arXiv:2209.15280v3 [cs.CV] 12 Mar 2023

Learning Transferable Spatiotemporal Representations from Natural Script Knowledge

Ziyun Zeng[1,2 *] Yuying Ge[3 *] Xihui Liu[3]
Bin Chen[4 ✉] Ping Luo[3] Shu-Tao Xia[1] Yixiao Ge[2 ✉]
[1] Tsinghua University [2] Applied Research Center (ARC), Tencent PCG
[3] The University of Hong Kong [4] Harbin Institute of Technology, Shenzhen
[*] equal contribution ✉ corresponding authors
zengzy21@mails.tsinghua.edu.cn yuyingge@hku.hk xihuiliu@eee.hku.hk
chenbin2021@hit.edu.cn pluo@cs.hku.hk xiast@sz.tsinghua.edu.cn yixiaoge@tencent.com

Abstract

*Pre-training on large-scale video data has become a common recipe for learning transferable spatiotemporal representations in recent years. Despite some progress, existing methods are mostly limited to highly curated datasets (*e.g.*, K400) and exhibit unsatisfactory out-of-the-box representations. We argue that it is due to the fact that they only capture pixel-level knowledge rather than spatiotemporal semantics, which hinders further progress in video understanding. Inspired by the great success of image-text pre-training (*e.g.*, CLIP), we take the first step to exploit language semantics to boost transferable spatiotemporal representation learning. We introduce a new pretext task, Turning to Video for Transcript Sorting (TVTS), which sorts shuffled ASR scripts by attending to learned video representations. We do not rely on descriptive captions and learn purely from video,* i.e.*, leveraging the natural transcribed speech knowledge to provide noisy but useful semantics over time. Our method enforces the vision model to contextualize what is happening over time so that it can re-organize the narrative transcripts, and can seamlessly apply to large-scale uncurated video data in the real world. Our method demonstrates strong out-of-the-box spatiotemporal representations on diverse benchmarks,* e.g.*, +13.6% gains over VideoMAE on SSV2 via linear probing. The code is available at* https://github.com/TencentARC/TVTS.

1. Introduction

The aspiration of representation learning is to encode general-purpose representations that transfer well to diverse downstream tasks, where self-supervised methodologies [9, 25] dominate due to their advantage in exploiting large-scale unlabeled data. Despite significant progress in learning representations of still images [23, 45], the real world is dynamic and requires reasoning over time. In this paper, we focus on *out-of-the-box spatiotemporal representation learning*, a more challenging but practical task towards generic video understanding, which aims to capture hidden representations that can be further used to conduct reasoning on broader tasks, *e.g.*, classification and retrieval.

There have been various attempts at self-supervised pre-training on video data from discriminative learning objectives [5, 8, 28] to generative ones [17, 51], where the core is context capturing in spatial and temporal dimensions. Though promising results are achieved when transferring the pre-trained models to downstream video recognition [22, 34, 50] via fine-tuning, the learned representations are still far away from out-of-the-box given the poor linearly probing results (see Figure 1(a)). Moreover, existing works mostly develop video models on the highly curated dataset with particular biases, *i.e.*, K400 [32]. Their applicability in the real world is questioned given the observed performance drops when training on a larger but uncurated dataset, YT-Temporal [57]. We argue that, to address the above issue, the rich spatiotemporal semantics contained in the video itself should be fully exploited. But current video models generally exploit visual-only perception (*e.g.*, pixels) without explicit semantics.

Recently, the success of CLIP [45] has inspired the community to learn semantically aware image representations that are better transferable to downstream tasks and scalable to larger uncurated datasets. It provides a feasible solution for improving spatiotemporal representation learning but remains two key problems. (1) The vision-language contrastive constraints in CLIP mainly encourage the understanding of static objects (noun contrast) and simple motions (verb contrast), while how to enable long-range tem-

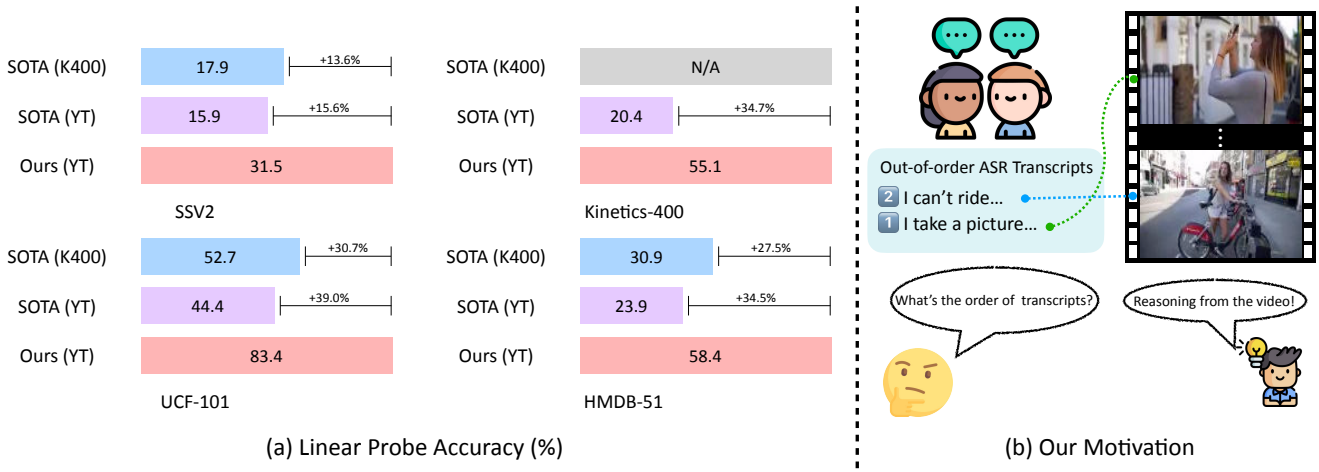


Figure 1. (a) We evaluate the transferability of spatiotemporal representations via linear probing on four video recognition datasets [22, 32, 34, 50], where the state-of-the-art method [51] underperforms. It performs even worse when pre-trained with a large-scale uncurated dataset, YT-Temporal [57]. (b) We encourage complex temporal understanding and advanced spatiotemporal representation learning with a new pretext task of sorting transcripts.

poral understanding with language supervision needs to be studied. (2) The quality of language supervision [49] is critical to the final performance of CLIP, however, it is hard to collect large-scale video data with literal captions that carefully describe the dynamic content over time. The ideal way for self-supervised learning is to learn useful knowledge purely from the data itself, which is also the philosophy followed by previous video pre-training methods [17, 51]. Fortunately, video data is naturally multi-modal with transcribed speech knowledge in the form of text (ASR), providing time-dependent semantics despite some noise.

To facilitate spatiotemporal understanding in large-scale uncurated data under the supervision of inherent script knowledge, we introduce a new pretext task for video pre-training, namely, **T**urning to **V**ideo for **T**ranscript **S**orting (TVTS). Intuitively, people sort out the order of events by temporal reasoning. As illustrated in Figure 1(b), given several unordered transcripts, it is difficult to reorganize the narrative by merely understanding the literal semantics. When the corresponding video is provided, it will be much easier to sort the transcripts by contextualizing what is happening over time. Whereas in neural networks, the temporal inference is embedded in spatiotemporal representations. Thus we believe that if the chronological order of transcripts can be correctly figured out via resorting to the correlated video representations, the video has been well understood.

