Token Mixing: Parameter-Efficient Transfer Learning from Image-Language to Video-Language

Yuqi Liu[1,2*], **Luhui Xu**[2], **Pengfei Xiong**[2], **Qin Jin**[1†]

[1] School of Information, Renmin University of China

[2] Tencent

{yuqi657, qjin}@ruc.edu.cn, luhuixu.cn@gmail.com, xiongpengfei2019@gmail.com

Abstract

Applying large scale pre-trained image-language model to video-language tasks faces two challenges. One is how to effectively transfer knowledge from static images to dynamic videos, and the other is how to cope with the prohibitive cost of fully fine-tuning due to the growing size of the model. Existing works that attempt to realize parameter-efficient image-language to video-language transfer learning can be categorized into two types: 1) appending a sequence of temporal transformer blocks after the 2D Vision Transformer (ViT), and 2) inserting a temporal block into the ViT architecture. While these two types of methods only require fine-tuning the newly added components, there are still many parameters to update, and they are only validated on a single video-language task. In this work, based on our analysis of the core ideas of different temporal modeling components in existing approaches, we propose a token mixing strategy to allow cross-frame interactions, which enables transferring from the pre-trained image-language model to video-language tasks through selecting and mixing a key set and a value set from the input video samples. As token mixing does not require the addition of any components or modules, we can partially fine-tune the pre-trained image-language model to achieve parameter-efficiency. We carry out extensive experiments to compare our proposed token mixing method with other parameter-efficient transfer learning methods. Our token mixing method outperforms other methods on both understanding tasks and generation tasks. Besides, our method achieves new records on multiple video-language tasks. The code is available at https://github.com/yuqi657/video_language_model.

Introduction

With the recent success of image-language pre-trained models (Li et al. 2022; Radford et al. 2021), many works (Luo et al. 2021; Fang et al. 2021; Liu et al. 2022; Wang et al. 2022b) have explored adopting an image-language pre-trained model to video-language tasks. These works usually face two challenges. One is how to effectively transfer knowledge learnt from 2D static images to 3D dynamic videos. The other is how to handle the expensive cost in model training or fine-tuning (e.g., BLIP-L/14 with 578M parameters), which would restrict the deployment of models in real-world applications. Therefore, parameter-efficient transfer learning from image-language to video-language has attracted much research attention recently.

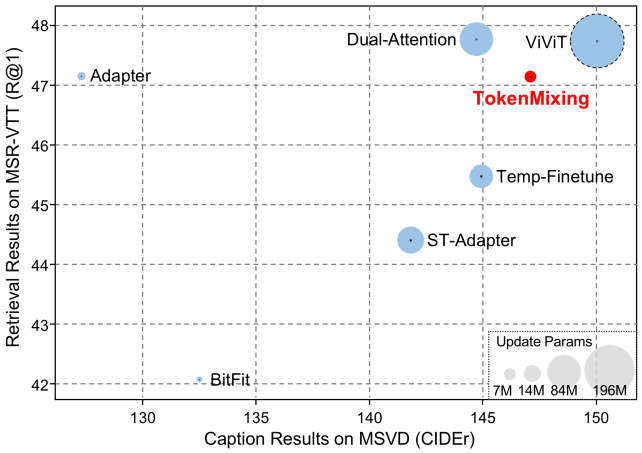

Figure 1: Different parameter-efficient tuning methods on video-language tasks. We compare our method with four partial fine-tuning methods including Dual-channel Attention (Hong et al. 2022), BitFit (Zaken, Ravfogel, and Goldberg 2021), ST-Adapter (Pan et al. 2022) and Adapter (Houlsby et al. 2019), Temporal Fine-tuning and a fully fine-tuning method ViViT (Arnab et al. 2021). Our method is effective in both video-text retrieval and video captioning tasks while has smaller updated parameters.

Motivated by works in the NLP area which have been proposed for parameter-efficient fine-tuning and achieved competitive performance by only fine-tuning a small set of parameters (Hong et al. 2022; Sung, Cho, and Bansal 2022; Pan et al. 2022), different approaches of parameter-efficient cross-modal transfer learning are proposed to leverage the power of pre-trained image-language models (Hong et al. 2022; Sung, Cho, and Bansal 2022; Pan et al. 2022). These works can be divided into two types: 1) appending a sequence of temporal transformer blocks after the original 2D Vision Transformer (ViT), or 2) inserting a temporal block into the original ViT architecture. Only the newly added

*This work was done when Yuqi was an intern at Tencent

†Corresponding Author

components need to be fine-tuned in both types of methods. For the first type, a straightforward way is to build several temporal transformer layers on top of the original spatial transformer layers. However, such methods introduce many new parameters and only achieve sub-optimal performance. For the second type, different plug-in components have been proposed. For example, VL-Adapter (Sung, Cho, and Bansal 2022) introduces three parameter-efficient tuning methods (Pan et al. 2022; Karimi Mahabadi, Henderson, and Ruder 2021; Mahabadi et al. 2021) for visual-language tasks. However, it treats image and video equally and only explores the cross-modal generation task. ST-Adapter (Pan et al. 2022) is further proposed to insert a 3D-Conv before the vision transformer block to fuse the spatio-temporal information in the video, and is verified on the video classification task. The newly designed component, however, still incurs many parameters (54M). CogVideo (Hong et al. 2022) inserts a temporal attention in parallel with the original spatial attention, which causes more parameters (85M). Furthermore, it is worth pointing out that all these methods are only validated on a single task, so their generalization ability has not been well verified.

