LEARNING TRANSFERABLE SPATIOTEMPORAL REPRESENTATIONS FROM NATURAL SCRIPT KNOWLEDGE

Anonymous authors
Paper under double-blind review

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 commonsense, which is far away from cognition-level 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. Furthermore, rather than the simple concept learning in vision-caption contrast, we encourage cognition-level temporal commonsense reasoning via narrative reorganization. The advantages enable our model to contextualize what is happening like human beings and seamlessly apply to large-scale uncurated video data in the real world. Note that our method differs from ones designed for video-text alignment (*e.g.*, Frozen) and multimodal representation learning (*e.g.*, Merlot). Our method demonstrates strong out-of-the-box spatiotemporal representations on diverse video benchmarks, *e.g.*, +13.6% gains over VideoMAE on SSV2 via linear probing.

1 INTRODUCTION

The aspiration of representation learning is to encode general-purpose representations that transfer well to diverse downstream tasks, where self-supervised methodologies (He et al., 2020; Chen et al., 2020) dominate due to their advantage in exploiting large-scale unlabeled data. Despite significant progress in learning representations of still images (He et al., 2022b; Radford et al., 2021), 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 with cognitive capabilities.

There have been various attempts at self-supervised pre-training on video data from discriminative learning objectives (Chen et al., 2021a; Huang et al., 2021a; Behrmann et al., 2021) to generative ones (Tong et al., 2022; Feichtenhofer et al., 2022), 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 (Goyal et al., 2017; Soomro et al., 2012; Kuehne et al., 2011) via fine-tuning, the learned representations are still far away from out-of-the-box due to the unsatisfactory 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 (Kay et al., 2017). Their applicability in the real world is questioned given the observed performance drops when training on a larger but uncurated dataset, YT180M (Zellers et al., 2021). We argue that all you need to address the above issues is cognition-level spatiotemporal understanding like human beings. But current video models generally exploit visual-only perception (*e.g.*, pixels) without explicit semantics.

Recently, the success of CLIP (Radford et al., 2021) has inspired the community to learn semantically aware image representations that are better transferable to downstream tasks and scalable

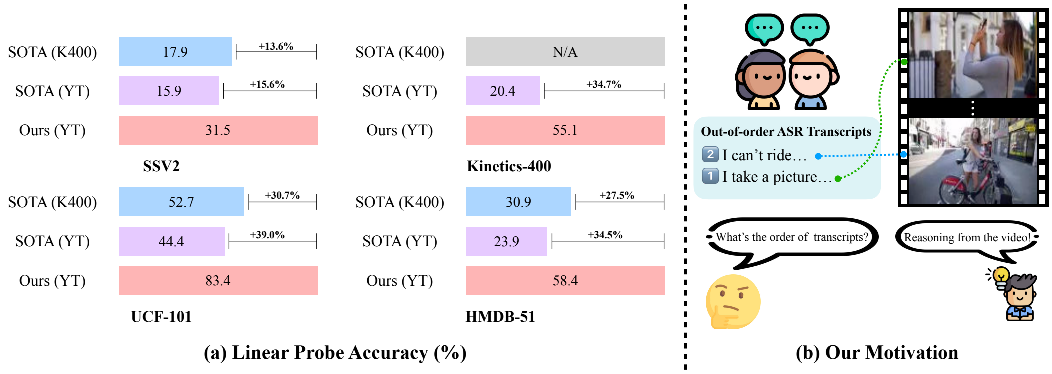


Figure 1: (a) We evaluate the transferability of spatiotemporal representations via linear probing on four video recognition datasets (Goyal et al., 2017; Kay et al., 2017; Soomro et al., 2012; Kuehne et al., 2011), where the state-of-the-art method (Tong et al., 2022) underperforms. It performs even worse when pre-training with a large-scale uncurated dataset, YT-Temporal (Zellers et al., 2021). (b) We encourage complex temporal understanding and advanced spatiotemporal representation learning with a new pretext task of sorting transcripts via temporal commonsense reasoning.

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 temporal understanding and temporal commonsense reasoning with language supervision needs to be studied. (2) The quality of language supervision (Santurkar et al., 2022) 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 data itself, which is also the philosophy followed by previous video pre-training methods (Tong et al., 2022; Feichtenhofer et al., 2022). Fortunately, video data is naturally multi-modality with transcribed speech knowledge in the form of text (ASR), providing time-dependent semantics despite some noise.

