requirements [3, 43]. One key factor for learning good representation is using contrastive loss to align multi-modal features [3, 28, 30, 43, 47]. We further design a multimodal temporal contrastive loss to conduct fine-grained alignment between long-form videos and paragraph. The power of the pre-training model is largely dependent on the amount of training data, some works built large-scale video-language datasets [30, 43, 47], we build a long-from video-paragraph dataset based on [43]. There are several works have explored long-form video-language pre-training, HERO [22] uses pre-extracted features, while MELORT [47] uses an image encoder to separately encode frames which ignores joint spacial temporal representation. Different from them, we use a video Transformer backbone and end-to-end pre-training on large-scale long-form video-paragraph dataset.

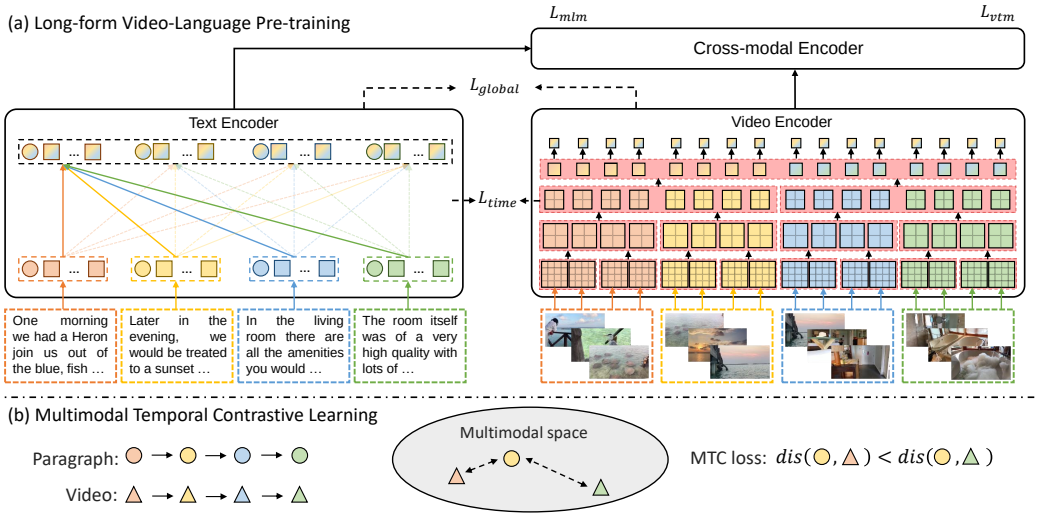

Figure 2: The framework of (a) long-form video-language pre-training (LF-VLP) and illustration of (b) multimodal temporal contrastive (MTC) learning. (a) LF-VLP consists of a text encoder, a video encoder and a cross-modal encoder. In the text encoder, attention computing is first within each sentence then the whole paragraph. The pink boxes in video encoder illustrate the proposed hierarchical temporal window attention (HTWA) mechanism. (b) the MTC loss aligns two sequence of representations (e.g., clip and sentence representations in our case), the distance of two element's representation is smaller when they are closer in time.

3 Approach

In this section, we first show the overall architecture of the proposed Long-Form Video-Language Pre-training (LF-VLP) model in Sec. 3.1. Then we explain our proposed Multimodal Temporal Contrastive loss (MTC) for learning cross-modal temporal relationship in Sec. 3.2, followed by the designed Hierarchical Temporal Window Attention (HTWA) mechanism for efficient video encoder in Sec. 3.3. Finally, we introduce the pre-training pipeline with target pre-training tasks in Sec. 3.4.

3.1 Model Architecture

As illustrated in Fig. 2, our proposed long-form video-language pre-training model (LF-VLP) consists of three parts: a video encoder E_V, a text encoder E_T and a cross-modal encoder E_C. With a video and a paragraph as input, we first pass them to video encoder E_V and text encoder E_T for embedding learning, respectively. Then we concatenate visual and language embeddings as input to E_C, where further cross-modal joint learning is conducted. The details of these encoders are as follows.

Text Encoder. The text encoder E_T is based on Transformer network [39]. We divide it into two parts for sentence-level and paragraph-level encoding. In the first several layers, self-attention is conducted within word tokens from the same sentence, and sentence embedding can be learned individually. In higher layers, we add segment embedding to distinguish each sentence and the attention computation is extended to all word tokens of the paragraph to output paragraph representation.

