Stitching Segments and Sentences towards Generalization in Video-Text Pre-training

Fan Ma[1]*, Xiaojie Jin[2]†, Heng Wang[2], Jingjia Huang[2], Linchao Zhu[1], Yi Yang[1]†

[1] Zhejiang University
[2] Bytedance Inc.
mafan@zju.edu.cn, {jinxiaojie, heng.wang, huangjingjia}@bytedance.com, {zhulinchao, yangyics}@zju.edu.cn

Abstract

Video-language pre-training models have recently achieved remarkable results on various multi-modal downstream tasks. However, most of these models rely on contrastive learning or masking modeling to align global features across modalities, neglecting the local associations between video frames and text tokens. This limits the model's ability to perform fine-grained matching and generalization, especially for tasks that selecting segments in long videos based on query texts. To address this issue, we propose a novel stitching and matching pre-text task for video-language pre-training that encourages fine-grained interactions between modalities. Our task involves stitching video frames or sentences into longer sequences and predicting the positions of cross-model queries in the stitched sequences. The individual frame and sentence representations are thus aligned via the stitching and matching strategy, encouraging the fine-grained interactions between videos and texts. in the stitched sequences for the cross-modal query. We conduct extensive experiments on various benchmarks covering text-to-video retrieval, video question answering, video captioning, and moment retrieval. Our results demonstrate that the proposed method significantly improves the generalization capacity of the video-text pre-training models.

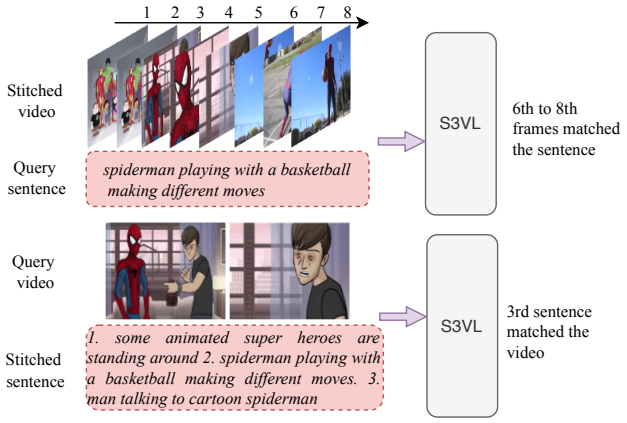

Figure 1: Stitching and matching for video-language pre-training (S3VL). The upper part is to predict the start and end frames from the stitched video given the text query. The lower part is to match the video with merged long sentences. The stitched videos and sentences are from different video-text pairs.

Introduction

Video-language pre-training is a burgeoning research area that aims to learn universal representations from large-scale multi-modal data (Radford et al. 2021; Fu et al. 2021; Wang et al. 2022b, 2023). These pre-trained representations facilitate various downstream video-related tasks, such as video question answering (QA) (Xu et al. 2016; Jin et al. 2023), text-to-video retrieval (Anne Hendricks et al. 2017; Maharaj et al. 2017), and video captioning (Lin et al. 2022; Xu et al. 2017). By leveraging massive video-text pairs to align cross-modal features, the pre-trained video-language models have demonstrated remarkable performance on diverse applications (Radford et al. 2021; Xu et al. 2021a).

Contrastive learning is a common technique for video-language pre-training that aims to align global features of videos and texts (Radford et al. 2021; Lei et al. 2021). However, this approach ignores the fine-grained associations between video frames and text tokens, which may result in misalignment due to irrelevant frames in the video. For example, a party video with the caption "the person is cutting cake" may also include frames of children running around the table. Global alignment of the entire video with the sentence may not only degrade the performance of text-to-video matching, but also limit the generalization capability of the model when applied to video downstream tasks that requires temporal relationships in long form videos. Therefore, fine-grained alignment between relevant texts and video frames is crucial for learning generic multi-modal representations.

Masked language modeling (MLM) has been used recently to achieve fine-grained interactions via predicting the masked element with unmasked visual and text features (Fu et al. 2021; Ge et al. 2022a; Wang et al. 2022b). However, MLM can not assure the alignment between different modalities along the temporal dimension, which is essential for many reasoning tasks such as video QA and captioning.

*This work was done during Fan's internship at Bytedance.
†Corresponding authors.

LocVTP (Cao et al. 2022) attempts to address this problem by dividing the video into several clips and extracting phrases from sentences for contrastive learning, but the objective relies on pseudo alignment due to the lack of video segment annotations. LF-VILA (Sun et al. 2022) employs a proprietary long-form video-text dataset that comprises several video-text pairs in each video with temporal alignment. However, existing public datasets containing video-text pairs, such as WebVID (Bain et al. 2021a) and YT-Temporal-180M (Zellers et al. 2021), mainly consist of short videos. Long form video datasets with segment annotations are scarce and often suffer from noise issues, as exemplified by HowTo100M (Miech et al. 2019). Therefore, it is challenging to achieve the reliable video-text alignment without accurate annotations, which limits the generalization ability on video downstream tasks.