To enable cognition-level 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 commonsense inference. 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, temporal commonsense 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.

Specifically, besides the video encoder, we employ a text encoder to embed script representations from ASR transcripts, together with a parametric module SortFormer to realize the pretext task of TVTS. Given an input video and its successive transcripts, we randomly shuffle the order of the sentences. The encoded script representations are then fed into SortFormer to predict their actual orders by attending to the video representations via cross-attention layers. The order prediction is cast as a K-way classification task, where K is the number of transcripts. The pretext task indirectly regularizes the video encoder to properly capture contextualized spatiotemporal representations to provide enough knowledge for transcript ordering. Besides TVTS, a global video-transcript contrastive loss is preserved to ease the reordering task via learning semantically meaningful video representations. Only the video encoder is utilized for downstream video recognition tasks for fair comparisons.

The usage of language supervision is related to video-text alignment (Bain et al., 2021; Ge et al., 2022a) and multimodal representation learning (Zellers et al., 2021; Fu et al., 2021) 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 (Lei et al., 2022) and fail to encode strong out-of-the-box temporal representations.

(2) Multimodal representation learning methods aim to match different modalities in the temporal dimension for mutual complementation and fusion. Though (Zellers et al., 2021) also reorders ASR scripts, it is tailored for learning joint representations across modalities rather than spatiotemporal representations in our work. It does not work when used directly for our task, discussed in Sec. 3.4.

To summarize, our contributions are three-fold. (1) We are the first to study cognition-level spatiotemporal representation learning. 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. (2) We introduce a novel pretext task for video pre-training, namely, Turning to Video for Transcript Sorting (TVTS). It promotes the capability of the video encoder in learning transferable video representations to perform temporal commonsense reasoning. (3) 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 (Goyal et al., 2017). 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 (Ranasinghe et al., 2022) aligns several views from the same video with different spatial and temporal resolution for video-invariant representations. RSPNet (Chen et al., 2021a), ASCNet (Huang et al., 2021a), and LongShortView (Behrmann et al., 2021) 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 (He et al., 2022a) methods (Tong et al., 2022; Feichtenhofer et al., 2022) 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 (Lei et al., 2022). 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, leading to weak cognitive capabilities. 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.

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 (Bain et al., 2021), MCQ (Ge et al., 2022a), and MILES (Ge et al., 2022b) 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 carefully describing the dynamic content over time. The second category works on joint representation learning across modalities mainly for VQA. For example, Merlot (Zellers et al., 2021) 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 mutual complementation and fusion in a joint encoder, rather than learn better spatiotemporal representations with cognitive capabilities.

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 (Radford et al., 2021) 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 (Jia et al., 2021), 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.