Video Encoder. Our video encoder is also stacked by Transformer layers. In particular, we design a hierarchical temporal window attention (HTWA) mechanism for efficient attention computation. Given a long-form video which has M clips, we sample N frames from each clip and divide each raw frame into $H \times W$ patches. Then the $M \times N \times H \times W$ patches are encoded by E_V. With our designed HTWA mechanism, the temporal window is gradually expanded, so that we can get hierarchical feature maps with different temporal receptive fields. In addition, video features with the

same temporal window size as clip frame number N in the middle layer can be utilized as the clip representation for fine-grained alignment with sentences.

Cross-modal Encoder. The cross-modal encoder E_C adopts self-attention to learn the joint relation between visual and language modalities. Visual and language embeddings from the output of E_V and E_T are concatenated as the input to E_C. Three pre-training tasks are used for the optimization of joint learning. In particular, we propose a novel multimodal temporal contrastive loss to enable learning of temporal relationship between different modalities in temporal dimension.

3.2 Multimodal Temporal Contrastive Learning

Contrastive learning is widely used in previous multimodal pre-training works to align different modalities such as image-language and video-language. The goal of this loss is to pull the representation of matched pairs close to each other and push unmatched pairs away from each other. However, when aligning long-form videos and paragraphs, the vanilla contrastive loss neglects the temporal relation between clips and sentences, which is important for capturing the complex temporal dynamics in long-form videos. To better learn the temporal relationship between different modalities, we propose a multimodal temporal contrastive loss, which can be applied to align sequences in different modalities such as video and paragraph. We assume that the distance of two elements' representation in different modalities should be consistent with their temporal distance. Specifically in our model, MTC encourages the video clip embedding to be more similar to its neighbor sentence embedding than sentences of long distance in the same paragraph. For example in Fig. 2, given two sequence of representation $v_i = \{v_{i,1}, v_{i,2}, ..., v_{i,M}\}$ and $t_i = \{t_{i,1}, t_{i,2}, ..., t_{i,M}\}$ from the i-th sample, we first sample an anchor set $\mathcal{K}$ of k representations from v_i, then we sample a set of representations $\mathcal{V}$ from t_i. For each $v_{i,p}$ in $\mathcal{K}$, we treat t_{i,q^+} as positive, where $|p - q^+| = \min(|p - q|), t_{i,q} \in \mathcal{V}$. We also randomly sample some representations from $t_j, j \neq i$ as $\mathcal{N}$ which is used to stable the training. Then the MTC loss is calculated by applying an InfoNCE loss:

$$\mathcal{L}^p_{MTC}(v_i, t_i) = -\log \frac{\exp\left(s(v_{i,p}, t_{i,q^+})\right)}{\sum_{t_{i,q} \in \mathcal{V}} \exp\left(s(v_{i,p}, t_{i,q})\right) + \sum_{t_j \in \mathcal{N}} \exp\left(s(v_{i,p}, t_j)\right)}, \tag{1}$$

$$\mathcal{L}_{MTC}(v_i, t_i) = \frac{1}{k} \sum_{t_{i,p} \in \mathcal{K}} \mathcal{L}^{i,p}_{MTC}, \tag{2}$$

where $s(f_1, f_2) = f_1^T \cdot f_2/\tau$, τ is the temperature.

We obtain clip representations of v_i and sentence representations t_i from the output of the first part of video encoder and text encoder, respectively. Then the MTC loss can be obtained by:

$$\mathcal{L}^{v2t}_{time} = -\frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_{MTC}(v_i, t_i),\ \mathcal{L}^{t2v}_{time} = -\frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_{MTC}(t_i, v_i). \tag{3}$$

The temporal contrastive loss $\mathcal{L}_{time}$ is the average of $\mathcal{L}^{v2t}_{time}$ and $\mathcal{L}^{t2v}_{time}$.

3.3 Hierarchical Temporal Window Attention

Directly feeding all sampled frames of a long-form video to vanilla Transformer network for global self-attention learning will heavily increase the computational cost. The cost is quadratic with respect to the number of frames. One possible way is to apply a small fixed temporal window. However, such a design will neglect to learn relationship between different clips of one video, which is essential for long-form video understanding. To capture long-range dependency in long-form videos efficiently and effectively, we propose a Hierarchical Temporal Window Attention (HTWA) mechanism by gradually increasing temporal window size in Transformer layers.