In this work, we attempt to overcome both the aforementioned challenges. Our aim is to transfer knowledge from the pre-trained image-language model to video-language tasks in a parameter-efficient way without adding specific structure to the original architecture. We notice that the core idea of previous methods mentioned above is to exchange information between tokens across different frames. Rather than inserting additional components, we believe that cross-frame interactions can be achieved by designing certain token-level operations. To be specific, we propose a token mixing strategy to enhance cross-frame interaction in some layers of original 2D ViT, and only fine-tune these token mixed layers. We analyze several parameter-efficient training techniques and benchmark different methods that attempt to achieve parameter-efficient transfer learning, including adapter-base methods (Sung, Cho, and Bansal 2022; Pan et al. 2022), bias-only method (Zaken, Ravfogel, and Goldberg 2021) and architecture modification methods (Arnab et al. 2021; Hong et al. 2022). As shown in Fig. 1, our token mixing strategy with partially fine-tuning achieves a better trade-off between efficiency and performance. Compared to other methods, token mixing is more task-agnostic, and is effective in both video-language understanding (e.g. video retrieval) and video-language generation (e.g. video captioning). More specifically, with only 7M updated parameters in the visual encoder, Token Mixing Tuning achieves 146.92 CIDEr on MSVD and 47.1 R@1 on MSRVTT. These results demonstrate that a good cross-frame interaction strategy is effective to transfer image-language knowledge to video-language tasks in an efficient way. We hope our study will inspire future research on transferring pre-trained image-language model to video-language tasks in an architecture-efficient and parameter-efficient way. We summarize our contributions as follows:

- We propose a novel parameter-efficient strategy called Token Mixing to transfer image-language models to video-language tasks.
- We establish a benchmark for video-language tasks by comprehensively experimenting with a variety of fine-tuning methods.
- With only a minor adaption on the image-language pre-trained model, our token mixing method achieves superior or on par performance on multiple video-language tasks, including video captioning on MSRVTT, MSVD, and VATEX benchmarks and video retrieval on MSRVTT and LSMDC benchmarks.

Related Works

Tranfer Learning from Image to Video

Various works (Arnab et al. 2021; Bertasius, Wang, and Torresani 2021; Luo et al. 2021; Liu et al. 2022) have explored transferring knowledge from image pre-trained models to video. ViViT (Arnab et al. 2021) develops four pure-transformer architectures for video classification. X-ViT (Bulat et al. 2021) proposes space-time mixing attention to reduce computation and memory cost for video recognition task. CogVideo (Hong et al. 2022) uses a dual-attention mechanism to model temporal information. In this work, we opt for plain visual transformer architecture and design an effective space-time mixing attention schemes, to transfer knowledge from image to video.

Parameter-efficient Transfer Learning

Transfer Learning has been explored in different fields (Guo et al. 2020; Sun et al. 2022; Tang et al. 2022; Xiao et al. 2022) in past years. Recently, with the growing size of pre-trained models, parameter-efficient tuning has received increasing attentions. Parameter-efficient tuning approaches can be divided into several types: 1) **Vanilla adapter**. Adapter (Houlsby et al. 2019) is first introduced in NLP community, which aims to fine-tune a huge pre-trained model (i.e. GPT-3 (Brown et al. 2020)) efficiently. VL-Adapter (Sung, Cho, and Bansal 2022) further uses it in multi-modal tasks. ST-Adapter (Pan et al. 2022) replaces original Adapter layer with a 3D-Conv to enable spatio-temporal modeling ability. 2) **Low-rank factorization**. LoRA (Hu et al. 2021) proposes trainable rank decomposition matrices to reduce fine-tuning parameters. 3)**No-Adapter**. BitFit (Zaken, Ravfogel, and Goldberg 2021) shows that only fine-tuning bias term is effective.

Video-Language Task

Video-language tasks involve video-language understanding tasks (e.g. video retrieval) and video-language generation tasks (e.g. video captioning). Most existing approaches (Cao et al. 2022; Chen, Liu, and Albanie 2021; Wang et al. 2022b; Liu et al. 2021; Yang et al. 2021; Suin and Rajagopalan 2020; Hou et al. 2020; Chen, Liu, and Albanie 2021; Ryu et al. 2021; Lin, Gan, and Wang 2021; Xu et al. 2019; Chen et al. 2019; Zhang, Song, and Jin 2022) explore task specific modules for different tasks. For example, for the video retrieval task, HiT (Liu et al. 2021) and Hunyuan_tvr (Min et al. 2022) use a hierarchical matching strategy for cross-modal interaction. For the video captioning task, Open-book