To enhance detailed matching between videos and texts, we introduce a novel pre-training task where video segments and sentences are stitched into longer sequences for matching during training. Our approach, as illustrated in Fig. 1, entails two objectives: 1) predicting the temporal positions of frames in the stitched video that aligns with to the textual description; 2) identifying the sentence in a stitched paragraph that matches a short video. For the first task, we stitch frame features from different videos within a training batch to produce the long form video. Subsequently, we integrate the stitched frame features with text features and predict the correct segment boundary for the sentence. Our approach enables individual frame features to learn from both language and visual contexts during pre-training, thus facilitating fine-grained alignment between individual frames and sentences without necessitating detailed annotations. Similarly, we form long paragraphs by stitching multiple sentences and predict the matching sentence for each short video. The results of extensive experimentation on several downstream tasks demonstrate that our proposed stitching and matching method during pre-training is superior. Our model has significantly improved zero-shot and fine-tuned text-to-video retrieval performance on three datasets. Moreover, our model has exhibited a strong generalization capability on four downstream tasks.

In summary, our contributions are three-fold.

- We introduce a novel task for video-text pre-training that involves stitching video segments and sentences into long sequences and predicting fine-grained boundaries to facilitate interactions between video frames and texts.
- We present innovative and effective methods for stitching and matching frame-level and word-level features, which enhance the interactions between them and boost the generalization capability of the pre-trained model.
- We conducted comprehensive experiments on four downstream tasks to demonstrate the superiority of our pre-training task. Our method outperforms the pre-trained model with contrastive learning and mask language modeling on all downstream tasks.

Related Work

Video-Language Pre-training

Pre-training video-language models on multi-modal data has become a popular approach to improve their performance on various downstream tasks, such as text-to-video retrieval (Xu et al. 2016), video question answering (Xu et al. 2016, 2017), and video captioning (Lin et al. 2022; Xu et al. 2016). Most existing models use contrastive learning to align videos and texts in a common feature space (Xu et al. 2021a; Huang et al. 2023; Jin et al. 2023), but this only achieves coarse-grained alignment. The contrastive learning is used to align features from different modalities in a global manner. Recent methods also use masked language modeling (MLM) to predict masked signals and enable multi-modal interactions (Li et al. 2020; Wang et al. 2022b; He et al. 2023; Liu et al. 2023), but this still does not capture fine-grained alignment. LocVTP (Cao et al. 2022) proposes a clip-phrase contrastive objective, but it relies on pseudo supervision that may not reflect the true matching. LF-ViLA (Sun et al. 2022) uses a new long-form dataset with temporal annotations, but this dataset is not publicly available and such annotations are difficult to obtain in practice. We present a novel stitching and matching task that does not require any annotations and can enhance the fine-grained alignment of video and text features.

Data Augmentation

Data augmentation enhances the diversity and quantity of training data for various tasks, thereby improving model generalization. It involves modifying the original data in different ways. For instance, image augmentation can blend two images into one with a mixed soft label (Zhang et al. 2018; Verma et al. 2019; Ma et al. 2022). Text augmentation can generate new texts by replacing, inserting or deleting words, or by using back translation or paraphrasing (Wei and Zou 2019; Zhang et al. 2020a). Video augmentation can alter the temporal or spatial dimensions of videos by adjusting the speed, order or duration of frames according to the video categories (Yun et al. 2020; Xu et al. 2021b). However, most existing methods are task-specific and need domain knowledge or human supervision to produce augmented data. We propose a general method that stitches frame and sentence features for matching to align multi-modal representations without prior knowledge.

Method

Overview of S3VL

Our framework consists of two encoders for extracting video and text features separately and one multi-modal encoder for integrating both visual and text features, as shown in Fig. 2. Given the i-th video-text pair, the visual encoder produces the video features $\mathbf{v}_i = \{\mathbf{v}_i^k\}_{k=1}^{T}$ where T is the number of frames. The text encoder takes a tokenized text sequence as input, with a [CLS] text token inserted at the beginning (represented by the circle in the figure), and generates the token features $\mathbf{s}_i = \{\mathbf{s}_i^k\}_{k=0}^{L+1}$ where L is the number of tokens in each sentence.

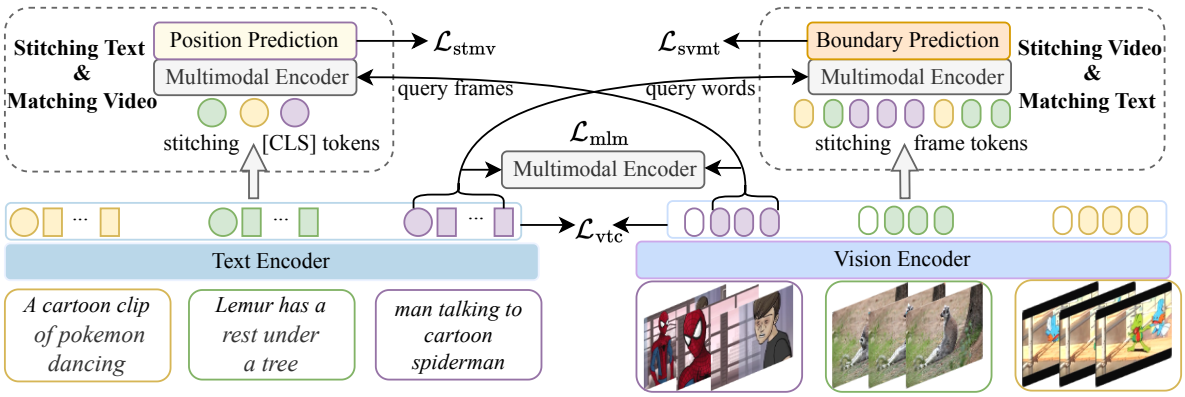

Figure 2: The framework of stitching video segments and sentences for video-text pre-training (S3VL). There are four pre-training objectives in the figure: $\mathcal{L}_{\text{vtc}}$ to globally align the video and text features, $\mathcal{L}_{\text{mlm}}$ to predict masked word tokens given the video and other language context, $\mathcal{L}_{\text{stmv}}$ to predict position of the sentence that matches the video segments, and $\mathcal{L}_{\text{svmt}}$ to predict boundary of the segment that matches the text description. We use same color for the paired video and text. The frame tokens are concatenated with the text tokens to form the input for the multi-modal encoder.