We realize the pretext task of TVTS by performing joint attention among the encoded video spatiotemporal representations and the extracted ASR transcript representations. Specifically, given an input video and its successive transcripts, we randomly shuffle the order of the sentences. Subsequently, we concatenate the encoded script representations and the video representations and perform self-attention to predict the actual orders of the shuffled transcripts by fully understanding the spatiotemporal semantics in the video. The order prediction is cast as a K-way classification task, where K is the number of transcripts. The pretext task indirectly regularizes our model to properly capture contextualized spatiotemporal representations to provide enough knowledge for transcript ordering.

The usage of language supervision is related to video-text alignment [4, 20] and multimodal representation learning [18, 57] methods, however, we are completely different. (1) Video-text alignment methods focus on retrieval tasks and are devoted to associating the vision patterns with language concepts. They are generally single-frame biased [35] and fail to encode strong out-of-the-box temporal representations. (2) Multimodal representation learning methods aim to learn fused representations across modalities rather than vision-only spatiotemporal representations in our work. Moreover, different from our pretext task that aims to optimize spatiotemporal video representations, [57] sorts video frames by taking the features of individual frames as inputs without temporal modeling, *i.e.*, learning video representations only at the image level. As [57] points out, its ordering pretext task is not critical for downstream tasks (performance even drops) and primarily serves as an interface to query the model about temporal events.

To summarize, our contributions are three-fold. (**i**) We exploit the rich semantics from script knowledge which is naturally along with the video, rendering a flexible pre-training method that can easily apply to uncurated video data in the real world. (**ii**) We introduce a novel pre-

text task for video pre-training, namely, Turning to Video for Transcript Sorting (TVTS). It promotes the capability of the model in learning transferable spatiotemporal video representations. (**iii**) We conduct comprehensive comparisons with advanced methods. Our pre-trained model exhibits strong out-of-the-box spatiotemporal representations on downstream action recognition tasks, especially the relatively large-scale and the most challenging SSV2 [22]. We also achieve state-of-the-art performances on eight common video datasets in terms of fine-tuning.

2. Related Work

Spatiotemporal representation learning. Dominant video representation learning works have two categories, *i.e.*, discriminative- and generative-based methods. **(i)** The discriminative-based methods aim at mining unique representations within videos. For example, SVT [46] aligns several views from the same video with different spatial and temporal resolution for video-invariant representations. RSPNet [8], ASCNet [28], and LongShortView [5] utilize the appearance and temporal consistency of videos as the supervision. They use different augmentations of videos to construct positive and negative pairs to learn correspondences along the spatial and temporal dimensions. **(ii)** The generative-based methods try to reconstruct visual information from corrupted inputs. For example, MAE-based [24] methods [17, 51] use pixel values of video frames as supervision by masking raw videos with an extremely high ratio and reconstructing them.

Previous works are mainly trained on highly curated datasets, *e.g.*, Kinetics-400, HMDB51, and UCF101, where the temporal motions are not significant [35]. This leads to a "spatial bias", thus weakening the transferability to real-world uncurated datasets due to the lack of long-term temporal reasoning. Besides, existing works merely use visual supervision without explicit semantic information. Compared to them, our work leverages natural language derived from the video itself, *i.e.*, the ASR transcripts, as the supervision. Benefiting from the rich spatiotemporal information, our learned video representations have stronger transferability to downstream tasks.

Video-text pre-training. Existing video-text pre-training work can be divided into two categories. The first category aims to learn video-text alignment for retrieval. For example, Frozen [4], MCQ [20], and MILES [21] generally adopt two separate encoders to extract video and text representations, then align them with contrastive loss. However, they only align videos with a global video caption, thus neglecting the fine-grained temporal information. Furthermore, they rely on clean captions, which are difficult to scale up, and it is actually hard to collect large-scale video data with captions describing the dynamic content over time. The second category works on joint representation learning across modalities mainly for VQA. For example, MERLOT [57] adopts a joint encoder to match the captions with the corresponding video frames and put scrambled video frames into the correct order. It aims to match different modalities in the temporal dimension to achieve multi-modality fusion in a joint encoder, rather than learn better spatiotemporal representations.

Image representation learning by language supervision. Recently, there have been a bunch of successful tries in utilizing language supervision to enhance image representation learning. For example, CLIP [45] utilized 400M image-text pairs collected from the Internet and adopt the contrastive loss to align the image and its corresponding text. The superior performance on downstream image classification tasks revealed that learning directly from the raw text about images is a promising alternative that leverages a much broader source of supervision. ALIGN [31], uses a larger but noisier uncurated dataset and shows similar results to CLIP. Nevertheless, these methods only utilize language supervision to improve spatial learning, without exploring temporal learning, which hinders them from properly learning out-of-the-box video representations.

3. Method

In this work, we introduce a novel pretext task, **Turning to Video for Transcript Sorting (TVTS)** to learn the transferable spatiotemporal video representation by leveraging the rich semantics from script knowledge. In this section, we first introduce the pretext TVTS in Sec. 3.1 and our pre-training objectives in Sec. 3.2. We then describe the model architecture in Sec. 3.3.

3.1. Turning to Video for Transcript Sorting

As shown in Fig. 2, we perform the pretext task of TVTS to learn transferable spatiotemporal representations of videos. Given the observation that it will be much easier to sort the ASR transcripts by contextualizing what is happening over time in the video, we first randomly shuffle several consecutive ASR transcripts and extract their representations. We then perform joint attention among the transcript representations and video representations to sort the transcripts in the correct order via capturing contextualized spatiotemporal representations of videos.

Sample and Shuffle. Given a video V and its corresponding ASR transcripts with word-level timestamps $\{(w_i, s_i)\}_{i=1}^{N_{\text{asr}}}$, where N_{asr} denotes the word number, w_i and s_i denote the i-th word and its timestamp respectively, we randomly choose a starting time s_{begin} and sample K consecutive transcripts, each with a duration of l (in seconds), and an interval of 1s between adjacent transcripts,

$$\begin{aligned} S_k &= s_{\text{begin}} + (k-1) * (l+1), \quad E_k = S_k + l \\ T_k &= \{w_i | S_k \leq s_i \leq E_k\}, \quad k \in \{1, \cdots, K\} \end{aligned} \tag{1}$$