A temporal window is used to restrict attention computing between frames in temporal dimension. Given a 3-D input tensor of $T' \times H' \times W'$ patches, where T' is the number of time steps and $H' \times W'$

denotes the number of spatial patches. We divide the input tensor to M_l windows along T' dimension, where l denotes the Transformer layer. Then multi-head attention is applied within each window, and the output is the combination of all attention windows as follows:

$$\begin{aligned} a_i &= MHA\left(z_i\right), i \in [1, 2, ..., M_l], \\ a &= Concat\left(a_1, a_2, ..., a_{M_l}\right), \end{aligned} \tag{4}$$

where z_i is the token embeddings belong to i-th window, MHA is multi-head attention, a_i is the attention weights of i-th window, and a is the attention weights with the shape of $T' \times H' \times W'$.

Compared with short-form videos, long-form videos have two distinguished characteristics that we should consider for better understanding. First, there are large motion gaps (dynamics) between frames that are sparsely sampled. Second, it is important to build long-range relationship to understand the whole video. Considering the above properties, we start to build connections between adjacent frames within small window size (e.g., 2) in the first few layers. The temporal window size is gradually increased to capture longer-range dependency as the semantics learned become high-level. Specifically, we use the temporal window size equal to the number of frames in a video in the last several layers, so that the full context can be attended to learn the global relationship.

3.4 Pre-training Pipeline

We adopt a two-stage pre-training as previous works did [43] since modal-independent design enables to provide powerful single-modality embedding for downstream tasks. In the first stage, video encoder and text encoder are learned independently with a video-text alignment task. In the second stage, text embedding and video embedding are concatenated as input to the cross-model encoder for joint representation learning. We adopt Masked Language Modeling (MLM) and Video-Text Matching (VTM) which are widely used as pre-training tasks to learning cross-modal interaction.

Video-Text Alignment. Specifically, we first train the text-encoder and video-encoder using contrastive loss to align textual and visual representations. In addition to the proposed multimodal temporal contrastive loss, we also adopt a standard contrastive loss on the global representation of long-form video and paragraph. The global contrastive loss is calculated as:

$$\mathcal{L}_{global}^{v2t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(s(V_i, T_i)\right)}{\sum_{j=1}^{B}\exp\left(s(V_i, T_j)\right)}, \ \mathcal{L}_{global}^{t2v} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(s(T_i, V_i)\right)}{\sum_{j=1}^{B}\exp\left(s(T_i, V_j)\right)}, \tag{5}$$

where V_i and T_i are the representation of i-th video and paragraph in the batch, respectively. The global alignment loss $\mathcal{L}_{global}$ is the average of $\mathcal{L}_{global}^{v2t}$ and $\mathcal{L}_{global}^{t2v}$.

The combination of multimodal temporal contrastive loss and global contrastive loss is used as the pre-training objective for the first stage:

$$\mathcal{L}_{stage1} = \mathcal{L}_{global} + \lambda_1 \mathcal{L}_{time}, \tag{6}$$

where λ_1 denotes the weight of MTC compared with global contrastive loss.

Masked Language Modeling. We follow the previous vision-language pre-training works to mask word tokens and predict the ground-truth labels from the output of cross-modal encoder, which integrates the context of other textual tokens and visual tokens:

$$\mathcal{L}_{mlm} = -\mathbb{E}_{(\mathcal{W},\mathcal{V})}\log p\left(w_i \mid \mathcal{W}_{\setminus i}, \mathcal{V}\right), \tag{7}$$

where $\mathcal{W}$ denotes the word tokens, $\mathcal{V}$ denotes the visual tokens, and w_i denotes the masked token. We adopt the same masking strategy and prediction method as BERT [11].

Video-Text Matching. To fuse the textual and visual information and generate a cross-modal representation, we use VTM as one pre-training task. VTM predicts whether the input paragraph and video are matched. We randomly replace the aligned video to a sampled negative with a probability of 0.5. We use a projection layer on the top of [CLS] embedding to predict a two-class matching logits y, and then compute negative log likely-hood loss as the VTM loss:

$$\mathcal{L}_{vtm} = -\mathbb{E}_{(\mathcal{W},\mathcal{V})}\log p\left(y \mid \mathcal{W}, \mathcal{V}\right). \tag{8}$$

Table 1: Result of paragraph-to-video retrieval on ActivityNet Captions dataset [19].