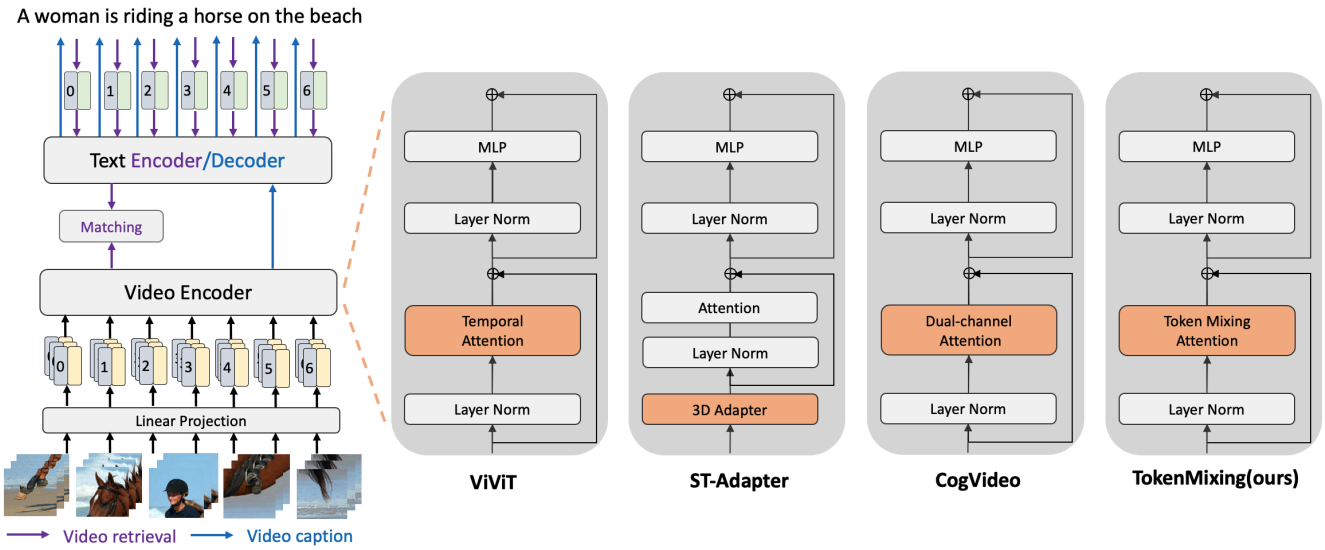

Figure 2: Architecture of Cross-modal model, which consists of three key components: the text encoder or decoder, the video encoder, and the text-video matching or text generation. The purple route is for the text-video retrieval task and the blue route is for the video captioning task. The video encoder can have different options as shown on the right, including ViViT (Arnab et al. 2021) , ST-Adapter (Pan et al. 2022) , CogVideo (Hong et al. 2022) and our proposed Token Mixing.

(Zhang et al. 2021) uses a retrieval-copy-generation strategy for generation. ORG-TRL (Zhang et al. 2020) uses object relational graph and teacher-recommended learning to enhance generation quality.

Apart from task specific modules, we believe a powerful video encoder can bring performance gain in any video-language tasks. Image-language pre-trained models (Li et al. 2022; Radford et al. 2021; Fei et al. 2022; Gu et al. 2022) have shown great power in image modeling. In this work, instead of designing task specific modules, we aim to utilize the power of pre-trained image-language model and transfer knowledge from pre-trained image encoder to video encoder with minimal adaptions.

Method

Our goal is to transfer large scale image-language pre-trained model to downstream video-language tasks in a parameter-efficient way. Specifically, we aim to effectively and efficiently inherit knowledge from the pre-trained image-language model with minimal adaptation while making it task-agnostic (i.e. it should be effective for both understanding and generation tasks). In principle, our proposed method can be applied to any image-language pre-trained models. We adopt BLIP (Li et al. 2022) for demonstration in this work. Our exploration may provide new insight in transfer learning from image-language to video-language.

Architecture

Cross-Modal Framework. The image-language works (Li et al. 2022; Radford et al. 2021; Fei et al. 2022) normally adopt the framework as shown in Fig. 2, which contains an visual encoder, a text encoder/decoder, and a task specific head (e.g. matching head for retrieval and language head for text generation). Both visual encoder and text encoder/decoder consist of consecutive layers of transformers. Our model follows such architecture as well.

Visual Transformer. Visual Transformer (ViT) (Dosovitskiy et al. 2021) is proposed for image modeling. Recent multi-modal pre-trained models (Li et al. 2022; Radford et al. 2021) usually choose a vanilla ViT as the image encoder. Formally, given a sequence of visual tokens $X_{in} = \{x_s | s = 0, 1, ..., S-1\}$, where S represents the sequence length, a transformer block computes its output X_{out} by a consecutive Multi-Head Attention (MHA) and Feed-Forward Network (FFN) as follows:

$$\widetilde{X} = X_{in} + \text{MHA}\left(\text{LN}\left(X_{in}\right)\right), \tag{1}$$

$$X_{\text{out}} = \widetilde{X} + \text{FFN}(\text{LN}(\widetilde{X})). \tag{2}$$

The MHA is computed as:

$$\widetilde{x}_s = \sum_{s'=0}^{S-1} \text{Softmax}\left\{\left(q_s \cdot k_{s'}\right)/\sqrt{D}\right\} v_{s'}, \tag{3}$$

where $q_s, k_s, v_s \in \mathbb{R}^D$ represent query, key, value projected from input x_s and D is the dimension of hidden states.

Temporal Information Integration. A vanilla ViT does not have the ability of cross-frame interaction. Several approaches attempt to integrate temporal information into ViT. Given a sequence of video tokens $X = \{x_{s,t} | s = 0, 1, ..., S-1, t = 0, 1, ..., T-1\}$, where S represents the sequence length and T represents the video length. A straight forward method is appending several temporal transformer blocks on top of the spatial transformer blocks, and only

these newly added transformer blocks need to be tuned. The MHA in these temporal transformer blocks is computed as:

$$\widetilde{x}_{s,t} = \sum_{t'=0}^{T-1} \text{Softmax}\left\{ \left(q_{s,t} \cdot k_{s,t'} \right) / \sqrt{D} \right\} v_{s,t'}, \quad (4)$$

Dual-channel attention (Hong et al. 2022) inserts a temporal MHA in parallel with spatial MHA, and a learnable parameter α is used to balance the spatial and temporal information, which is calculated as:

$$\begin{aligned} \tilde{x}_{s,t} = \alpha & \sum_{s'=0}^{S-1} \text{Softmax}\left\{ \left(q_{s,t} \cdot k_{s',t} \right) / \sqrt{D} \right\} v_{s',t} \\ +(1-\alpha) & \sum_{t'=0}^{T-1} \text{Softmax}\left\{ \left(q_{s,t} \cdot k_{s,t'} \right) / \sqrt{D} \right\} v_{s,t'}, \end{aligned} \quad (5)$$

and only the newly added temporal MHA need to be tuned.