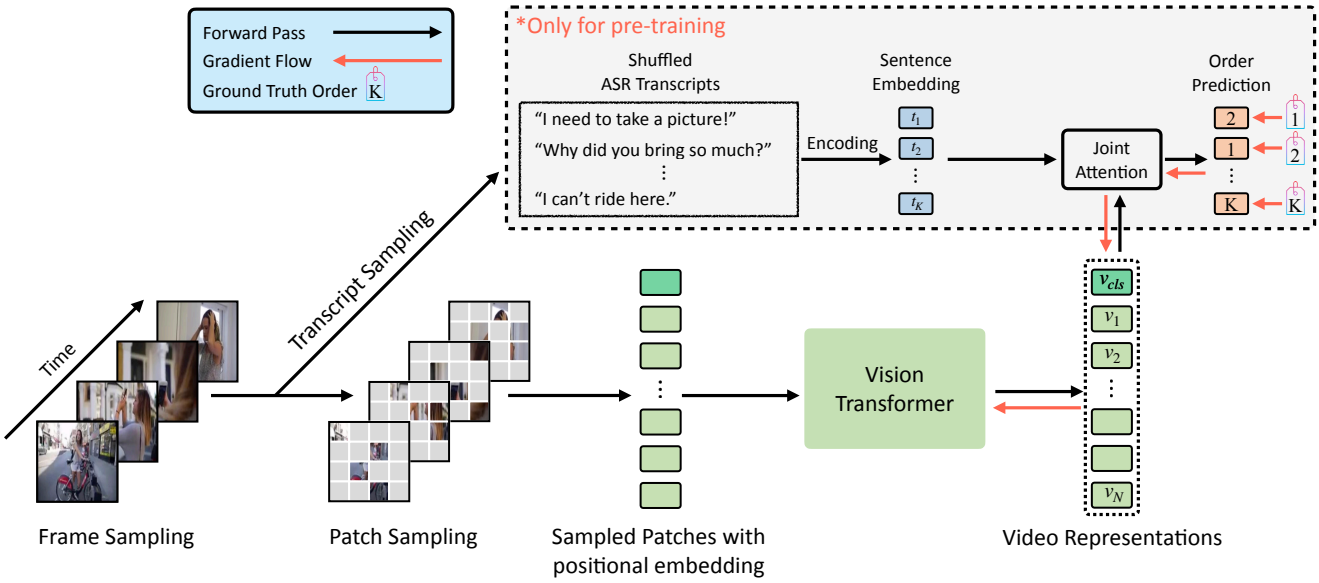


Figure 2. Our pre-training pipeline. We first sample K consecutive ASR transcripts, and a video clip consisting of M frames within the span of the transcripts. We randomly sample frame patches as the input of a vision transformer for the video representations. We then shuffle the transcripts and extract the representation of each transcript. We perform joint attention among the transcript and the video representations to predict the actual order of each transcript, which is optimized with a cross-entropy objective.

where S_k and E_k denote the beginning and ending time of the k-th transcript. We consecutively sample K transcripts with an interval of 1s and collect all words within $[S_k, E_k]$ for the k-th transcript. Finally, we randomly shuffle the transcripts as $\{T_{o_i}\}_{i=1}^{K}$, which means that the i-th transcript in this shuffled sequence is actually the o_i-th transcript in the original ordered sequence.

As for the video, we sample a clip between the beginning and ending time of all K transcripts, *i.e.*, $[S_1, E_K]$, which contains M frames as $\{F_i\}_{i=1}^{M}$. Specifically, we follow TSN [54] to divide $[S_1, E_K]$ into M segments with equal length and randomly sample 1 frame from each segment. After that, we get a video clip with M frames and K shuffled transcripts along the span of the video clip.

Sorting Transcripts. Given the shuffled transcripts $\{T_{o_i}\}_{i=1}^{K}$ and the corresponding video clip $\{F_i\}_{i=1}^{M}$, we first feed the transcripts in parallel to encode *unordered* text representations $\{t_{o_i}\}_{i=1}^{K}$. We then randomly sample the frame patches by masking a large proportion of the video clip among the spatial and temporal dimension as the input of a vision transformer to encode video representations $\{v_j\}_{j=0}^{N}$, where N denotes the number of the unmasked video patches, and v_0 is the representation of the [CLS] token. It is worth noting that *we do not add the extra [MASK] token, and we have no explicit reconstruction target*, which is different from previous works [17, 51]. We sample video frame patches as a means of data augmentation since it provides corrupted knowledge for our model to perform the pretext task of TVTS. Such a strategy also reduces the computational cost during pre-training as the attention is calculated on fewer patches.

We then concatenate the text representations of the shuffled transcripts $\{t_{o_i}\}_{i=1}^{K}$ and the video representations of the sampled video clip $\{v_j\}_{j=0}^{N}$, and perform multi-head self-attention among them. Our model attempts to sort the transcripts in the correct order by attending to the text features of all transcripts and the visual features of the unmasked video clip. We model the prediction of the transcript orders as a K-way classification task for each transcript. The first K output representations after the joint attention are further fed into a linear classifier to predict the order $p \in \mathbb{R}^K$, where p_j denotes the probability that the transcript is the j-th transcript in the original ordered sequence. For the transcript T_{o_i}, the ground truth classification label should be o_i.

The pretext task of TVTS regularizes the model to contextualize what is happening over time, so that it can provide enough knowledge for our model to figure out the chronological order of the shuffled transcripts. It improves the capability of the model to learn spatiotemporal representations that can be transferred to downstream tasks. We compare our method with other works that also adopt an ordering-based pretext task for pre-training in Sec. 4.5.

3.2. Pre-training Objectives

Besides the pretext task of TVTS, we use a global video-transcript contrastive objective. It aligns the features of the video clip and the averaged features of K transcripts so that

the video and transcript representations are in the same feature space for performing the joint attention to predict transcript orders. We combine two objectives to optimize the entire model in an end-to-end manner.

The first one is a cross-entropy objective $\mathcal{L}_{\text{sort}}$, which supervises our model to predict the correct order of the transcripts, and is formulated as below,

$$\begin{aligned} \mathcal{L}_{\text{sort}} &= -\frac{1}{K}\sum_{i=1}^{K}\log \text{softmax}(\hat{p}^i) \\ s.t. \quad \text{softmax}(\hat{p}^i) &= \frac{\exp(p^i_{o_i})}{\sum_{j=1}^{K}\exp(p^i_j)}, \end{aligned} \tag{2}$$

where p^i_j denotes the probability that the *i*-th transcript in the shuffled sequence is the *j*-th transcript in the original ordered sequence and o_i is the ground truth order in the original ordered sequence.

The second one is the global video-transcript contrastive objective $\mathcal{L}_{\text{base}}$, formulated as a bidirectional InfoNCE [42],

$$\begin{aligned} \mathcal{L}_{\text{base}} &= \text{NCE}(\hat{t},\hat{v}) + \text{NCE}(\hat{v},\hat{t}) \\ s.t. \quad \text{NCE}(q,k) &= -\log\frac{\exp(q^\top k_+/\tau)}{\sum_{i=1}^{B}\exp(q^\top k_i/\tau)}, \end{aligned} \tag{3}$$

where $\hat{t}$ and $\hat{v}$ denote the global text and video representation. We average the [CLS] token representation of all K transcripts as $\hat{t}$, *i.e.*, $\hat{t} \leftarrow \frac{1}{K}\sum_{i=1}^{K} t_i$, and use the [CLS] token representation of the video clip as $\hat{v}$, *i.e.*, $\hat{v} \leftarrow v_0$.

Our overall pre-training objective combines the two objectives, *i.e.*, $\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda\mathcal{L}_{\text{sort}}$, where λ is a hyper-parameter to balance the two losses. In our implementation, we set $\lambda = 2$ to roughly scale the gradient magnitudes of $\mathcal{L}_{\text{base}}$ and $\mathcal{L}_{\text{sort}}$ to be the same for efficient training.