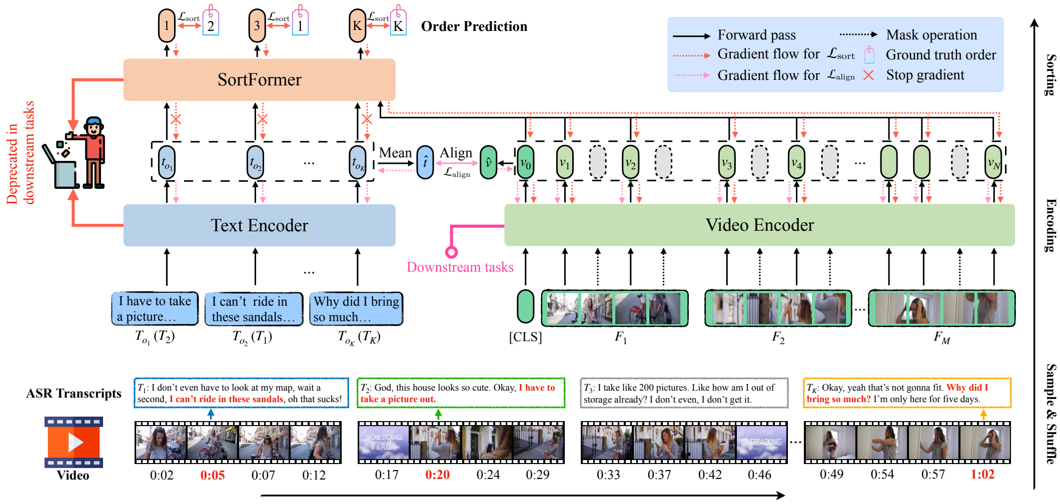


Figure 2: Our pre-training pipeline. We first sample K consecutive ASR transcripts, and a video clip consisting of M frames within the span of the transcripts. Then we shuffle the transcripts and encode their representations through a text encoder. Next, we feed the clip with the correct frame order into the video encoder for the video representations. Finally, the text and video representations are concatenated and fed into SortFormer to predict the order of each transcript. We adopt a contrastive objective $\mathcal{L}_{\text{align}}$ to align the semantics between vision and text, and a cross-entropy objective $\mathcal{L}_{\text{sort}}$ to train SortFormer to correctly sort the shuffled transcripts. Only the video encoder is adopted for embedding spatiotemporal representations in downstream tasks.

3 Method

We introduce a novel pretext task, **Turning to Video for Transcript Sorting (TVTS)**, with a parametric module SortFormer, to learn the transferable spatiotemporal video representation. 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 architecture of three components in Sec. 3.3. At last, we clarify the differences between our method and other work that use ordering for representation learning in Sec. 3.4.

3.1 Turning to Video for Transcript Sorting

As shown in Fig. 2, TVTS is performed using a parametric module SortFormer 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. SortFormer is then trained to sort the transcripts in the correct order via resorting to the video representations from the video encoder.

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,

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

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 (Wang et al., 2016) to divide $[S_1, E_K]$ into M segments with equal length and randomly sample 1 frame from each segment. After the above operations, 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 mask a large proportion of the video clip among the spatial and temporal dimension as the input of the video encoder 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 (Tong et al., 2022; Feichtenhofer et al., 2022). We mask the video clip as a means of data augmentation since it provides corrupted knowledge for SortFormer to realize 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 feed them into SortFormer to perform multi-head self-attention. SortFormer attempts to sort the order of the transcripts 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. The first K output representations of SortFormer $\{z_{o_i}\}_{i=1}^{K}$ are further fed into a linear classifier to predict the order $p_i \in \mathbb{R}^K$, where p_i denotes the probability that the transcript is the *i*-th transcript in the original ordered sequence. For the transcript T_{o_i}, the groundtruth classification label should be o_i.

The pretext task of TVTS regularizes the video encoder to contextualize what is happening over time, so that it can provide enough knowledge for SortFormer to figure out the chronological order of the shuffled transcripts. It improves the capability of the video encoder to learn spatiotemporal representations that enable temporal commonsense reasoning.

3.2 Pre-training Objectives

Besides the pretext task of TVTS, we use a global video-transcript contrastive objective, which aligns the features of the video clip and the averaged features of K transcripts. It aims to ease the pretext task of TVTS through learning semantic-aware video representations. We combine two objectives to optimize the entire model in an end-to-end manner. The first one is the global video-transcript contrastive objective $\mathcal{L}_{\text{align}}$, formulated as a bidirectional InfoNCE (Oord et al., 2018),

$$\mathcal{L}_{\text{align}} = \text{NCE}(\hat{t}, \hat{v}) + \text{NCE}(\hat{v}, \hat{t}) \quad s.t.\ \text{NCE}(q, k) = -\log \frac{\exp(q^\top k_+/\tau)}{\sum_{i=1}^{B} \exp(q^\top k_i/\tau)}, \tag{2}$$

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.*, $\hat{v} \leftarrow v_0$.

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

$$\mathcal{L}_{\text{sort}} = -\frac{1}{K}\sum_{i=1}^{K} \text{softmax}(\hat{p}^i) \quad s.t.\ \text{softmax}(\hat{p}^i) = \frac{\exp(p^i_{o_i})}{\sum_{j=1}^{K} \exp(p^i_j)}, \tag{3}$$

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 groundtruth order in the original ordered sequence.

Our overall pre-training objective combines the two objectives, *i.e.*, $\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda\mathcal{L}_{\text{sort}}$, 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 $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{sort}}$ to be the same for efficient training.