Method	Pre-training dataset	Video Length	R@1 ↑	R@5 ↑	R@50 ↑	MedR ↓
HSE [48]	-	-	20.5	49.3	-	-
ClipBERT [20]	CC3M	-	21.3	49.0	-	6.0
HD-VILA [43]	HD-VILA-100M	13.4s	27.4	56.0	93.1	4.0
Support Set [32]	HowTo100M	3.6s	29.2	61.6	94.7	3.0
TACo [45]	HowTo100M	3.6s	30.4	61.2	93.4	3.0
LF-VLP (Ours)	LF-VILA-8M	100.2s	**35.3**	**65.4**	**95.0**	**3.0**

Table 2: Result of paragraph-to-video retrieval on two datasets. * denotes results by our re-implementation.

(a) Result on DiDeMo dataset [1].

Method	R@1 ↑	R@5 ↑	R@10 ↑
FSE [48]	13.9	36.0	-
ClipBERT [20]	20.4	48.0	69.0
HD-VILA [43]	26.0	54.8	69.0
Frozen [3]	31.0	59.8	72.4
LF-VLP (Ours)	**35.0**	**64.5**	**75.8**

(b) Result on QuerYD dataset [31].

Method	R@1 ↑	R@5 ↑	R@10 ↑
MOEE [29]	11.6	30.2	43.2
CE [25]	13.9	37.6	78.3
TeachText [10]	14.4	37.7	50.9
Frozen* [3]	53.8	75.7	82.7
LF-VLP (Ours)	**66.2**	**82.5**	**87.6**

In the second stage, we freeze the video-encoder and text-encoder as [43] to accelerate training. The overall loss of stage 2 is the combination of MLM and VTM:

$$\mathcal{L}_{stage2} = \mathcal{L}_{mlm} + \lambda_2 \mathcal{L}_{vtm}, \tag{9}$$

where λ_2 is the weight of VTM in consideration of MLM.

4 Experiments

In this section, we first introduce the pre-training details, and then show experiments of utilizing pre-trained model on downstream paragraph-to-video retrieval and long-form video QA tasks to verify the effectiveness of our proposed LF-VLP.

4.1 Pre-training Details

Pre-training Dataset. To facilitate research on long-form video understanding, We build a large-scale long-form video-paragraph dataset based on HD-VILA-100M [43], which is an existing large-scale video-language dataset with diverse categories. It contains 100 million clip-text pairs derived from 3.3 million YouTube videos. We keep continuous clips with at least 4 clips and construct a dataset with 8.5 million long-from videos and corresponding transcripts, namely LF-VILA-8M. The average duration of each video is 100.2 seconds and the average number of words of each paragraph is 307.9. We provide additional statistics and examples in the supplementary material. The dataset will be made public to the research community.

Implementation Details. During pre-training, our model samples 4 consecutive clip-sentence pairs as input. We uniformly sample 8 frames from each clip and resize the frames to 192×320. we use WordPiece tokenizer like BERT to split each sentence to tokens with a max length of 50. For video encoder, we use Swin-Transformer [26] as backbone and integrate our proposed HTWA for frame sequence. Temporal window sizes are set to five stages: 2, 4, 8, 16, 32, respectively. We use 8×8 patches and use a fixed spatial window of 3×5, the output feature is down-sampled by 64 times to 3×5. We adopt a 12-layer Transformer network for text encoder, with 8 layers for the first part and 4 layers for the second part. We also use a 12-layer Transformer network for cross-modal encoder. The

Table 3: Results of text-to-video on Condensed Movie dataset [2] from official leaderboard.

Method	Geometric Mean	R@1 ↑	R@5 ↑	R@10 ↑
MoEE [29]	5.88	1.94	7.84	13.38
TeachText [10]	23.15	12.08	27.40	37.45
LF-VLP (Ours)	**26.40**	**13.56**	**32.47**	**41.79**

Table 4: Results of video question answering tasks.

(a) ActivityNet QA [46].

Method	Acc
MAR-VQA [50]	34.6
CoMVT [34]	38.8
VQA-T [44]	38.9
MERLOT [47]	**41.4**
LF-VLP (Ours)	39.9

(b) How2QA [22].