Following previous pre-training methods (Huang et al. 2023; Lei et al. 2021), we employ contrastive loss $\mathcal{L}_{\text{vtc}}$ on the paired video and text embedding to align the cross-modal representations globally:

$$\begin{aligned}\mathcal{L}_{\text{vtc}} = &-\frac{1}{2B}\sum_i log\frac{\exp(\text{sim}(\mathbf{v}_i,\mathbf{s}_i)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{v}_i,\mathbf{s}_j)/\tau)} \\ &-\frac{1}{2B}\sum_i log\frac{\exp(\text{sim}(\mathbf{s}_i,\mathbf{v}_i)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{s}_i,\mathbf{v}_j)/\tau)},\end{aligned} \tag{1}$$

where $\mathbf{v}_i$ and $\mathbf{s}_i$ are the i-th paired video and sentence representations, and τ is a temperature parameter that controls the sharpness of the distribution. sim(,) is the cosine similarity and B is the number of text-video pairs.

To integrate both the visual and text features, we project all the frame and text tokens into a common embedding space and concatenate them into a new sequence as the multi-modal input. We then apply masked language modeling $\mathcal{L}_{\text{mlm}}$ on the concatenated sequence to predict masked language tokens given frames and context words:

$$\mathcal{L}_{\text{mlm}} = -\frac{1}{B}\sum_i\sum_k logP(s_i^k|s_i^{<k}, s_i^{>k}; \mathbf{v}_i), \tag{2}$$

where s_i^k is the masked token in the text sequence, $s^{<k}$ is the sequence before s_i^k, and $s_i^{>k}$ is the sequence after s_i^k. For the video captioning task that generates one word token at a time, we use the casual mask and remove the $s_i^{>k}$ in Eq. (2) to predict the masked token.

The contrastive learning and masked language modeling only learn the coarse alignment between videos and texts, limiting the generalization capability on downstream tasks that require fine-grained matching between frames and sentences. To address this issue, we propose a novel stitching and matching task as shown in Fig. 2. Our proposed method enhances the detailed alignment between modalities and improves the model generalization capability.

Stitching and Matching

The frame features are interacted with text features in the multi-modal encoder. However, this interaction between videos and sentences remains somewhat rudimentary as the alignment between video and text representations are not guaranteed. To address this limitation, this section proposes a stitching and matching task that enables fine-grained alignment across different modalities. As accurate annotations indicating which segments in videos correspond to specific text descriptions are nearly unavailable, we stitch short video segments or sentences into longer sequences to provide supervision. Subsequently, the model predicts the temporal boundaries of the stitched longer video tokens based on the provided text descriptions, and identifies the matched positions within the stitched sentences based on the short video segments. Each of the objectives is discussed in detail in the following sections.

Matching text in stitched videos. This task is to predict the temporal boundary of the text-matched segment in a composed long video sequence. During pre-training, we concatenate the short videos in a training batch to form a long video sequence. Only one segment in this sequence corresponds to the text query.

Stitching video segments. We stitch frames from different videos in two ways. The first method randomly permutes the order of videos in a training batch and concatenates their frame features while preserving the temporal order within each video as shown in Fig. 2. This results in a sequence of BT frame tokens. We add frame position embeddings to these tokens and combine them with word tokens from the sentence to form the multi-modal input. Another method to create long video sequences is to sample frame tokens from different videos and insert them into a background sequence. We select K_p frame tokens from the video that matches the text and K frame tokens from the other videos, where K is the desired length. We denote the frame tokens that matches the text as the positive frame tokens and the irrelevant tokens

as the negative tokens. We randomly insert the K_p positive frame tokens into K background tokens. By stitching video segments, we establish the temporal boundary of the text-matched segment and represent it as (s_v, e_v).

Boundary prediction. We use a multi-modal encoder to enable cross-modal interactions between frame and text features. To enhance these interactions, we design a boundary prediction task that requires locating the text-matched segment in the video. We use a matching head with two linear layers to predict boundaries for updated frame features $\mathbf{p}_{\text{svmt}} \in \mathbb{R}^{BT\times 2}$ where the subscript svmt is the abbreviation for stitching video and matching text task. The model produces the probability of each frame belonging to the starting and ending boundaries of a video clip that corresponds to the text. We use the softmax function to generate the boundary. The objective of matching text in the stitched video is then written as:

$$\mathcal{L}_{\text{svmt}} = -\log \text{softmax}(\mathbf{p}^0_{\text{svmt}})^{s_v} - \log \text{softmax}(\mathbf{p}^1_{\text{svmt}})^{e_v}, \tag{3}$$

where $\mathbf{p}_{\text{svmt}} = [\mathbf{p}^0_{\text{svmt}}, \mathbf{p}^1_{\text{svmt}}]$ denotes the start and end predictions. In addition to the boundary prediction loss in Eq. (3), we can also use the updated [CLS] text token from multi-modal encoder to directly regress the start and end positions. Alternatively, we could predict the matching probability for text in all frame tokens. However these attempts ignores the interactions between individual frame tokens.