3.3. Model Architecture

The vision transformer takes a video clip as input, which consists of M frames of resolution $H \times W$, and outputs video representations. We follow [51] to adopt cube embeddings, where each token corresponds to a cube of size $2 \times 16 \times 16$. This yields $\frac{M}{2} \times \frac{H}{16} \times \frac{W}{16}$ 3D tokens. Then we add divided space-time embedding to the token sequence, where tokens within the same frame obtain the same temporal embedding, and tokens within the same spatial location of different frames obtain the same spatial embedding. In this way, the vision transformer learns the positional information of the cubes. Next, we follow BERT [12] to add a learnable [CLS] token at the beginning of the token sequence for global video representations. Then we mask a portion of video tokens without [MASK] token replacement, as stated in Sec. 3.1. We adopt a standard ViT [15] architecture to encode video representations. The unmasked N video tokens as well as the [CLS] token are fed into the vision transformer, and joint space-time attention [3] is performed among the whole unmasked token sequence.

We use a DistilBERT [48] to extract the representations of ASR transcripts. We adopt two stacked bidirectional transformer blocks to predict the order of each transcript by performing joint attention among the transcript and the video representations. Within each block, multi-head self-attention is performed among all the video and text tokens, *i.e.*, all transcript-video tokens interact with each other.

4. Experiments

4.1. Pre-training Datasets

We pre-train our model on the large-scale **YT-Temporal** dataset [57] containing 6M YouTube videos with ASR transcripts and word-level timestamps.

4.2. Downstream Tasks

Action Recognition. We evaluate our pre-trained model on four common video datasets: (a) **Something-Something V2** (SSV2) [22], (b) **Kinetics-400** (K400) [32], (c) **UCF-101** [50], (d) **HMDB-51** [34]. Our evaluation is two-fold: (i) We conduct *zero-shot* **video-to-video retrieval** and *linear probe* **classification** on SSV2 to evaluate the transferability of the learned video representation. The former aims to retrieve videos of the same category as a query video, and the latter freezes the visual encoder and only optimizes a linear classifier. (ii) We *fully fine-tune* our pre-trained model on the training set of the four datasets to evaluate the action recognition capability. See Appendix for details.

Text-to-Video Retrieval. Beyond action recognition, we further evaluate retrieval performance on four benchmarks to see if the improved semantic-aware video representation can benefit retrieval tasks: (a) **MSR-VTT** [56] (b) **DiDeMo** [2] (c) **MSVD** [7] (d) **LSMDC** [47]. We adopt Recall@K (R@K) and Median Rank (MedR) as the evaluation metric. See Appendix for details.

4.3. Implementation Details

We follow recent works [4, 20] to adopt the pre-trained DistilBERT [48] to extract transcript representations. The vision transformer is a vanilla ViT-Base [15] with patch size P=16 and hidden state dimension D=768, and is initialized with ImageMAE-Base [23]. We set the temperature parameter τ to be 0.05. We pre-train our model on the YT-Temporal dataset sampling 16 frames for 20 epochs. We randomly mask 75% tokens within each frame. The input frame is first resized to 256×256, then we apply RandomCrop during training and CenterCrop during inference. The final input resolution of each frame is 224×224. For downstream tasks, we sample 16 frames for action recognition following [51] and 4 frames for text-to-video retrieval following [4]. More hyper-parameters are listed in Appendix.

Target →	None	Transcript	Video		
Dataset ↓	Baseline	Ours	VCOP [55]	MERLOT [57]	MERLOT-like
UCF-101	81.2 (↓2.2)	**83.4**	79.1 (↓4.3)	74.9 (↓8.5)	80.1 (↓3.3)
HMDB-51	56.5 (↓1.9)	**58.4**	54.2 (↓4.2)	49.6 (↓8.8)	55.4 (↓3.0)

Table 1. Comparison with methods that use ordering-based pretext tasks for pre-training. We report top-1 accuracy under the linear probe classification protocol on UCF-101 and HMDB-51. The model pre-trained only with $\mathcal{L}_{\text{base}}$ serves as the baseline. All models are pre-trained on the YT-Temporal dataset for a fair comparison.

Name	$\mathcal{L}_{\text{base}}$	$\mathcal{L}_{\text{sort}}$	sg	SSV2	Kinetics-400
$\text{M}_{\text{scratch}}$	×	×	-	64.5 (↓4.0)	75.4 (↓3.4)
M_{base}	✓	×	-	67.0 (↓1.5)	77.8 (↓1.0)
$\text{M}_{\text{sort}\setminus\text{sg}}$	×	✓	×	failed	failed
M_{sort}	×	✓	✓	failed	failed
$\text{M}_{\text{ours}\setminus\text{sg}}$	✓	✓	×	66.2 (↓2.3)	76.5 (↓2.3)
M_{ours}	✓	✓	✓	**68.5**	**78.8**

Table 2. The top-1 accuracy under the fine-tuning protocol on SSV2 and Kinetics-400, w.r.t. different pre-training objectives, where $\mathcal{L}_{\text{sort}}$ trains the pretext task of TVTS. sg denotes stopping gradients of $\mathcal{L}_{\text{sort}}$ towards encoding transcript representations.

Dataset	None	Sort Modeling		
	Baseline	Pairwise	Factorial	K-way
SSV2	67.0 (↓1.5)	67.4 (↓1.1)	67.2 (↓1.3)	**68.5**
K400	77.8 (↓1.0)	78.1 (↓0.7)	78.0 (↓0.8)	**78.8**

Table 3. The top-1 accuracy under the fine-tuning protocol on SSV2 and Kinetics-400, w.r.t. different ways to model the sorting of transcripts. "Pairwise" predicts the relative order for all transcript pairs, and "Factorial" performs K!-classification for all possible orders. Our method uses "K-way" classification to predict the order of each transcript. The model pre-trained with $\mathcal{L}_{\text{base}}$ only serves as the baseline.

4.4. Ablation Study

Pre-training Objectives. To demonstrate the effectiveness of our pretext task TVTS, we pre-train models with different objectives on the YT-Temporal dataset and evaluate them on SSV2 and K400. The results are listed in Table 2, in which $\text{M}_{\text{scratch}}$ denotes that we directly fine-tune the ImageMAE [23] initialized model. We have the following observations: (**i**) M_{base} outperforms $\text{M}_{\text{scratch}}$, which indicates that the natural language can be a promising supervision for video representation learning. (**ii**) Compared to M_{base}, M_{ours} further boosts performance by 1.5%, which demonstrates that TVTS can effectively regularize our model to learn transferable spatiotemporal representations. (**iii**) $\text{M}_{\text{ours}\setminus\text{sg}}$ drops performance, because when the gradients of $\mathcal{L}_{\text{sort}}$ flow towards encoding transcript representations, the model optimizes the transcript representations to ease the ordering task rather than enhance spatiotemporal representations to provide enough knowledge for transcript sorting. (**iv**) Both $\text{M}_{\text{sort}\setminus\text{sg}}$ and M_{sort} failed because the parametric module for sorting hardly converges when feeding the misaligned video and transcript representations from distinct latent spaces.