ST-Adapter (Pan et al. 2022) inserts a 3D-Conv into the vanilla transformer block to integrate spatio-temporal information. The video input is firstly computed by the 3D-Conv, and then fed into a spatial only transformer, which is computed as:

$$X = X + \text{Conv3D}(X), \quad (6)$$

$$\widetilde{x}_{s,t} = \sum_{s'=0}^{S-1} \text{Softmax}\left\{ \left(q_{s,t} \cdot k_{s',t} \right) / \sqrt{D} \right\} v_{s',t}. \quad (7)$$

The core idea behind these methods is to make MHA directly or indirectly process tokens from other frames. However, all these methods introduce new components into the vanilla pre-trained model, and thus add many parameters that need to be updated. Besides, MetaFormer (Yu et al. 2022) provides us insight that the general architecture of transformers is the essential part. We therefore aim to find a more efficient way to let MHA attend to tokens from other frames without introducing much computation complexity.

Token Mixing Attention

To enable a token in a frame to interact with tokens from other frames more efficiently with minimal adaption of MHA, we propose a new token-level operation namely token mixing attention.

Looking closely at Eq. 4 and Eq. 5, we see that for these temporal information integrated visual encoders, each token $q_{s,t}$ can access to $k_{s',t'}$ and $v_{s',t'}$ from other frames, thus integrating temporal information via such cross-frame token interactions. However, we believe that cross-frame interactions can be achieved by designing certain token-level operations. Specifically, we select tokens from the video to form a key set $K = \{k_{s',t'} | s' \in [0, S-1] \cap \mathbb{Z}, t' \in [0, T-1] \cap \mathbb{Z}\}$ and a value set $V = \{v_{s',t'} | s' \in [0, S-1] \cap \mathbb{Z}, t' \in [0, T-1] \cap \mathbb{Z}\}$, where s' and t' are selected indices, as shown in Fig. 3. To avoid increasing the computation complexity, we restrict the number of elements in set K, V to

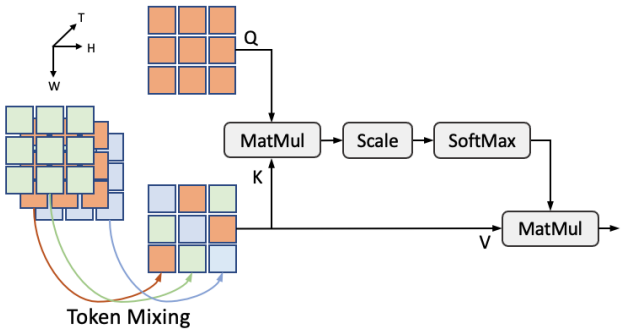

Figure 3: Illustration of token mixing attention operation. 'T, W, H' refer to video temporal dimension, width dimension and height dimension. We select tokens from adjacent frames to form the key set K and value set V.

equal to S. Then for each $q_{s,t}$, we compute the attention with the selected K, V set:

$$\widetilde{x}_{s,t} = \sum_{k_{s',t'} \in K, v_{s',t'} \in V} \text{Softmax}\left\{ \left(q_{s,t} \cdot k_{s',t'} \right) / \sqrt{D} \right\} v_{s',t'}, \quad (8)$$

In this way, without introducing extra components, token mixing attention empowers the model with the ability of cross-frame interaction.

Selection Strategy. Since we aim to make minimal adaption to MHA without adding any components, we do not design complex token selection module. We simply apply uniform sampling strategy. To be specific, for each position $s \in \{0, 1, ..., S-1\}$, we select token from $T (\text{mod } s)$ frame.

Parameter-Efficient Tuning. The main idea of our work is to transfer knowledge from image-language pre-trained model to video-language tasks in an efficient way. So we only modify one transformer layer and update parameters in this modified transformer layer.

Experiments

Experiment Settings

Datasets. We evaluate our model on video captioning tasks and video retrieval tasks. For **video captioning** task, we use widely adopted benchmarks, MSRVTT (Xu et al. 2016), VATEX (Wang et al. 2019), MSVD (Chen and Dolan 2011). For **video retrieval** task, we choose MSRVTT (Xu et al. 2016) and LSMDC (Rohrbach et al. 2017). The MSRVTT dataset contains 10,000 video clips with 20 captions per video. The VATEX dataset contains 34,991 video clips with several captions per video, with longer video duration and captions than MSRVTT. MSVD is relatively smaller, with only 1,970 videos and each video has about 40 captions. LSMDC is a dataset in the movie domain, which contains 118,081 video-text pairs extracted from 202 movies.