Matching video in stitched sentences. We also match a short video clip in the stitched sentence to align multi-modal representations. The text have both relevant and irrelevant descriptions for the video clip. We stitch word tokens from different sentences into a long sequence and predict the position of the video segment that matches the text.

Stitching sentences. We use two methods to stitch sentences. The first one is to shuffle and concatenate the tokens in each sentence in the training batch. Then, we use the video segment to predict start and end positions in the longer text tokens. The sampling strategy is not suitable for stitching sentences, as it may alter the semantic information by missing or repeating some tokens. For example, omitting the word "not" may reverse the meaning. We also propose to stitch only the [CLS] text token in every sentence, as shown in Fig. 2. For one video and B merged [CLS] text tokens, we form the multi-modal input containing $B + T$ tokens.

Position prediction. For the first text stitching method, we use the same boundary prediction loss in Eq. (3) for the shuffled sentences. For the stitched [CLS] text tokens, we only predict the position of the video segment that matches the text. The training objective for matching video is given by:

$$\mathcal{L}_{\text{stmv}} = -\log \text{softmax}(\mathbf{p}_{\text{stmv}})^{m_v}, \tag{4}$$

where $\mathbf{p}_{\text{stmv}} \in \mathbb{R}^B$ is the logit prediction of text position and m_v is the index of sentence that matches the video.

Pre-training Objectives

We use four losses to pre-train the model as shown in Fig. 2. The text-video contrastive is to coarsely project video and text into the common feature space, while the objective of stitching and matching is to enhance fine-grained visual language interaction for video-text alignment. With the mask language modeling, our pre-training objective is formed via:

$$\mathcal{L} = \mathcal{L}_{\text{vtc}} + \alpha\mathcal{L}_{\text{mlm}} + \beta\mathcal{L}_{\text{sm}}, \tag{5}$$

where $\mathcal{L}_{\text{sm}} = \mathcal{L}_{\text{svmt}} + \mathcal{L}_{\text{stmv}}$ denotes the stitching and matching objective. The α and β are the hyper-parameters to balance each pre-training tasks, which are set to 1.

Experiments

Datasets and Downstream Tasks

Pre-training datasets. Following recent work (Huang et al. 2023), we use the WebVid (Bain et al. 2021b) and the Google Conceptual Captions (Sharma et al. 2018) as the training data. The static image is treated as the video with only single frame during the pre-training.

Downstream tasks. We evaluate our method on four popular downstream tasks. (1) **Text-to-video retrieval** on three datasets: MSR-VTT (Xu et al. 2016), DiDeMo (Anne Hendricks et al. 2017), and LSMDC (Maharaj et al. 2017). (2) **Video question answering** on MSR-VTT (Xu et al. 2016) and MSVD (Xu et al. 2017). (3) **Video captioning** on MSR-VTT (Xu et al. 2016) and MSVD (Xu et al. 2017). (4) **Video moment retrieval with language** is to predict temporal boundary in a long video for language query. We conduct experiments on DiDeMo to testify the pre-training models.

Implementation Details

We adopt VideoSwin (Liu et al. 2022b) as the video encoder with pre-trained weights on the Kinetics-400 dataset (Kay et al. 2017), and pre-trained BERT-base model as the text encoder. The multi-modal encoder is initialized from the last three layers of the BERT-base model. We pre-train our model for 40 epochs, using a batch size of 2048 on 64 NVIDIA V100 GPUs. We use AdamW (Loshchilov and Hutter 2019) optimizer with a weight decay 0.005 and betas (0.9, 0.98). The learning rate is first set to 5e-5 and then decays by 10 times following a cosine annealing decay schedule. All video frames are resized to 224×224, and 8 frames are randomly sampled in a video while the temporal order is preserved. During pre-training, all words in the sentence is random masked with 15% probability to enable the mask language modeling in both normal and causal attentions. For the retrieval task, we only fine-tune the uni-modal encoders with the contrastive learning. For both video QA and video captioning tasks, we adopt the casual mask in both text and multi-modal encoders to generate both answers and descriptions. For moment retrieval with language tasks, we use the pre-trained visual encoder to extract video features first and adopt the off-the-shelf algorithms to train corresponding models with our extracted features. We have also used CLIP model as the encoder, which is pre-trained with image-text pairs, and further training the CLIP with the multi-modal encoder with more video-text pairs. The experiments with pre-trained image-text encoders are present in the supplementary material.