Sort Modeling. We explore different ways to model the order prediction of the shuffled transcripts. Besides using K-way classification of each transcript, we also tried *Pairwise*, which sorts the transcripts by predicting the relative orders of the $K(K-1)/2$ transcript pairs, and *Factorial*, which predicts an overall ordering distribution by performing a K!-way classification (K! possible orders given K transcripts). As listed in Table 3, both *Pairwise* and *Factorial* drop performance, because the former ignores the overall relationship among the transcripts while the latter imposes the same penalty on the results when different number of transcripts are sorted incorrectly. But they still outperform the baseline, indicating that sorting transcripts does benefit spatiotemporal representation learning. Our separate K-way classification modeling achieves the best performance.

4.5. Comparison with Ordering-based Pre-training

MERLOT [57] also adopts an ordering-based pretext task, but has a totally different approach and purpose. MERLOT reorders scrambled video frames given the representations of every single frame and the ordered ASR transcripts with a joint encoder, and reserves the joint encoder for downstream multimodal tasks such as VQA. Specifically, MERLOT predicts the relative order of two video frames by binary classification. MERLOT aims to promote the joint encoder in learning multi-modal representations rather than spatiotemporal representations of videos. Since the visual encoder takes a single frame as input, it only achieves semantic understanding at the single-frame level without temporal reasoning among frames. As MERLOT points out, the ordering pretext task is not critical for downstream tasks (performance even drops) and primarily serves as an interface to query the model about temporal events. We further tailor MERLOT for our architecture as MERLOT-like, which sorts K shuffled video frames by K-way classification with the knowledge of the ordered tran-

Method	Venue	Pre-train Dataset	Zero-shot Video-to-video Retrieval			Linear Probe
			R@1	R@5	R@10	
Spatiotemporal representation learning method(s)						
CVRL [44]	CVPR'21	Kinetics-400	-	-	-	11.4 (↓20.1)
MViT [16]	ICCV'21	Kinetics-400	-	-	-	19.4 (↓12.1)
SCVRL [14]	CVPRW'22	Kinetics-400	-	-	-	13.8 (↓17.7)
SVT [46]	CVPR'22	Kinetics-400	11.3 (↓3.4)	30.7 (↓7.7)	41.1 (↓9.4)	18.3 (↓13.2)
SVT† [46]	CVPR'22	YT-Temporal	9.9 (↓4.8)	26.2 (↓12.2)	36.3 (↓14.2)	18.0 (↓13.5)
VideoMAE [51]	NeurIPS'22	Kinetics-400	7.9 (↓6.8)	18.6 (↓19.8)	26.5 (↓24.0)	17.9 (↓13.6)
VideoMAE† [51]	NeurIPS'22	YT-Temporal	7.2 (↓7.5)	17.6 (↓20.8)	25.6 (↓24.9)	15.9 (↓15.6)
Video-text alignment method(s)						
Frozen‡ [4]	ICCV'21	CC3M, WebVid-2M	10.4 (↓4.3)	28.5 (↓9.9)	38.7 (↓11.8)	17.5 (↓14.0)
MCQ‡ [20]	CVPR'22	CC3M, WebVid-2M	10.4 (↓4.3)	28.6 (↓9.8)	38.5 (↓12.0)	18.0 (↓13.5)
MILES‡ [21]	ECCV'22	CC3M, WebVid-2M	10.3 (↓4.4)	28.4 (↓10.0)	38.4 (↓12.1)	18.6 (↓12.9)
Image representation learning method(s)						
CLIP [45]	ICML'21	WIT	10.5 (↓4.2)	28.8 (↓9.6)	38.8 (↓11.7)	16.4 (↓15.1)
Ours	CVPR'23	YT-Temporal	**14.7**	**38.4**	**50.5**	**31.5**

Table 4. Transferability evaluation on SSV2. We report Recall@K for zero-shot video-to-video retrieval and top-1 accuracy for linear probe classification, where video-to-video retrieval aims to retrieve videos of the same category as a query video. † denotes pre-training on YT-Temporal for a fair comparison, and ‡ denotes the use of official pre-trained weights for evaluation.

scripts. Similar to MERLOT, VCOP [55] also predicts the order of the shuffled video clips, but only pre-trains on videos without language semantics.

As listed in Table 1, all models that sort shuffled video clips/frames as the pretext task (*i.e.*, VCOP, MERLOT) perform even worse than the baseline counterpart without sorting. It indicates that **sorting shuffled videos in pre-training is infeasible and counterintuitive for improving spatiotemporal representations**, because it does not regularize the video encoder for spatial and temporal reasoning given only a single frame or a short video segment as input. They mainly regularize an extra module beyond the video encoder to figure out the chronological order of the videos. By contrast, our pretext task regularizes the model to sort transcripts via reasoning among the video representations, which enforces the model to capture contextualized spatiotemporal representations so that it can provide enough knowledge for transcript ordering.

4.6. Main Results

4.6.1 Action Recognition

Out-of-the-box Representations. To explore the transferability of the learned video representation, we evaluate zero-shot video-to-video retrieval and linear probe classification. We compare our proposed method with seven state-of-the-art methods, including: (a) Five video representation learning methods, *i.e.*, CVRL [44], MViT [16], SCVRL [14], SVT [46], and VideoMAE [51]. (b) Three video-text alignment methods, *i.e.*, Frozen [4], MCQ [20], and MILES [21]. (c) One image representation learning method with natural language supervision, *i.e.*, CLIP [45]. We average frame features as its video representation.

Method	Backbone	Pre-train Dataset	SSV2	K400
TSM [38]	R50 × 2	ImageNet-1K	66.0	-
Vi^{2}CLR [13]	S3D	Kinetics-400	-	71.2
CORP [27]	R3D-50	Kinetics-400	48.8	-
MoCo v3 [10]	ViT-B	Kinetics-400	62.4	-
TANet [41]	R50 × 2	ImageNet-1K	66.0	-
MViT [16]	ViT-B	Kinetcis-400	64.7	78.4
TimeSformer [6]	ViT-B	ImageNet-21K	59.5	78.3
RSANet [33]	R50	ImageNet-1K	66.0	-
SVT [46]	ViT-B	Kinetics-400	59.2	78.1
VideoMAE† [51]	ViT-B	YT-Temporal	67.9	78.2
Frozen‡ [4]	ViT-B	CC3M, WebVid2M	55.1	76.9
MCQ‡ [20]	ViT-B	CC3M, WebVid2M	51.5	77.8
MILES‡ [21]	ViT-B	CC3M, WebVid2M	54.1	77.4
OmniVL [52]	ViT-B	*Enormous Datasets	61.6	79.1
CLIP [45]	ViT-B	WIT	36.3	75.2
Ours	ViT-B	YT-Temporal	68.5	78.8
Ours	ViT-B	YT-Temporal CC3M, WebVid2M	**69.1**	**79.8**

Table 5. The top-1 accuracy under the fine-tuning protocol on SSV2 and Kinetics-400. OmniVL adopts a mixture of eight datasets. † denotes pre-training on YT-Temporal, and ‡ denotes the use of official pre-trained weights for evaluation.