3.3 Model Architecture

Video Encoder. The video encoder takes a video clip as input, which consists of M frames of resolution $H \times W$, and outputs video representations. We follow (Tong et al., 2022) 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 VideoFormer learns the positional information of the cubes. Next, we follow BERT (Devlin et al., 2018) 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 (Dosovitskiy et al., 2020) architecture as the video encoder. The unmasked N video tokens as well as the [CLS] token are fed into the video encoder, and perform joint space-time attention (Arnab et al., 2021) among the whole unmasked token sequence.

Text Encoder. The text encoder takes the ASR transcripts as inputs and outputs text representations. We adopt DistilBERT (Sanh et al., 2019) as the text encoder. The final transcript representations are taken from the [CLS] token, which is concatenated at the beginning of the input text.

SortFormer. SortFormer takes the concatenated transcript-video representations as inputs and outputs the order for each transcript. It consists of two stacked bidirectional transformer blocks. 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. The output transcript representations are further fed into a linear classifier to perform K-way classification, which produces the order prediction.

3.4 The Pretext Task of Ordering in Representation Learning

Merlot (Zellers et al., 2021) also adopts an ordering-based pretext task, but has a totally different approach and purpose. Specifically, Merlot reorders scrambled video frames given the ordered ASR transcripts with a joint encoder, and reserves the joint encoder for downstream multimodal tasks such as VQA. Merlot aims to promote the joint encoder in learning joint representations across modalities rather than spatiotemporal representations. By contrast, TVTS sorts shuffled ASR transcripts via resorting to the ordered video clip using a proxy SortFormer. It aims to improve the capability of the video encoder in learning transferable spatiotemporal video representations by forcing the video encoder to provide enough knowledge for transcript ordering. When the ordering task in Merlot is directly applied to our method, it achieves poor performance as shown in Sec. 4.5.

4 Experiments

4.1 Pre-training Datasets

We pre-train our model on the large-scale **YT-Temporal** dataset (Zellers et al., 2021), which contains 6M YouTube videos with ASR transcripts and word-level timestamps. We downloaded 5M videos for pre-training and abandon the rest. We also follow recent works (Ge et al., 2022a;b) to jointly post-pretrain our model on **Google Conceptual Captions** (CC3M) and **WebVid-2M**. Both of 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 $\mathcal{L}_{\text{align}}$.

4.2 Downstream Tasks

Action Recognition. We evaluate our pre-trained model on four common video datasets: (a) **Something-Something V2** (SSV2) (Goyal et al., 2017), (b) **Kinetics-400** (Kay et al., 2017), (c) **UCF-101** (Soomro et al., 2012), (d) **HMDB-51** (Kuehne et al., 2011). Our evaluation is two-fold: (i) We conduct *zero-shot* video retrieval and *linear* action recognition on the SSV2 dataset to evaluate the transferability of the learned video representation, where the former aims to retrieve videos of the same category as a query video, and the latter freezes the video encoder and only optimizes a linear classifier. (ii) We *fully fine-tune* our pre-trained model with label supervision on the training set of the four datasets to evaluate the recognition capability. See Appendix A.1.1 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** (Xu et al., 2016) (b) **DiDeMo** (Anne Hendricks et al., 2017) (c) **MSVD** Chen & Dolan (2011) (d) **LSMDC** (Rohrbach et al., 2015). We adopt Recall@K (R@K) and Median Rank (MedR) as the evaluation metric. See Appendix A.1.2 for details.

Method	Venus	Pre-train Dataset	Zero-shot Video Retrieval			Linear Probe
			R@1	R@5	R@10	
Spatiotemporal representation learning method(s)						
CVRL	CVPR'21	Kinetics-400	-	-	-	11.4 (↓20.1)
MViT	ICCV'21	Kinetics-400	-	-	-	19.4 (↓12.1)
SCVRL	CVPR'22	Kinetics-400	-	-	-	19.4 (↓12.1)
SVT	CVPR'22	Kinetics-400	-	-	-	18.3 (↓13.2)
VideoMAE	NeurIPS'22	Kinetics-400	7.9 (↓6.8)	18.6 (↓19.8)	26.5 (↓24.0)	17.9 (↓13.6)
VideoMAE	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	ICCV'21	CC3M, WebVid-2M	10.4 (↓4.3)	28.5 (↓9.9)	38.7 (↓11.8)	17.5 (↓14.0)
MCQ	CVPR'22	CC3M, WebVid-2M	10.4 (↓4.3)	28.6 (↓9.8)	38.5 (↓12.0)	18.0 (↓13.5)
MILES	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	Arxiv'21	400M Web Data	10.5 (↓4.2)	28.8 (↓9.6)	38.8 (↓11.7)	16.4 (↓15.1)
Ours	-	YT-Temporal	**14.7**	**38.4**	**50.5**	**31.5**

Table 1: Transferability evaluation on the SSV2 dataset. We report Recall@K for zero-shot video retrieval and top-1 accuracy for linear probe classification, where video retrieval aims to retrieve videos of the same category as a query video.