Method	Pre-training data	MSR-VTT			DiDeMo			LSMDC		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot										
Frozen (Bain et al. 2021b)	W2M+CC3M	18.7	39.5	51.6	21.1	46.0	56.2	9.3	22.0	30.1
VIOLET (Fu et al. 2021)	W2M+CC3M+Y180M	25.9	49.5	59.7	23.5	49.8	59.8	-	-	-
ALPRO (Li et al. 2022a)	W2M+CC3M	24.1	44.7	55.4	23.8	47.3	57.9	-	-	-
LocVTP (Cao et al. 2022)	W2M+CC3M	22.1	48.0	55.3	-	-	-	-	-	-
MCQ (Ge et al. 2022a)	W2M+CC3M	26.0	46.4	56.4	25.6	50.6	61.1	12.2	25.9	32.2
OA-Trans (Wang et al. 2022a)	W2M+CC3M	23.4	47.5	55.6	23.5	50.4	59.8	-	-	-
Miles (Ge et al. 2022b)	W2M+CC3M	26.1	47.2	56.9	27.2	50.3	63.6	11.1	24.7	30.6
Clover (Huang et al. 2023)	W2M+CC3M	25.8	49.6	60.1	28.0	53.5	65.1	13.8	28.1	38.3
RegionLearner (Yan et al. 2023)	W2M+CC3M	22.2	43.3	52.9	-	-	-	-	-	-
S3VL (Ours)	W2M+CC3M	**27.2**	**50.1**	**60.3**	**31.3**	**56.4**	**67.6**	**14.9**	**31.2**	**39.5**
Fine-tune										
HD-VILA (Sun et al. 2022)	L8M	28.8	57.4	69.1	35.6	65.3	78.0	17.4	34.1	44.1
Frozen (Bain et al. 2021b)	W2M+CC3M	31.0	59.5	70.5	31.0	59.8	72.4	15.0	30.8	39.8
All-in-one (Wang et al. 2022b)	W2M+H100M	37.9	68.1	77.1	32.7	61.4	73.5	-	-	-
VIOLET (Fu et al. 2021)	W2M+CC3M+Y180M	34.5	63.0	73.4	32.6	62.8	74.7	16.1	36.6	41.2
ALPRO (Li et al. 2022a)	W2M+CC3M	33.9	60.7	73.2	35.9	67.5	78.8	-	-	-
LocVTP (Cao et al. 2022)	W2M+CC3M	36.5	64.3	76.8	-	-	-	-	-	-
OA-Trans (Wang et al. 2022a)	W2M+CC3M	35.8	63.4	76.5	34.8	64.4	75.1	18.2	34.3	43.7
Miles (Ge et al. 2022b)	W2M+CC3M	37.7	63.6	73.8	36.6	63.9	74.0	17.8	35.6	44.1
Lavender (Li et al. 2022b)	W2M+CC3M	37.8	63.8	75.0	47.4	74.7	82.4	22.2	**43.8**	**53.5**
MCQ (Ge et al. 2022a)	W2M+CC3M	37.6	64.8	75.1	37.0	62.2	73.9	17.9	35.4	44.5
Clover (Huang et al. 2023)	W2M+CC3M	38.6	67.4	76.2	45.1	74.3	82.2	22.7	42.0	52.6
RegionLearner (Yan et al. 2023)	W2M+CC3M	36.3	63.9	72.5	-	-	-	-	-	-
S3VL (Ours)	W2M+CC3M	**41.0**	**68.2**	**77.7**	**48.6**	**76.1**	**85.4**	**23.2**	42.2	51.3

Table 1: Text-to-video retrieval comparison on MSR-VTT, DiDeMo and LSMDC under the zero-shot and fine-tune setups. W2M, C3M, H100M, H8M, Y180M are abbreviations for WebVid2M (Bain et al. 2021b), CC3M (Sharma et al. 2018), HowTo100M (Miech et al. 2019), LF-VILA-8M (Sun et al. 2022), YT-Temporal-180M (Zellers et al. 2021), respectively. Higher Recall@k indicate better performance. The best performance is masked in bold under each setting.

Comparison to Prior Arts

Text-to-video retrieval. Tab. 1 illustrates the text-to-video retrieval results on MSR-VTT (Xu et al. 2016), DiDeMo (Anne Hendricks et al. 2017), and LSMDC (Maharaj et al. 2017) datasets under zero-shot and fine-tuning settings. Our proposed method significantly outperforms the previous approaches among all the datasets. Notably, the performance improvement with zero-shot evaluation demonstrates the stronger generalization ability of our method. Our S3VL achieves the highest recall on four datasets under the zero-shot setting. In detail, our method outperforms Clover (Huang et al. 2023) by 1.4% on MSR-VTT, 3.3% on DiDeMo, 1.1% on LSMDC in Recall@1. Moreover, our proposed method surpasses VIOLET by a large margin on both MSR-VTT and DiDeMo, even though VIOLET is pre-trained with more text-video pairs.

When fine-tuned on the three datasets, S3VL also shows superiority over the compared methods. Our method outperforms the compared methods across all the metrics on MSR-VTT and DiDeMo with a clear improvement. Compared to videos in MSR-VTT, videos in DiDeMo contain more frames and diverse scenes. The noticeable improvement on the DiDeMo also suggest that our method better matches long videos with texts. Compared to LocVTP (Cao et al. 2022) that leverages pseudo fine-grained alignment information, our model achieves much higher results under both zero-shot and fine-tune settings.

Video question answering. We evaluate our method on two open-ended video question answering datasets and compare it with several methods (Tab. 2). These include JustAsk (Yang et al. 2021), ALPRO (Li et al. 2022a), VIOLET (Fu et al. 2021), All-in-one (Wang et al. 2022b), Clover (Huang et al. 2023) and Lavender (Li et al. 2022b). Unlike All-in-one and Clover that use classification loss for the open-ended QA, our method generates answers without restricting the categories. Our method achieves the best performance on the MSR-VTT and exceeds Lavender by 0.4%.