The results are listed in Table 4 and we have the following observations: **(i)** Our method surpasses all baselines by

MSR-VTT			DiDeMo			MSVD			LSMDC		
Method	R@1	MedR	Method	R@1	MedR	Method	R@1	MedR	Method	R@1	MedR
MMT [19]	26.6	4.0	CE [40]	16.1	8.3	NoiseEst [1]	20.3	6.0	NoiseEst [1]	6.4	39.0
SupportSet [43]	30.1	3.0	ClipBert [36]	20.4	6.0	SupportSet [43]	28.4	4.0	MMT [19]	12.9	19.3
Frozen [4]	31.0	3.0	Frozen [4]	31.0	3.0	Frozen [4]	45.6	2.0	Frozen [4]	15.0	20.0
Ours	**34.6**	**3.0**	Ours	**32.4**	**3.0**	Ours	**45.9**	**2.0**	Ours	**17.2**	**17.0**

Table 6. The R@1 and MedR under the fine-tuning protocol on MSR-VTT, DiDeMo, MSVD, and LSMDC for text-to-video retrieval.

Method	Backbone	UCF-101	HMDB-51
BE [53]	I3D	87.1	56.2
CMD [29]	R(2+1)D-26	85.7	54.0
Vi^{2}CLR [13]	S3D	89.1	55.7
ASCNet [28]	S3D-G	90.8	60.5
TEC [30]	S3D-G	88.2	63.5
LSFD [5]	C3D	79.8	52.1
MCN [39]	R3D	89.7	59.3
TCLR [11]	R(2+1)D-18	84.3	54.2
SVT [46]	ViT-B	93.7	67.2
VideoMAE† [51]	ViT-B	94.2	68.4
Frozen‡ [4]	ViT-B	91.4	65.6
MCQ‡ [20]	ViT-B	92.9	65.1
MILES‡ [21]	ViT-B	92.1	66.8
Ours	ViT-B	**95.1**	**70.5**

Table 7. The top-1 accuracy under the fine-tuning protocol on UCF-101 and HMDB-51. † denotes pre-training on YT-Temporal, and ‡ denotes the use of official pre-trained weights for evaluation.

a large margin under all evaluation metrics, which indicates that our learned video representation has stronger transferability that can be used for out-of-domain video recognition. **(ii)** Previous video representation learning works yield weak transferability with only visual supervision. It implies that merely exploiting visual-only perception without explicit semantics hinders spatiotemporal understanding. Furthermore, we observe a significant performance drop on VideoMAE when it pre-trains the model on the large-scale uncurated dataset, *i.e.*, YT-Temporal. By contrast, our pre-trained model achieves promising results, which indicates that TVTS can successfully apply to real-world uncurated video data by exploiting rich semantics from script knowledge. **(iii)** Our method also outperforms video-text alignment works by a large margin. We infer that these works only focus on alignment between global video and caption representation without exploring fine-grained temporal information. On the contrary, our proposed TVTS regularizes the model to learn transferable spatiotemporal video representations. **(iv)** Benefiting from large-scale language supervision, image-based CLIP achieves competitive performance. But it is still worse than our model because we fully exploit the rich semantics from script knowledge.

Fine-tuning Transferability. We evaluate our model under the fine-tuning protocol on SSV2, Kinetics-400, UCF-101, and HMDB-51. Besides pre-training on YT-Temporal, we further follow recent works [20, 21] to jointly post-pretrain our model on **Google Conceptual Captions** (CC3M) and **WebVid-2M**. Their texts are harvested from the web in the form of a single caption. Since there is no timestamp-annotated text on CC3M and WebVid-2M, we only adopt the contrastive object. As listed in Table 5 and Table 7, the recognition capability of our model is comparable to previous works as we achieve state-of-the-art or competitive accuracy, while retaining strong transferability. Additionally, the video-text alignment methods show inferior performance on SSV2 since they are devoted to associating the vision patterns with language concepts, without fully exploiting the temporal information. By contrast, our TVTS achieves satisfactory performance via strengthening the learning of spatiotemporal representations.

4.6.2 Text-to-Video Retrieval

As we preserve a global video-transcript contrastive loss to ease the ordering task via learning semantically meaningful video representations, it is natural to ask if the semantic-aware video representations can also benefit retrieval. Hence we conduct text-to-video retrieval under the fine-tuning protocol. As reported in Table 6, our model achieves SOTA performance. The promising results show that our TVTS can also learn the association between video patterns and language semantics.

5. Conclusion

In this work, we for the first time leverage script knowledge that is naturally tied to the video to facilitate spatiotemporal representation learning. We introduce a novel pretext task dubbed *Turning to Video for Transcript Sorting* (TVTS), which regularizes the model to learn transferable video representations for spatial and temporal reasoning. Extensive evaluations on downstream video tasks show the great superiority of our method.

Acknowledgement. This work is supported in part by the National Natural Science Foundation of China under grant 62171248, the PCNL KEY project (PCL2021A07), Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005), Guangdong Basic and Applied Basic Research Foundation under grant 2021A1515110066, the GXWD 20220811172936001, the Shenzhen Science and Technology Innovation Commission (Research Center for Computer Network (Shenzhen) Ministry of Education), and Shenzhen Science and Technology Program under Grant JCYJ20220818101012025.

References

[1] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6644–6652, 2021. 8, 12, 13

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 5, 12

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 5

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 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*, pages 1728–1738, 2021. 2, 3, 5, 7, 8, 12, 13

[5] Nadine Behrmann, Mohsen Fayyaz, Juergen Gall, and Mehdi Noroozi. Long short view feature decomposition via contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9244–9253, 2021. 1, 3, 8

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 7

[7] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 5, 12

[8] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. Rspnet: Relative speed perception for unsupervised video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1045–1053, 2021. 1, 3

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 7

[11] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022. 8

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[13] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1502–1512, 2021. 7, 8

[14] Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. Scvrl: Shuffled contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4132–4141, 2022. 7

[15] 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. *arXiv preprint arXiv:2010.11929*, 2020. 5

[16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 7

[17] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 1, 2, 3, 4

[18] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2

[19] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020. 8, 12, 13

[20] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 2, 3, 5, 7, 8, 12

[21] Yuying Ge, Yixiao Ge, Xihui Liu, Alex Jinpeng Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. Miles: Visual bert pre-training with injected language semantics for video-text retrieval. *arXiv preprint arXiv:2204.12408*, 2022. 3, 7, 8, 12

[22] 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 *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1, 2, 3, 5, 12

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 5, 6

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 16000–16009, 2022. 3

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[26] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. 12

[27] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949, 2021. 7

[28] Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu, Mingkui Tan, and Errui Ding. Ascnet: Self-supervised video representation learning with appearance-speed consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8096–8105, 2021. 1, 3, 8

[29] Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, and Rong Jin. Self-supervised video representation learning by context and motion decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13886–13895, 2021. 8

[30] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9970–9980, 2021. 8

[31] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[32] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 5, 12

[33] Manjin Kim, Heeseung Kwon, Chunyu Wang, Suha Kwak, and Minsu Cho. Relational self-attention: What's missing in attention for video understanding. *Advances in Neural Information Processing Systems*, 34:8046–8059, 2021. 7

[34] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, 2011. 1, 2, 5, 12

[35] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 2, 3

[36] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 8, 12, 13

[37] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 12, 13

[38] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 7

[39] Yuanze Lin, Xun Guo, and Yan Lu. Self-supervised video representation learning with meta-contrastive network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8239–8249, 2021. 8

[40] 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. 8, 12, 13

[41] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13708–13718, 2021. 7

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[43] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 8, 12, 13