			MSVD				VATEX				MSRVTT			
Methods	UP	Mem	B4	M	R	C	B4	M	R	C	B4	M	R	C
Fully														
ViViT	226.5*	28.6	67.4	45.1	82.1	149.1	38.5	26.3	53.4	69.9	48.8	32.5	65.5	67.8
Partial														
Dual Attn	84.9	25.9	65.7	44.3	81.3	144.7	36.6	24.5	51.5	60.8	46.4	31.7	64.2	63.7
ST-Adapter	54.8	21.9	66.5	44.3	81.5	141.7	35.6	24.6	51.5	59.2	45.7	31.7	64.1	64.2
Temp FT	28.3	22.5	66.5	44.3	81.2	144.7	35.4	24.2	50.9	58.4	46.1	31.8	64.2	63.1
Adapter	1.9	21.3	59.9	42.7	78.8	127.3	32.1	22.8	49.1	50.8	36.9	29.9	59.9	52.1
BitFit	**0.3**	**18.2**	62.1	43.2	79.9	132.5	32.4	22.9	49.3	51.2	39.5	30.2	60.9	55.1
Mix (ours)	7.1	20.2	**67.0**	**44.9**	**82.2**	**146.9**	**38.5**	**26.1**	**53.2**	**69.5**	**48.9**	**32.9**	**65.5**	**67.9**

Table 1: Video captioning results of different fine-tuning methods on MSVD, VATEX and MSRVTT. 'Fully' represents 'Fully Fine-tuning', 'Partial' represents 'Partial Fine-tuning', 'Dual Attn' is short for 'Dual-channel Attention', 'Temp FT' represents 'Temporal Fine-tune', 'Mix' represents 'Token Mixing', 'UP' refers to 'number of Updated Parameters in the video encoder', and 'Mem' refers to 'Memory Usage per GPU', '*' indicates that parameters in the text decoder are also updated.

	Method	UP(M)	R@1	R@5	R@10
MSRVTT	**Fully**				
	ViViT	226.51*	47.6	73.4	81.8
	Partial				
	Temp Fine	28.31	45.5	70.6	79.8
	Adapter	1.88	47.1	**72.4**	79.4
	BitFit	0.26	42.1	64.8	75.3
	Mix (ours)	7.07	**47.1**	70.8	**80.5**
LSMDC	**Fully**				
	ViViT	226.51*	26.3	45.9	54.1
	Partial				
	Temp Fine	28.31	21.3	**40.2**	**50.1**
	Adapter	1.88	22.0	40.1	50.1
	BitFit	0.26	20.6	37.8	47.1
	Mix (ours)	7.07	**22.1**	39.9	48.8

Table 2: Video retrieval results of different fine-tuning methods on MSR-VTT and LSMDC. We only compare to methods with similar number of updated parameters due to limited space.

Evaluation Metrics. For video captioning, we choose BLEU-4 (B4) (Papineni et al. 2002), METEOR (M) (Banerjee and Lavie 2005), ROUGE (R) (Lin 2004), and CIDEr (C) (Vedantam, Lawrence Zitnick, and Parikh 2015) as evaluation metrics. For video retrieval, we choose Recall at K (R@K, higher is better) as the evaluation metrics. R@K calculates the fraction of correct videos among the top K retrieved videos. We use K=1,5,10 in our experiments.

Architecture Details. In most of our experiments, we choose BLIP(ViT-B/16) (Li et al. 2022) as our default base backbone model, where input frames are split into 16×16 patch sequence and then input to a 12-layer visual transformers. The text is encoded/decoded by a 12-layer transformer. The feature dimension of both text and video is 768. For video captioning, tokens from all frames are flatten and input to the cross-attention in the decoder. For video retrieval, we mean pool the [CLS] tokens from all frames to calculate the similarity score with the text representation. We use partially fine-tuning setting (i.e. only fine-tune layer with token mixing strategy) in the following experiments unless explicitly stated.

Other Baselines

Fully Fine-Tuning. Fully fine-tuning is the straight forward method when applying pre-trained models to downstream tasks. We build a few temporal transformer layers on top of the original spatial transformer layers, and their weights are initialized by the pre-trained spatial transformer, similar to the settings in ViViT (Arnab et al. 2021). In this setting, all parameters are updated.

Temporal Fine-tuning. The architecture of temporal fine-tuning is the same as fully fine-tuning. However, we only tune the parameters in the newly added blocks in this setting.

Dual-channel Attention. Like CogVideo (Hong et al. 2022), we only fine-tune the newly added temporal attention module. The weights of temporal attention are initialized by the spatial attention weights.

Adapter. VL-Adapter (Sung, Cho, and Bansal 2022) has explored Adapter and its variants in visual-language tasks. We insert adapter into each visual transformer layer and only tune the parameters in the adapter.

ST-Adapter. ST-Adapter(Pan et al. 2022) uses a 3D-Conv to fuse spatio-temporal information. We follow the setting of ST-Adapter.

Bias-only Fine-Tuning. It is proposed in BitFit (Zaken, Ravfogel, and Goldberg 2021). In bias-only fine-tuning, we only tune the bias in each module while keeping other parameters in the model frozen.

Parameter-Efficient Fine-tuning Benchmark

Tab. 1 and Tab. 2 present results of different fine-tuning methods on video captioning and video retrieval tasks, respectively. We compare the number of updated parameters in the visual encoder, and the performance of different methods. All models are initialized using BLIP (ViT-B/16), and we use the same setting (e.g. batch size) in all experiments for fair comparison. The linear projection layer and language

Video Captioning on MSVD					
	TM	B4	M	R	C
PF	×	65.7	44.3	81.6	146.3
	✓	**67.0**	**44.9**	**82.2**	**146.9**
FF	×	67.9	45.3	82.2	150.1
	✓	**68.2**	**45.6**	**82.4**	**151.2**

Text-to-Video Retrieval on MSRVTT					
	TM	R1	R5	R10	rsum
PF	×	46.1	70.1	79.6	195.8
	✓	**47.1**	**70.8**	**80.5**	**198.4**
FF	×	47.6	72.1	80.5	200.2
	✓	**48.5**	**72.3**	**81.5**	**202.3**

Table 3: Impact of token mixing on the performance of video-language tasks. FF represents 'Fully Fine-tune', 'PF' represents 'Partially Fine-tune', 'TM' represents 'Token Mixing', and 'rsum' is the sum of R@K.

head are both fine-tuned in all experiments since these modules are important in transfer learning as shown in the ablation study. We discuss our main observations as follows.