Video captioning. We evaluate our method on video captioning task in Tab. 2. The causal mask is used during pre-training and fine-tuneing. And 60% of words are masked for captioning task during fine-tuning. Our model outperforms all other models on both MSR-VTT and MSVD datasets, demonstrating the effectiveness of the proposed S3VL.

Moment retrieval with natural language. We evaluate our method on the moment retrieval task that predicts the

Method	MSR-VTT VQA	MSR-VTT Cap.	MSVD VQA	MSVD Cap.
JuskAsk (Yang et al. 2021)	41.5	-	46.3	-
ALPRO (Li et al. 2022a)	42.1	-	45.9	-
VIOLET (Fu et al. 2021)	43.9	-	47.9	-
All-in-one (Wang et al. 2022b)	44.3	-	47.9	-
Clover (Huang et al. 2023)	43.9	-	51.9	-
PMT (Peng et al. 2023)	41.8	-	40.3	-
SwinBERT (Lin et al. 2022)	-	53.8	-	120.6
Lavender (Li et al. 2022b)	44.2	58.0	**55.4**	142.9
RegionLearner (Yan et al. 2023)	38.6	-	39.3	-
STOA-VLP (Zhong et al. 2023)	43.2	60.2	50.8	131.8
S3VL (Ours)	**44.7**	**61.9**	53.9	**148.2**

Table 2: Video question answering and captioning comparison on MSR-VTT and MSVD under the open-ended setting. Cap. denotes the video captioning task. We report the accuracy and the highest performance is masked in bold.

Method	Rank@0.5	Rank@0.7	AVG
2D-TAN (Zhang et al. 2020b)	42.8	23.2	33.0
LocVTP (Cao et al. 2022)	41.2	**24.8**	33.0
UMT (Liu et al. 2022a)	48.3	**29.3**	38.8
S3VL (Ours)	**49.6**	28.8	**39.7**

Table 3: Moment retrieval comparison on Charades-STA. The Rank@0.5 denotes the top-1 retrieval results with temporal IoU greater than 0.5. AVG indicates the average score of two metrics.

temporal boundary in a video for the text description. We use 2D-TAN (Zhang et al. 2020b) as the baseline model and compare it with LocVTP (Cao et al. 2022) that uses the same pre-training datasets. Tab. 3 shows the results on DiDeMo. Our method surpasses LocVTP by a large margin on Rank@0.5, indicating that our extracted features are more accurate in finding the temporal boundaries.

Analysis

To evaluate our design choices, we perform ablation experiments on WebVid1M, a subset of WebVid with one million pairs of videos and texts.

Pre-training objective. As shown in Tab. 4, we evaluate our proposed objectives in Eq. (3) and Eq. (4) on four tasks, including text-to-video retrieval, video question answering and captioning. Compared to the baseline which uses contrastive learning and masked language modeling, our model significantly improves the zero-shot retrieval performance by 1.6% on recall@1. The captioning task's performance can be enhanced by combining the input text, thereby increasing its impact. Similarly, stitching video clips together further improves the retrieval task. This is because retrieval relies on the similarity between global video and text representations, and stitching the global video features helps to reduce the distance between these two modalities. On

$\mathcal{L}_{\text{vtc}}$	$\mathcal{L}_{\text{mlm}}$	$\mathcal{L}_{\text{svmt}}$	$\mathcal{L}_{\text{stmv}}$	Retrieval	VQA	Cap.	MR
✓				21.2	41.2	53.6	41.8
✓	✓			20.8	42.7	56.9	41.6
✓	✓	✓		22.1	**43.1**	57.7	43.0
✓	✓		✓	21.9	42.7	57.5	42.8
✓	✓	✓	✓	**22.4**	43.0	**58.1**	**43.6**

Table 4: Effect of pretraining tasks on downstream tasks. The recall@1, accuracy, CIDEr, and Rank@0.5 are reported in the zero-shot text-to-video retrieval, video question answering, video captioning tasks, and moment retrieval respectively. MR denotes the moment retrieval task.

Strategy	Retrieval	VQA	Cap.	MR
Shuffling	21.7	42.9	**58.1**	42.3
Sampling	21.5	42.8	56.9	42.8
HardSampling	**22.1**	**43.1**	57.7	**43.6**

Table 5: Analysis of video stitching strategies. Results on text-to-video retrieval, video question answering, video captioning and moment retrieval tasks are reported.

all tasks, our model pre-trained with stitching and matching outperforms the baseline model. This suggests that our design obtains superior generalization capability.

Stitching video segments. We adopt three strategies in Tab. 5 to stitching individual video features into a long video sequence. The quality of the merged sequence is essential for improving generalization capabability of pre-trained model. The *Shuffling* strategy only changes the order of different videos where the order of frame features in the same video is preserved. *Sampling* strategy requires the max length to specify how many frames should be sampled. In this experiment, we only use boundary prediction loss $\mathcal{L}_{\text{svmt}}$ in Eq. (3) and specify the max sample number K to 128, and the minimum and the maximum number of positive frames K_p to 1 and 32. The *HardSampling* denotes that frame features are only sampled from those most similar videos in a batch. From Tab. 5, we observe that the *HardSampling* strategy acquires the highest zero-shot retrieval performance. It shows that merging frame tokens from similar videos for detailed matching is more helpful to the retrieval task.