[44] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 7

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3, 7

[46] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2874–2884, 2022. 3, 7, 8

[47] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 5, 12

[48] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 5

[49] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning. *arXiv preprint arXiv:2207.07635*, 2022. 2

[50] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos

in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 2, 5, 12

[51] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 1, 2, 3, 4, 5, 7, 8, 12

[52] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *arXiv preprint arXiv:2209.07526*, 2022. 7

[53] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11804–11813, 2021. 8

[54] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 4

[55] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 6, 7

[56] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5, 12

[57] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 1, 2, 3, 5, 6, 13

A. Downstream Datasets

A.1. Action Recognition

The statistics of our downstream action recognition datasets are listed as follows: (a) **Something-Something V2** (SSV2) [22] is a large-scale dataset that shows humans performing pre-defined basic actions with everyday objects. It consists of 169K training videos and 20K validation videos belonging to 174 fine-grained action classes. (b) **Kinetics-400** [32] contains 240K training videos and 20K validation videos belonging to 400 classes. (c) **UCF-101** [50] contains 9.5K/3.5K training and validation videos with 101 action classes. (d) **HMDB-51** [34] contains 3.5K/1.5K training and validation videos with 51 action classes.

A.2. Text-to-Video Retrieval

The statistics of our downstream text-to-video retrieval datasets are listed as follows: (a) **MSR-VTT** [56] contains 10K YouTube videos with 200K descriptions. Following [4], we train on the training and validation set consisting of 9K videos and evaluate on the 1K-A test set. (b) **MSVD** [7] contains 1,970 YouTube videos with 80K descriptions, where each video has around 40 sentences. We adopt the official split [4], in which 1200, 100, and 670 videos are used for training, validation, and testing respectively. (c) **DiDeMo** [2] contains 10K Flickr videos with 40K sentences. We follow [4, 20, 21] to evaluate paragraph-to-video retrieval, *i.e.*, we concatenate all sentences for a video to form a single query. Specifically, we directly use the whole video without cropping the localized moments (as done by [4, 20, 21]). (d) **LSMDC** [47] consists of 118,081 video clips harvested from 202 movies. We adopt the split of [4], where the validation and test set has 7,408 and 1,000 videos respectively.

B. Implementation Details

As some of the YT-Temporal dataset's video sources, *e.g.*, YouTube, are overlapped with those of downstream datasets, we have carefully checked that there is no data leakage between pre-training and downstream datasets by extracting respective frame features with CLIP, calculating their similarity between frame features, and manually examining those with similarity above the threshold.

Our training hyper-parameters are listed in Table 8 and Table 9. We mostly follow the setting of [51] for convenience. Carefully tuning these parameters may yield better performance.

config	pre-train	post-pretrain
optimizer	AdamW	
learning rate	1×10^{-4}	
batch size	1024	800
training epochs	20	12
training frames	16	$1 + 4$
masking ratio	75%	0
input size	224×224	
patch size, P	16	
data augmentation	RandomCrop	
hidden state dimension, D_h	768	
common space dimension, D	256	
temperature parameter, τ	0.05	

Table 8. The pre-train and post-pretrain setup.

config	linear probe	fine-tuning
optimizer	SGD	AdamW
learning rate	0.1	0.001
batch size	384	384
training epochs	100	50 (SSV2), 100 (Others)
training frames	16	
clips $\times$ crops	5×3 (K400), 2×3 (Others)	
data augmentation	CenterCrop	

Table 9. The linear probe and fine-tuning setup.

C. Additional Experiments

C.1. Full Results for Text-to-Video Retrieval

We compare our method with seven state-of-the-art methods [1, 4, 19, 36, 37, 40, 43]. The full Recall@K and MedR results are reported in Table 10. Our model achieves state-of-the-art or competitive performance on all datasets. It shows that our TVTS is capable of learning the association between video patterns and language semantics.

C.2. SVO-Probes Test

Our model can also be well transferred to understand static images and reason about the dynamic context behind them. To evaluate such an ability, we conduct experiments on the recently proposed SVO Probes [26], a zero-shot test benchmark for *subject*, *verb*, and *object* understanding in the image field. In SVO Probes, each sentence is tied with a positive and a negative image, in which the positive image has consistent semantics, *i.e.* subject, verb, and object, with the sentence, while the negative image substitutes one of the three concepts but keeps the remaining two unchanged. The objective is to test whether a model can correctly identify the positive image given a query sentence. We treat it as a text-image retrieval task, *i.e.* given the text and image embedding, if their cosine similarity surpasses a certain threshold ρ, we consider the image positive. We report the

MSR-VTT				
Method	R@1	R@5	R@10	MedR
NoiseEst [1]	17.4	41.6	53.6	8.0
MMT [19]	26.6	57.1	69.6	4.0
SupportSet [43]	30.1	58.5	69.3	3.0
Frozen [4]	31.0	59.5	70.5	3.0
Ours	**34.6**	**61.5**	**72.2**	**3.0**

DiDeMo				
Method	R@1	R@5	R@10	MedR
HERO [37]	2.1	-	11.4	-
CE [40]	16.1	41.1	82.7	8.3
ClipBert [36]	20.4	48.0	60.8	6.0
Frozen [4]	31.0	59.8	**72.4**	3.0
Ours	**32.4**	**59.8**	71.7	**3.0**

LSMDC				
Method	R@1	R@5	R@10	MedR
NoiseEst [1]	6.4	19.8	28.4	39.0
MMT [19]	12.9	29.9	40.1	19.3
Frozen [4]	15.0	30.8	39.8	20.0
Ours	**17.2**	**32.8**	**41.7**	**17.0**

MSVD				
Method	R@1	R@5	R@10	MedR
NoiseEst [1]	20.3	49.0	63.3	6.0
SupportSet [43]	28.4	60.0	72.9	4.0
Frozen [4]	45.6	**79.8**	**88.2**	2.0
Ours	**45.9**	76.7	85.4	**2.0**

Table 10. The full results for text-to-video retrieval on MSR-VTT, DiDeMo, LSMDC, and MSVD.

ρ	0.2			0.25			0.3		
Method	subj	obj	verb	subj	obj	verb	subj	obj	verb
Frozen	0.56	0.61	0.54	0.58	0.66	0.56	0.62	0.72	0.58
Ours	**0.59**	**0.65**	**0.59**	**0.64**	**0.70**	**0.62**	**0.68**	**0.76**	**0.63**

Table 11. Experiments on SVO Probes, a recently proposed benchmark for the subject, verb, and object understanding in static images. Our pre-trained model can better reason about the dynamic context behind the given images. We do not compare with SOTA spatiotemporal representation learning methods, *e.g.*, VideoMAE, since they cannot perform text-to-video retrieval.