A simple token mixing strategy is sufficient. Among all compared parameter-efficient tuning methods, our token mixing strategy achieves a better trade-off between parameter efficiency and performance. It achieves the best performance on video captioning tasks against its competitors, and achieves comparable results on video retrieval tasks. Our token mixing method even performs on par with the fully fine-tuning method (146.9 vs. 149.1 on MSVD and 69.5 vs. 69.9 on VATEX) while updating far less parameters (7.07M vs. 226.51M) on video captioning tasks. We also find that fully fine-tuning is more useful in video retrieval tasks.

Spatio-temporal information is important. Temporal Fine-tune, ST-Adatper, Dual-Attention and Toking-Mixing all involve cross-frame interaction, thus they all have the capability to model spatio-temporal information. BitFit and Adapter do not have the temporal modeling ability. We observe that methods with spatio-temporal modeling capability perform better, indicating that temporal modeling is critical in transfer learning from image-language to video-language.

Ablation Study

Influence of Token Mixing Strategy. Tab. 3 ablates the effectiveness of token mixing strategy. We compare the token mixing strategy with the original spatial only vision transformer block under both fully and partially tuning settings. We observe that token mixing strategy improves the model performance consistently on both video captioning task and video retrieval task. The results show that our method is also beneficial under the fully fine-tuning setting.

Which module is more important? We explore how to be more parameter-efficient (i.e. which part in a transformer can be frozen) and the results are shown in Tab. 4. We observe that visual projection and language head play an important role in transfer learning. Thus if we only fine-tune the token mixing layer, the performance degrades significantly. Moreover, we observe that only fine-tuning the feed-forward network (FFN) part of token mixing layer, which corresponds to updating 66.2% parameters compared to fine-tuning the whole layer (4.7M vs. 7.1M), can achieve satisfactory performance. It inspires us that only fine-tuning the FFN part in a transformer layer may be sufficient for more parameter-efficiency.

		Token Mixing			
LP	LH	ATTN	FFN	UP	C
×	×	✓	✓	7.1	138.87
✓	×	✓	✓	7.1	141.15
×	✓	✓	✓	7.1	146.06
✓	✓	×	×	-	134.39
✓	✓	✓	×	2.4	144.45
✓	✓	×	✓	4.7	145.46
✓	✓	✓	✓	7.1	**146.92**

Table 4: Impact of fine-tuning different part on video captioning performance on MSVD. 'C' represents CIDEr for evaluation. 'LP' means 'linear projection', 'LH' represents 'language head', 'UP' represents 'number of Updated Parameters'

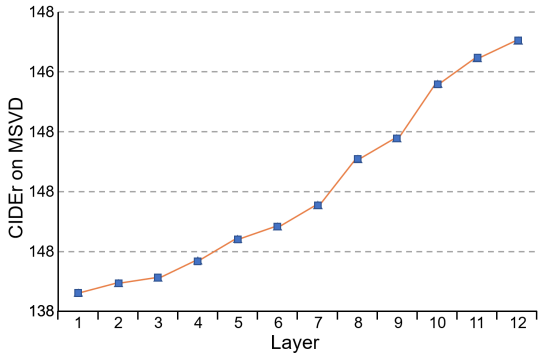

Figure 4: Impact of token mixing at different layers on video captioning. Experiments are done on MSVD.

Which layer is more important? We conduct experiments to explore plugging token mixing into which layer is more helpful. We only use our token mixing strategy in one layer in the experiments. As shown in Fig. 4, if we use token-mixing strategy in the deeper layer, the performance is better. We believe that the shallow layer is more important for modeling spatial information and performing cross-frame interaction in the deeper layer is more reasonable.

Comparison with SOTA

We compare our results with other state-of-the-art (SOTA) methods on both video captioning and video retrieval tasks. Both parameter-efficient fine-tuning and fully fine-tuning results of our method are compared in Tab. 5-6. For video captioning, we compare with ORG-TRL (Zhang et al. 2020), Open-book (Zhang et al. 2021), SwinBERT (Lin et al. 2022) and MVGPT (Seo et al. 2022). Please note that our method takes significantly fewer frames as input (e.g. 8 in our

	MSVD				VATEX				MSRVTT			
Methods	B4	M	R	C	B4	M	R	C	B4	M	R	C
ORG-TRL	54.30	36.40	73.90	95.20	-	-	-	-	43.60	28.80	62.10	50.90
Open-book	-	-	-	-	33.90	23.70	50.20	57.50	42.80	29.30	61.70	52.90
MV-GPT	-	-	-	-	-	-	-	-	48.92	38.66	64.00	60.00
SwinBERT	66.30	42.40	80.90	149.40	38.70	26.20	53.20	73.00	45.40	30.60	64.10	55.90
Token Mixing (P)	67.04	44.93	82.15	146.92	38.46	26.12	53.18	69.49	48.96	32.88	65.52	67.97
Token Mixing (F)	68.15	45.57	82.43	151.19	38.60	26.17	53.33	69.63	**49.40**	**33.01**	**65.79**	68.60
Token Mixing-L (P)	70.83	46.89	84.09	160.50	39.12	25.85	53.25	69.04	47.95	32.53	65.39	68.21
Token Mixing-L (F)	**71.16**	**47.44**	**84.22**	**162.71**	**39.90**	**26.47**	**53.84**	**73.44**	48.73	32.84	65.73	**69.48**

Table 5: Comparison to other SOTA methods on video captioning tasks. '-L' represents 'ViT-Large', 'P' represents 'partially fine-tune' and 'F' represents 'fully fine-tune'.