Stitching sentences. We also compare two stitching strategies for sentences in Tab. 6. The word merging is similar to the shuffling in video merging, where only the sentence order is rearranged. We retain the special [CLS] text token when merging words. We predict the start and end positions for the video in this merging strategy. *Stitching [CLS]* denotes that only the first token in each sentence is selected to combine different texts and the position prediction loss $\mathcal{L}_{\text{stmv}}$ in Eq. (4) is employed. We found that the model with *Stitching [CLS]* achieves a bit better retrieval performance from the experimental results. In text-to-video retrieval task, only the feature of [CLS] text token is used to align with the video clips. Therefore, stitching [CLS] token would be beneficial to align the representations between videos and texts.

Strategy	Retrieval	VQA	Captioning	MRetrieval
Stitching words	21.7	**42.9**	57.4	43.3
Stitching [CLS]	**21.9**	42.7	**57.5**	**43.6**

Table 6: Analysis of word stitching strategies. Results on text-to-video retrieval, video question answering, and video captioning tasks are reported.

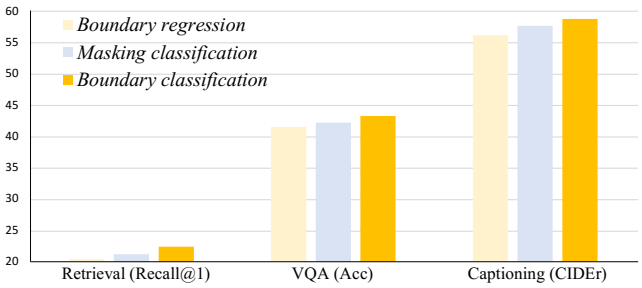

Figure 3: Boundary prediction with different strategies. Boundary classification denotes our method to produce the boundary probability at each position.

Boundary prediction. To predict precise segments in stitched long sequences, our model outputs the boundary probability at each updated token after the multi-modal encoder. We compare with several other methods for detailed matching on stitched videos as shown in Fig. 3. *Boundary regression* involves predicting the start and end frames of the matching segment on the [CLS] text token. *Masking classification* involves classifying each frame token as matching or not. Our model outperforms these baselines on all tasks, demonstrating the effectiveness of our pre-training strategy.

Generalization capability. We summarize the experimental results of our method with different backbones on four tasks in Tab. 7. Two visual backbones, including a video encoder SwinT (Liu et al. 2022b) and a frame encoder ViT (Dosovitskiy et al. 2021), are adopted to verify the stitching and matching (S&M) task for pre-training. On MSR-VTT, the model trained with stitching and matching strategy obtains significant improvement in the zero-shot text-to-video retrieval task for both video and image encoders. Both ViT and SwinT based models with S&M pre-training obtain about 1% performance gain on the mAP@0.5 metric. Our method also significantly improves the Rank@0.5 on Charades. Albeit ViT based model achieves lower performance than the SwinT based model, both visual encoders with the S&M pre-training obtain improvement on almost all tasks. This demonstrates that generic representations are well learned with our proposed stitching and matching task.

Qualitative examples. We extract frame and text features on the video in Charades-STA, and calculate the similarity between the visual and language features. We visualize the results in Fig. 4 to show the matching performance between models trained with and without stitching and matching. For the video with the description "person runs up the stairs",

Backbone	MSR-VTT Recall@1	MSR-VTT Acc	MSR-VTT CIDEr	Charades Rank@0.5
ViT (w/o S&M)	16.4	36.2	53.1	38.8
ViT (w S&M)	17.7	36.9	53.0	40.2
SwinT (w/o S&M)	20.8	42.7	56.9	41.8
SwinT (w S&M)	22.4	43.1	58.1	43.6

Table 7: Effect of stitching and matching pre-training task with different visual backbones. S&M denotes the stitching and matching task. Results on four tasks are reported.

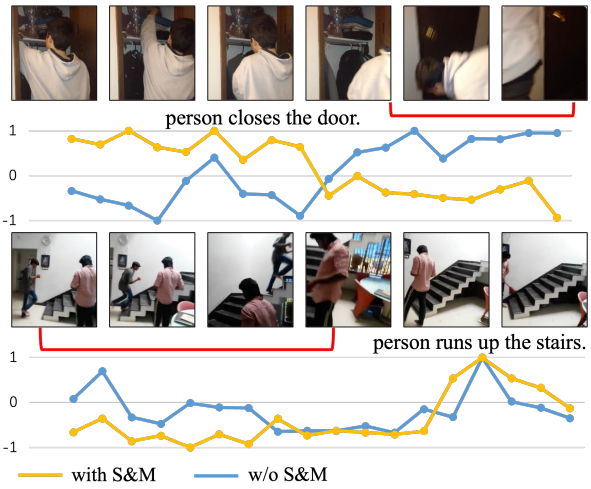

Figure 4: Visualization of similarity scores between frame and text features. Higher score indicates higher similarity between the frame and text features. The temporal ground-truth for the text description is marked with red.

although the similarity of the two models are close, only the model with S&M accurately localizes frames that match the caption. It shows that the cross modality features are better aligned in our method, and the temporal boundaries are more accurate for text descriptions.

Conclusion

We introduce a novel stitching and matching task for video-language pre-training to improve model generalization capability. Without any fine-grained annotations, we merge video and text features into longer sequences and use the multi-modal encoder to predict the precise boundary. The proposed task is simple yet effective to learn a generic mulitimodal representations. Extensive experiments on multiple tasks demonstrate the strength of the proposed method.