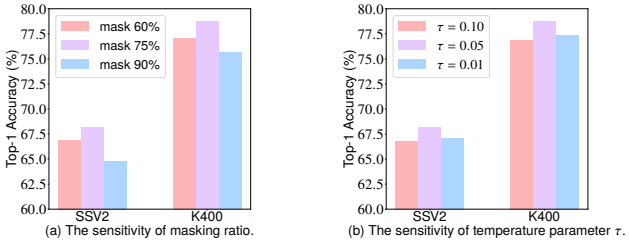
Name	Formulation	$\mathcal{L}_{\text{base}}$	$\mathcal{L}_{\text{sort}}$	SSV2	Gain
M_1	MERLOT	✓	×	66.2	+0.9
M_2		✓	✓	67.1	
M_3	Ours	✓	×	67.0	+1.5
M_4		✓	✓	**68.5**	

Table 12. The top-1 accuracy w.r.t. different contrastive formulation on SSV2 under the fine-tuning protocol.

Sort Source	Transcripts, K	Sort Module	Accuracy
T	4	RG	0.4%
T		SortTSF	0.5%
T + V		SortTSF	**21.5%**

Table 13. The sort accuracy w.r.t. different sort modules. T (V) denotes the transcript (video) representation; RG refers to random guessing, and SortTSF refers to the sort transformer.

precision results in terms of different values of ρ, shown in Table 11. Our model reaches higher precision on all concepts, which implies our learned spatiotemporal representations have strong out-of-the-box capabilities.



(a) The sensitivity of masking ratio. (b) The sensitivity of temperature parameter τ.

Figure 3. (a) The top-1 accuracy w.r.t. different masking ratio. (b) The top-1 accuracy w.r.t. different temperature parameter τ.

C.3. Ablation Study (Cont.)

Contrastive Formulation. Since MERLOT [57] formulates the contrastive objective by frame-transcript matching, we further investigate how much this change in the proposed approach from MERLOT contributes to the improved performance. Specifically, we replace the contrastive formulation of $\mathcal{L}_{\text{base}}$ with that of MERLOT, and the results are reported in Table 12. The accuracy slightly degrades due to mismatches between single frames and noisy transcripts, but the sorting task still boosts video representations, given the gains when plugging $\mathcal{L}_{\text{sort}}$.

Sort Accuracy. To prevent the model from learning shortcuts, *i.e.*, memorizing orders from text alone, we stop the gradients of sorting loss from flowing toward encoding transcript features. To verify it, we test the accuracy of transcript sorting using our pre-trained model in Table 13, where the expectation of random guessing accuracy is 0.4% ($1/4^4$). Sorting the text alone almost fails, while sorting text via resorting to video features achieves 21.5% accuracy. It implies the sorting task is solved by promoting video understanding instead of learning shortcuts.

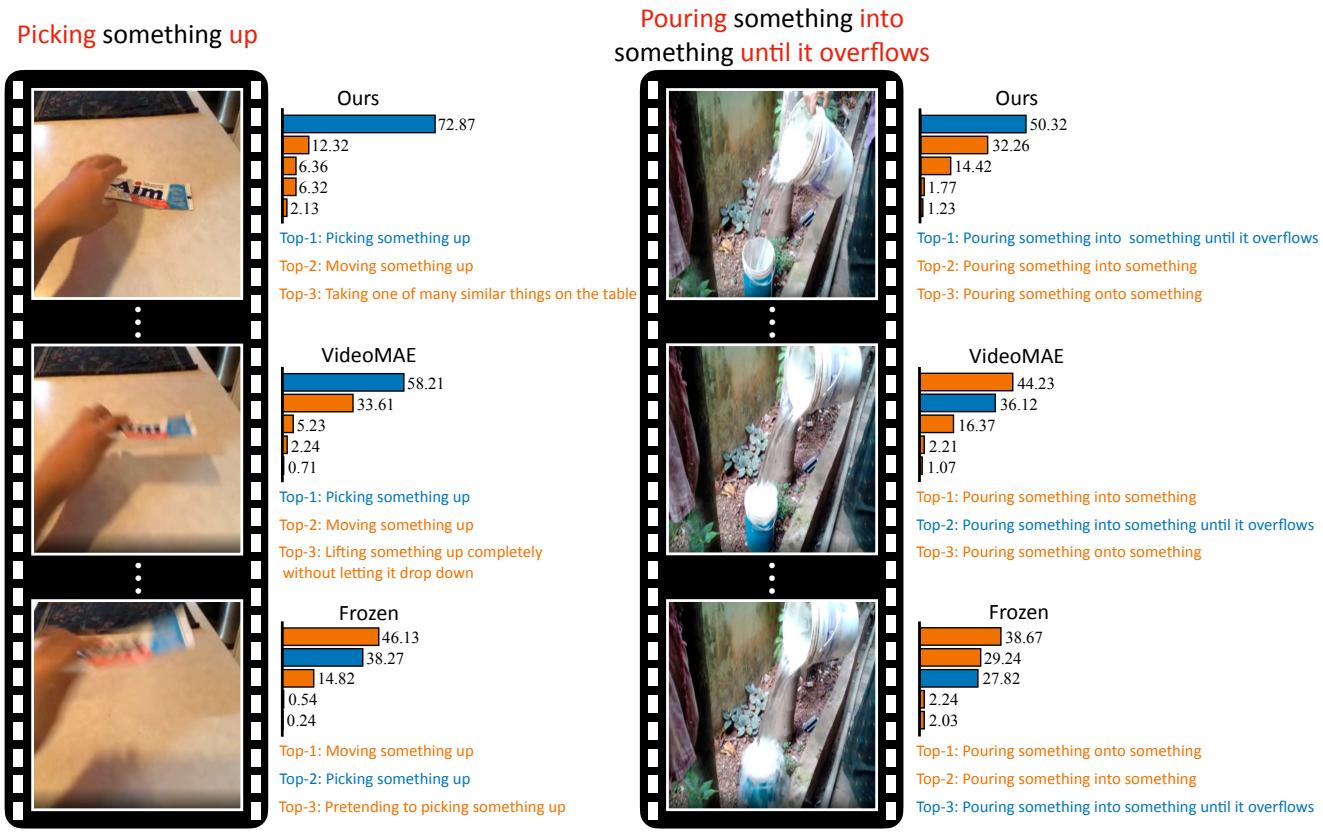


Figure 4. Visualization of the top-5 prediction scores on SSV2, we normalize the scores to make their summation 100%. The blue and orange rows denote the scores of the right and wrong classes, respectively.

Masking Ratio. We compare different masking ratios for TVTS in Figure 3(a). Both lower (60%) and higher (90%) masking ratio drop performance than our method with 75% ratio, because a lower masking ratio brings in temporal redundancy, while a higher ratio leads to the extremely limited knowledge to perform TVTS.

Temperature Parameter. We also investigate the influence of the temperature parameter τ in $\mathcal{L}_{\text{base}}$ in Figure 3(b). A smaller τ makes the model focus more on the hard negative samples, but it also increases the difficulty of convergence. We set $\tau = 0.05$ for its best performance.

Visualization. To demonstrate the superiority of our learned spatiotemporal representation intuitively, we randomly pick two videos in SSV2 and illustrate the top-5 prediction scores w.r.t. our method, VideoMAE and Frozen in Figure 4. Our method predicts the highest score for the right class. In the first column, we need to distinguish the action "picking" from other similar actions such as "moving", which requires fine-grained temporal reasoning ability. In the second column, the model must extract both the spatial and temporal information to classify the video as the category containing "into" and "until it overflows". Only our method classifies the video correctly, while VideoMAE and Frozen make mistakes due to a lack of spatiotemporal modeling ability.