MSRVTT text-to-video retrieval				
	Method	R@1	R@5	R@10
ViT-B	OmniVL	47.8	74.2	83.8
	TS2-NET	49.4	**75.6**	**85.3**
	Hunyuan_tvr	**49.7**	75.0	83.5
	Token Mixing (P)	47.1	70.8	80.5
	Token Mixing (F)	48.5	72.3	81.5
ViT-L	Hunyuan_tvr-L	49.5	74.2	**83.9**
	Token Mixing-L (P)	50.5	**74.2**	82.0
	Token Mixing-L (F)	**51.4**	72.9	80.5

LSMDC text-to-video retrieval				
	Method	R@1	R@5	R@10
ViT-B	TS2-NET	23.4	42.3	50.9
	Hunyuan_tvr	24.5	**44.1**	53.9
	Token Mixing (P)	22.1	39.9	48.8
	Token Mixing (F)	**25.3**	43.8	**54.0**
ViT-L	Hunyuan_tvr-L	27.1	45.1	53.4
	Token Mixing-L (P)	25.0	43.3	51.3
	Token Mixing-L (F)	**27.4**	**48.7**	**57.1**

Table 6: Comparison to other SOTA methods on video retrieval tasks. '-L' represents 'ViT-Large', 'P' represents 'partially fine-tune' and 'F' represents 'fully fine-tune'. All results reported without inverted softmax.

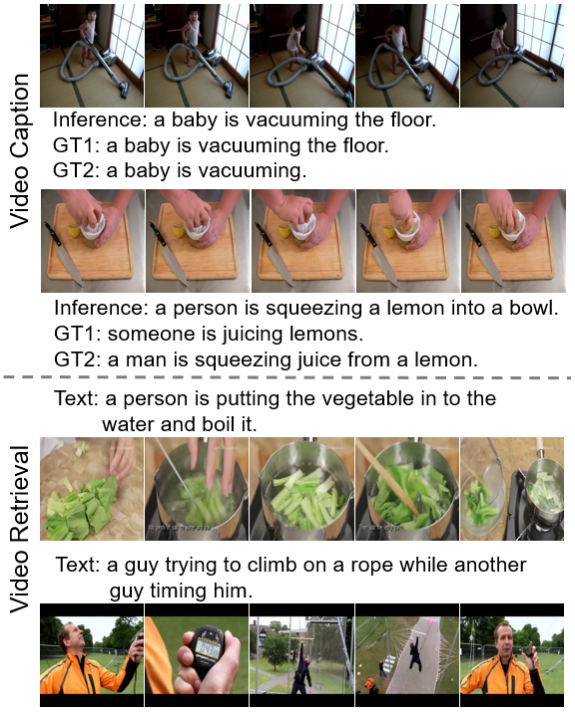

Figure 5: Cases of video captioning and video retrieval.

method vs. 64 in SwinBERT (Lin et al. 2022)). For video retrieval, we compare with TS2-Net (Liu et al. 2022), Hunyuan_tvr (Min et al. 2022), and video-language pre-trained model OmniVL(Wang et al. 2022a). All methods use the same frame resolution.We observe that with minor adaptions and only a small part of parameters updated, our method still achieves comparable performance. When fully fine-tuning our model with token mixing strategy, the performance gains consistently and achieves state-of-the-art results.

Qualitative Results

Some caption and retrieval cases are shown in Fig. 5. Our model can recognize visual objects (e.g. baby, lemon, vegetable, rope) and actions (e.g. vacuum, squeeze, boil, climb) to ensure correct caption generation or video retrieval.

Conclusion

In this paper, we address parameter-efficient transfer learning from image-language to video-language. We analyze different parameter-efficient methods at first, and then propose a novel parameter-efficient strategy called token mixing without introducing extra components into the pre-trained model structure. We benchmark several parameter-efficient fine-tuning methods on video-language tasks. The experimental results demonstrate that our token mixing strategy achieves the best trade-off of efficiency and performance. With only a minor adaption on the image-language pre-trained model, our token mixing method achieves new records on multiple video captioning benchmarks, including MSRVTT, MSVD, VATEX, and achieves comparable results on video retrieval tasks, including MSRVTT, LSMDC.

Acknowledgments

This work was partially supported by National Key R&D Program of China (No. 2020AAA0108600) and National Natural Science Foundation of China (No. 62072462).

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *ICCV*, 6836–6846.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Bulat, A.; Perez Rua, J. M.; Sudhakaran, S.; Martinez, B.; and Tzimiropoulos, G. 2021. Space-time mixing attention for video transformer. *NeurIPS*.

Cao, S.; Wang, B.; Zhang, W.; and Ma, L. 2022. Visual Consensus Modeling for Video-Text Retrieval. In *AAAI*.

Chen, D.; and Dolan, W. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *ACL*. Association for Computational Linguistics.

Chen, J.; Pan, Y.; Li, Y.; Yao, T.; Chao, H.; and Mei, T. 2019. Temporal deformable convolutional encoder-decoder networks for video captioning. In *AAAI*, 8167–8174.