Acknowledgements

This work is support by Major Program of National Natural Science Foundation of China (62293554) , the Fundamental Research Funds for the Central Universities (No. 226-2022- 00051), and China Postdoctoral Science Foundation (524000-X92302).

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 5803–5812.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021a. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021b. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1728–1738.

Cao, M.; Yang, T.; Weng, J.; Zhang, C.; Wang, J.; and Zou, Y. 2022. LocVTP: Video-Text Pre-training for Temporal Localization. *ArXiv*, abs/2207.10362.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Fu, T.-J.; Li, L.; Gan, Z.; Lin, K.; Wang, W. Y.; Wang, L.; and Liu, Z. 2021. VIOLET : End-to-End Video-Language Transformers with Masked Visual-token Modeling. *ArXiv*, abs/2111.12681.

Ge, Y.; Ge, Y.; Liu, X.; Li, D.; Shan, Y.; Qie, X.; and Luo, P. 2022a. Bridging Video-Text Retrieval With Multiple Choice Questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16167–16176.

Ge, Y.; Ge, Y.; Liu, X.; Wang, A. J.; Wu, J.; Shan, Y.; Qie, X.; and Luo, P. 2022b. MILES: Visual BERT Pre-training with Injected Language Semantics for Video-text Retrieval. *arXiv preprint arXiv:2204.12408*.

He, X.; Chen, S.; Ma, F.; Huang, Z.; Jin, X.; Liu, Z.; Fu, D.; Yang, Y.; Liu, J.; and Feng, J. 2023. VLAB: Enhancing Video Language Pre-training by Feature Adapting and Blending. arXiv:2305.13167.

Huang, J.; Li, Y.; Feng, J.; Sun, X.; and Ji, R. 2023. Clover: Towards A Unified Video-Language Alignment and Fusion Model. *CVPR 2023*.

Jin, Y.; Niu, G.; Xiao, X.; Zhang, J.; Peng, X.; and Yu, J. 2023. Knowledge-Constrained Answer Generation for Open-Ended Video Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8141–8149.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, A.; Suleyman, M.; and Zisserman, A. 2017. The Kinetics Human Action Video Dataset. *ArXiv*, abs/1705.06950.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7331–7341.

Li, D.; Li, J.; Li, H.; Niebles, J. C.; and Hoi, S. C. H. 2022a. Align and Prompt: Video-and-Language Pre-training with Entity Prompts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4943–4953.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Li, L.; Gan, Z.; Lin, K.; Lin, C.-C.; Liu, Z.; Liu, C.; and Wang, L. 2022b. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*.

Lin, K.; Li, L.; Lin, C.-C.; Ahmed, F.; Gan, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17928–17937.

Liu, Y.; Li, S.; Wu, Y.; Chen, C. W.; Shan, Y.; and Qie, X. 2022a. UMT: Unified Multi-modal Transformers for Joint Video Moment Retrieval and Highlight Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3032–3041.

Liu, Y.; Xu, L.; Xiong, P.; and Jin, Q. 2023. Token Mixing: Parameter-Efficient Transfer Learning from Image-Language to Video-Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1781–1789.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022b. Video Swin Transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3192–3201.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Ma, F.; Wu, Y.; Yu, X.; and Yang, Y. 2022. Learning With Noisy Labels via Self-Reweighting From Class Centroids. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11): 6275–6285.

Maharaj, T.; Ballas, N.; Rohrbach, A.; Courville, A.; and Pal, C. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6884–6893.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Peng, M.; Wang, C.; Shi, Y.; and Zhou, X.-D. 2023. Efficient End-to-End Video Question Answering with Pyramidal Multimodal Transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2038–2046.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*.

Sun, Y.; Xue, H.; Song, R.; Liu, B.; Yang, H.; and Fu, J. 2022. Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning. In *NeurIPS*.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, 6438–6447. PMLR.

Wang, A. J.; Ge, Y.; Cai, G.; Yan, R.; Lin, X.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022a. Object-aware Video-language Pre-training for Retrieval. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Wang, A. J.; Ge, Y.; Yan, R.; Ge, Y.; Lin, X.; Cai, G.; Wu, J.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022b. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*.

Wang, X.; Wang, W.; Shao, J.; and Yang, Y. 2023. Learning to Follow and Generate Instructions for Language-Capable Navigation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6382–6388. Hong Kong, China: Association for Computational Linguistics.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, 1645–1653.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021a. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Xu, M.; Perez-Rua, J.-M.; Escorcia, V.; Martinez, B.; Zhu, X.; Zhang, L.; Ghanem, B.; and Xiang, T. 2021b. Boundary-sensitive Pre-training for Temporal Localization in Videos. arXiv:2011.10830.

Yan, R.; Shou, M. Z.; Ge, Y.; Wang, J.; Lin, X.; Cai, G.; and Tang, J. 2023. Video-Text Pre-training with Learned Regions for Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3100–3108.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1666–1677.

Yun, S.; Oh, S. J.; Heo, B.; Han, D.; and Kim, J. 2020. Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34: 23634–23651.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. 2020a. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 11328–11339. PMLR.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2D Temporal Adjacent Networks forMoment Localization with Natural Language. In *AAAI*.

Zhong, W.; Zheng, M.; Tang, D.; Luo, X.; Gong, H.; Feng, X.; and Qin, B. 2023. STOA-VLP: Spatial-Temporal Modeling of Object and Action for Video-Language Pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3715–3723.