Dataset	#Frames	CLIP	Frozen	VideoMAE	SOTA	Ours
SSV2	16	36.3	55.1	68.2	68.2 (Tong et al., 2022)	**68.9**
K400		75.2	76.9	79.4	79.7 (Patrick et al., 2021)	**79.8**
UCF-101		90.3	88.7	94.2	94.2 (Tong et al., 2022)	**95.1**
HMDB-51		67.4	67.7	70.2	70.2 (Tong et al., 2022)	**70.5**

Table 2: Top-1 accuracy on action recognition benchmarks under the fine-tuning protocol.

4.3 Implementation Details

We follow recent works (Bain et al., 2021; Ge et al., 2022a) to adopt the pre-trained DistilBERT (Sanh et al., 2019) as the text encoder. The video encoder is a vanilla ViT-Base (Dosovitskiy et al., 2020) initialized with ImageMAE-Base He et al. (2022b). We first pre-train our model on the YT-Temporal dataset sampling 16 frames for 20 epochs. Then we jointly post-pretrain our model on the CC3M sampling 1 frame and the WebVid-2M sampling 4 frames for 12 epochs. We randomly mask 75% tokens within each frame for the YT-Temporal pre-training. For downstream tasks, we sample 16 frames for action recognition following (Tong et al., 2022) and 4 frames for text-to-video retrieval following (Bain et al., 2021). The detailed hyper-parameters are listed in Appendix A.2.

4.4 Main Results

4.4.1 Action Recognition

Out-of-the-box representations. To explore the transferability of the learned video representation, we evaluate zero-shot 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 (Qian et al., 2021), MViT (Fan et al., 2021), SCVRL (Dorkenwald et al., 2022), SVT (Ranasinghe et al., 2022), and VideoMAE (Tong et al., 2022). (b) Three video-text alignment methods, *i.e.*, Frozen (Bain et al., 2021), MCQ (Ge et al., 2022a), and MILES (Ge et al., 2022b). (c) One image representation learning method with natural language supervision, *i.e.*, CLIP (Radford et al., 2021). Specially, we average frame features as the CLIP video representation.

The results are listed in Table 1 and we have the following observations: **(i)** Our method surpasses all baselines by 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 can not

Method	Backbone	Pre-train Dataset	Params	GFLOPs	SSV2	K400
TSM (Lin et al., 2019)	R50 $\times$ 2	ImageNet-1K	49M	130 $\times$ 2 $\times$ 3	66.0	-
Vi^{2}CLR (Diba et al., 2021)	S3D	Kinetics-400	-	-	-	71.2
CORP (Hu et al., 2021)	R3D-50	Kinetics-400	32M	-	48.8	-
MoCo v3 (Chen et al., 2021b)	ViT-B	Kinetics-400	87M	-	62.4	-
TANet (Liu et al., 2021)	R50 $\times$ 2	ImageNet-1K	51M	99 $\times$ 2 $\times$ 3	66.0	-
MViT (Fan et al., 2021)	ViT-B	Kinetcis-400	37M	455 $\times$ 1 $\times$ 3	64.7	78.4
TimeSformer (Bertasius et al., 2021)	ViT-B	ImageNet-21K	121M	196 $\times$ 1 $\times$ 3	59.5	78.3
Motionformer (Patrick et al., 2021)	ViT-B	ImageNet-21K, K400	109M	370 $\times$ 1 $\times$ 3	66.5	79.7
RSANet (Kim et al., 2021)	R50	ImageNet-1K	24M	72 $\times$ 1 $\times$ 1	66.0	-
SVT (Ranasinghe et al., 2022)	ViT-B	Kinetics-400	87M	-	59.2	78.1
VideoMAE (Tong et al., 2022)	ViT-B	Kinetcis-400	87M	180 $\times$ 2 $\times$ 3	68.2	79.4
VideoMAE (Tong et al., 2022)	ViT-B	YT-Temporal	87M	180 $\times$ 2 $\times$ 3	67.9	78.2
Frozen (Bain et al., 2021)	ViT-B	CC3M, WebVid2M	114M	370 $\times$ 2 $\times$ 3	55.1	76.9
MCQ (Ge et al., 2022a)	ViT-B	CC3M, WebVid2M	87M	280 $\times$ 2 $\times$ 3	51.5	77.8
MILES (Ge et al., 2022b)	ViT-B	CC3M, WebVid2M	114M	370 $\times$ 2 $\times$ 3	54.1	77.4
OmniVL (Wang et al., 2022)	ViT-B	*Enormous Datasets	87M	-	61.6	79.1
CLIP (Radford et al., 2021)	ViT-B	400M Web Data	87M	281 $\times$ 2 $\times$ 3	36.3	75.2
Ours	ViT-B	YT-Temporal	87M	180 $\times$ 2 $\times$ 3	68.2	78.8
Ours	ViT-B	YT-Temporal CC3M, WebVid2M	87M	180 $\times$ 2 $\times$ 3	**68.9**	**79.8**