Chen, Q.; Liu, Y.; and Albanie, S. 2021. Mind-the-Gap! Unsupervised Domain Adaptation for Text-Video Retrieval. In *AAAI*, 1072–1080.

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. In *ICLR*.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Fei, N.; Lu, Z.; Gao, Y.; Yang, G.; Huo, Y.; Wen, J.; Lu, H.; Song, R.; Gao, X.; Xiang, T.; et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1): 1–13.

Gu, J.; Meng, X.; Lu, G.; Hou, L.; Niu, M.; Xu, H.; Liang, X.; Zhang, W.; Jiang, X.; and Xu, C. 2022. Wukong: 100 Million Large-scale Chinese Cross-modal Pre-training Dataset and A Foundation Framework. *arXiv preprint arXiv:2202.06767*.

Guo, Y.; Li, Y.; Wang, L.; and Rosing, T. 2020. Adafilter: Adaptive filter fine-tuning for deep transfer learning. In *AAAI*, 4060–4066.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*.

Hou, J.; Wu, X.; Zhang, X.; Qi, Y.; Jia, Y.; and Luo, J. 2020. Joint commonsense and relation reasoning for image and video captioning. In *AAAI*, 10973–10980.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *ICML*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Karimi Mahabadi, R.; Henderson, J.; and Ruder, S. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *NeurIPS*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Lin, K.; Gan, Z.; and Wang, L. 2021. Augmented partial mutual learning with frame masking for video captioning. In *AAAI*, 2047–2055.

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 *CVPR*, 17949–17958.

Liu, S.; Fan, H.; Qian, S.; Chen, Y.; Ding, W.; and Wang, Z. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *ICCV*, 11915–11925.

Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; and Jin, Q. 2022. TS2-Net: Token Shift and Selection Transformer for Text-Video Retrieval. In *ECCV*.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2021. CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval. *arXiv preprint arXiv:2104.08860*.

Mahabadi, R. K.; Ruder, S.; Dehghani, M.; and Henderson, J. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *ACL*.

Min, S.; Kong, W.; Tu, R.-C.; Gong, D.; Cai, C.; Zhao, W.; Liu, C.; Zheng, S.; Wang, H.; Li, Z.; et al. 2022. HunYuan_tvr for Text-Video Retrivial. *arXiv preprint arXiv:2204.03382*.

Pan, J.; Lin, Z.; Zhu, X.; Shao, J.; and Li, H. 2022. ST-Adapter: Parameter-Efficient Image-to-Video Transfer Learning for Action Recognition. *arXiv preprint arXiv:2206.13559*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rohrbach, A.; Torabi, A.; Rohrbach, M.; Tandon, N.; Pal, C.; Larochelle, H.; Courville, A.; and Schiele, B. 2017. Movie description. *IJCV*, 94–120.

Ryu, H.; Kang, S.; Kang, H.; and Yoo, C. D. 2021. Semantic grouping network for video captioning. In *AAAI*, 2514–2522.

Seo, P. H.; Nagrani, A.; Arnab, A.; and Schmid, C. 2022. End-to-end generative pretraining for multimodal video captioning. In *CVPR*, 17959–17968.

Suin, M.; and Rajagopalan, A. 2020. An efficient framework for dense video captioning. In *AAAI*, 12039–12046.

Sun, J.; Wei, D.; Ma, K.; Wang, L.; and Zheng, Y. 2022. Boost Supervised Pretraining for Visual Transfer Learning: Implications of Self-Supervised Contrastive Representation Learning. In *AAAI*.

Sung, Y.-L.; Cho, J.; and Bansal, M. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*, 5227–5237.

Tang, Y.; Kang, X.; Li, C.; Lin, Z.; and Ming, A. 2022. Transfer Learning for Color Constancy via Statistic Perspective. In *AAAI*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Wang, J.; Chen, D.; Wu, Z.; Luo, C.; Zhou, L.; Zhao, Y.; Xie, Y.; Liu, C.; Jiang, Y.-G.; and Yuan, L. 2022a. Omnivl: One foundation model for image-language and video-language tasks. *arXiv preprint arXiv:2209.07526*.

Wang, Q.; Zhang, Y.; Zheng, Y.; Pan, P.; and Hua, X.-S. 2022b. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 4581–4591.

Xiao, A.; Huang, J.; Guan, D.; Zhan, F.; and Lu, S. 2022. Transfer learning from synthetic to real LiDAR point cloud for semantic segmentation. In *AAAI*, 2795–2803.

Xu, H.; He, K.; Plummer, B. A.; Sigal, L.; Sclaroff, S.; and Saenko, K. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 9062–9069.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 5288–5296.

Yang, B.; Zou, Y.; Liu, F.; and Zhang, C. 2021. Non-autoregressive coarse-to-fine video captioning. In *AAAI*, 3119–3127.

Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; and Yan, S. 2022. Metaformer is actually what you need for vision. In *CVPR*, 10819–10829.

Zaken, E. B.; Ravfogel, S.; and Goldberg, Y. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *ACL*.

Zhang, Q.; Song, Y.; and Jin, Q. 2022. Unifying Event Detection and Captioning as Sequence Generation via Pre-training. In *ECCV*, 363–379.

Zhang, Z.; Qi, Z.; Yuan, C.; Shan, Y.; Li, B.; Deng, Y.; and Hu, W. 2021. Open-book video captioning with retrieve-copy-generate network. In *CVPR*, 9837–9846.

Zhang, Z.; Shi, Y.; Yuan, C.; Li, B.; Wang, P.; Hu, W.; and Zha, Z.-J. 2020. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, 13278–13288.