Method	Acc
CLIP [33] by [23]	69.3
CLIP-SF [23]	72.9
ResNet-SF [23]	74.3
HERO [22]	74.3
LF-VLP (Ours)	**76.1**

(c) VIOLIN [24].

Method	Acc
LXMERT [37]	66.3
Det-BERT [24]	67.8
GVE [8]	68.4
HERO [22]	68.6
LF-VLP (Ours)	**70.9**

weight of video encoder is initialized with Swin-Transformer pre-trained on ImageNet21K. We use the first 12 layers of BERT-Large to initialize the weight of text encoder, and the last 12 layers to initialize the weight of cross-modal encoder.

In pre-training, we use AdamW optimizer with a learning rate of 5e-5, and warm-up the learning rate for 1 epoch, followed by a linear decay, we use a weight decay of 0.05. We train our model with 32 NVIDIA Tesla V100 GPUs. For stage one, we use a batch size of 512 and train for 10 epochs. For stage two, we use a batch size of 1,536 and train for 10 epochs. We use the model from stage one for retrieval tasks since two-stream architecture is efficient for retrieval and widely used in previous works. Pretrained model from stage two is applied for video QA tasks. We excluded videos that have overlap with downstream tasks from the training dataset using YouTube IDs.

4.2 Paragraph-Video Retrieval

We conduct retrieval task on four paragraph-video retrieval datasets: **ActivityNet Captions** [19], **DiDeMo** [1] , **QuerYD** [31] and **Condensed Movie** [2]. Details of each dataset and implementation are in the supplementary material.

Results. Tab. 1~ 3 show the results of LF-VILA on 4 paragraph-to-video retrieval datasets. For the most widely used **ActivityNet Captions** [19] dataset, we surpass the SOTA model TACo [45] by **16.1%** on R@1. TACo is pre-trained on HowTo100M [30] using pre-extracted feature. This demonstrates the benefit of end-to-end training and utilization of long-form video dataset. Compared to **HD-VILA** [43] which is trained on 100M short-form video and sentence pairs, we use long-form videos with less data and achieve **28.8%** improvement in terms of R@1. This shows the effectiveness of our model LF-VLP in learning better alignment for long-form video and language. For **DiDeMo** [1], we also observe a significant improvement. In particular, we obtain **12.9%** improvement in terms of R@1 over the SOTA model Frozen [3]. On **QuerYD** [31] and **Condensed Movie** [2], we outperform the previous best models Frozen [3] and TeachText [10] with a relative **23.0%** and **12.3%** improvement, respectively. These two datasets are challenging due to their long videos. The result shows the value of long-form video-language pre-training and that our model LF-VLP can better understand the story line and temporal relations in long videos.

4.3 Video Question Answering

We conduct video QA task on three widely-used datasets for long-form video understanding: **ActivityNet-QA** [46], **How2QA** [22] and **VIOLIN** [24]. Details of each dataset and implementation are in the supplementary material.

Table 5: Ablation study on ActivityNet retrieval. We sample 1M pairs of data for pre-training.

(a) Analysis of pre-training tasks.

Pretain	ActivityNet Retrieval		
	R@1	R@5	R@50
w/o Pre-training	15.0	40.2	85.8
$\mathcal{L}_{global}$	26.1	56.7	92.7
$\mathcal{L}_{global} + \mathcal{L}_{time}$	**27.8**	**58.3**	**92.8**

(b) Analysis of temporal window (TW).

Method	TW#	Time	ActivityNet Retrieval		
			R@1	R@5	R@50
Fixed	4	0.91×	25.1	54.9	91.8
	32	1.49×	**26.2**	56.4	92.3
HTWA	[2-32]	1.00×	26.1	**56.7**	**92.7**

Results. Tab. 4 shows the results on 3 video question answering tasks. For **ActivityNet QA** [46], we outperform most previous works except MELORT [47]. Note that VQA-T [44] and MERLOT [47] are specifically designed for video QA. VQA-T [44] uses automatically generates 69M video-question-answer triplets from narrated videos for training. MERLOT [47] utilizes 180M pairs of data and it needs excessive computational cost for training (30K TPU hours), while we only need 3K GPU hours. For **How2QA** [22] and **VIOLIN** [24], we achieve the new SOTA. This illustrates the reasoning capability of our model on long-form videos by better capturing the long-range dependency between video clips and paragraphs.

4.4 Ablation Study

To validate the effectiveness of our proposed MTC loss and HTWA mechanism, we conduct ablation study on ActivityNet paragraph-to-video retrieval task with a subset of data to save resources. We randomly sample 1M video-paragraph pairs from the whole LF-VILA-8M for pre-training. **(1) Does pre-training benefit downstream tasks?** As Tab. 5a shows, our model improves R@1 by 11.1% after pre-training with the global alignment, which shows the benefit of pre-training. **(2) Is multimodal temporal contrastive loss helpful?** As shown in Tab. 5a, when combined with the MTC loss, the performance is further improved by 1.7%. **(3) Is hierarchical temporal window attention helpful?** In Tab. 5b, we compare our methods with video backbone using fixed window sizes. When we apply a small fixed window size (e.g.,4), the performance is relatively poor. This indicates the limitation of small attention window for modeling long-range dependency. When we increase the window size to 32 to cover a whole video, there is almost no improvement in performance, while the computational cost and training time increase significantly.

5 Conclusion

In this paper, we study video-language pre-training on a large-scale long-form video-paragraph dataset. To better align long-from videos and paragraphs, we propose a multimodal temporal contrastive (MTC) loss to capture the rich temporal relation between different modalities. In addition, we design a hierarchical temporal window attention (HTWA) mechanism to be applied with an image Transformer. Our proposed long-form video-language pre-training model (LF-VLP) combined with MTC and HTWA can learn effective multi-modal representation by capturing long-range dependency from long-form videos efficiently. Experiments on 7 long-form video-language understanding tasks verify the effectiveness of our model.

Limitation and Broader Impact. This paper has the broader impact on many video-language understanding applications such as video-text retrieval, video question-answering, etc. Since we apply two-stream architecture in the first stage, we can also utilize the single-modality features (i.e., video and language) from our model for even broader tasks. By learning vision-language representation from unlabeled videos and subtitles, our work may be easily extended and scaled to larger data. On the other hand, vision-language pre-training may learn biased or offensive content from user-generated video-subtitle data. This may cause improper understanding of videos. However, these concerns are general to the entire fields and are not amplified by this work.

References

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

[2] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*, pages 460–479, 2020.

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

[4] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Shot and scene detection via hierarchical clustering for re-using broadcast video. In *CAIP*, pages 801–811. Springer, 2015.

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, pages 813–824. PMLR, 2021.

[6] 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, 2015.

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[8] Junwen Chen and Yu Kong. Explainable video entailment with grounded visual evidence. In *ICCV*, pages 2001–2010, 2021.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020.

[10] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *ICCV*, pages 11583–11593, 2021.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pages 4171–4186, 2019.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICML*, 2021.

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[14] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016.

[15] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. COOT: Cooperative hierarchical transformer for video-text representation learning. In *NeurIPS*, pages 22605–22618, 2020.

[16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842–5850, 2017.

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.

[18] Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. *AAAI*, 2020.

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.

[20] 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 7331–7341, 2021.

[21] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *EMNLP*, pages 1369–1379, 2018.

[22] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, pages 2046–2065, 2020.

[23] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. *NeurIPS*, 2021.

[24] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *CVPR*, pages 10900–10910, 2020.

[25] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022.

[28] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020.

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

[30] 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 *ICCV*, pages 2630–2640, 2019.

[31] Andreea-Maria Oncescu, João F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP*, pages 2265–2269. IEEE, 2021.

[32] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2020.

[33] 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*, pages 8748–8763. PMLR, 2021.

[34] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *CVPR*, pages 16877–16887, 2021.

[35] Mykhailo Shvets, Wei Liu, and Alexander C Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *ICCV*, pages 9756–9764, 2019.

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, pages 568–576, 2014.

[37] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5100–5111, 2019.

[38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[40] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, pages 284–293, 2019.

[41] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *CVPR*, pages 1884–1894, 2021.

[42] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018.

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

[44] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021.

[45] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *ICCV*, pages 11562–11572, 2021.

[46] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019.

[47] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *NeurIPS*, 34, 2021.

[48] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, pages 374–390, 2018.

[49] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2914–2923, 2017.

[50] Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao. Multichannel attention refinement for video question answering. *TOMM*, 16:1–23, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]