Table 3: Top-1 accuracy under the fine-tuning protocol on SSV2 and Kinetics-400 (K400). OmniVL adopts a mixture of eight datasets. We report GFLOPs on SSV2 evaluation with 2 clips $\times$ 3 crops.

Method	UCF-101	HMDB-51	Method	UCF-101	HMDB-51
BE (Wang et al., 2021)	87.1	56.2	CMD (Huang et al., 2021b)	85.7	54.0
Vi^{2}CLR (Diba et al., 2021)	89.1	55.7	ASCNet (Huang et al., 2021a)	90.8	60.5
TEC (Jenni & Jin, 2021)	88.2	63.5	LSFD (Behrmann et al., 2021)	79.8	52.1
MCN (Lin et al., 2021)	89.7	59.3	TCLR (Dave et al., 2022)	84.3	54.2
SVT (Ranasinghe et al., 2022)	93.7	67.2	Frozen (Bain et al., 2021)	88.7	65.6
MCQ (Ge et al., 2022a)	92.9	65.1	MILES (Ge et al., 2022b)	92.1	66.8
VideoMAE (Tong et al., 2022)	94.2	70.2	Ours	**95.1**	**70.5**

Table 4: Top-1 accuracy under the fine-tuning protocol on UCF-101 and HMDB-51.

realize cognition-level 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 video encoder to learn transferable spatiotemporal video representations. **(iv)** Benefiting from large-scale language supervision, image-based CLIP achieves a comparable performance compared to video-based methods. But it is still worse than our model because our work fully exploits the rich semantics from script knowledge, which is naturally along with videos.

Fine-tuning transferability. We evaluate our model under the fine-tuning protocol. Table 2 gives the overall results, and the detailed comparisons are reported in Tables 3 and 4. 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.

4.4.2 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

MSR-VTT			DiDeMo			MSVD			LSMDC		
Method	R@1	MedR	Method	R@1	MedR	Method	R@1	MedR	Method	R@1	MedR
MMT	26.6	4.0	CE	16.1	8.3	NoiseEst	20.3	6.0	NoiseEst	6.4	39.0
SupportSet	30.1	3.0	ClipBert	20.4	6.0	SupportSet	28.4	4.0	MMT	12.9	19.3
Frozen	31.0	3.0	Frozen	31.0	3.0	Frozen	45.6	2.0	Frozen	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 5: The R@1 and MedR *w.r.t.* MSR-VTT, DiDeMo, MSVD, and LSMDC from left to right.

Dataset	scratch	*w/o* sort	pair sort	$K!$ sort	video sort	ours
SSV2	64.5	67.0	67.4	67.2	64.8	**68.2**
K400	75.4	77.8	78.1	78.0	75.6	**78.8**

Table 6: The top-1 accuracy of different pre-training objectives and sort proxies under fine-tuning. The "pair sort" and "$K!$ sort" are alternatives for transcript sorting. The "video sort" indicates reorganizing shuffled video clips according to transcript knowledge, which is introduced in Merlot (Zellers et al., 2021). It aims to align video clips and the corresponding transcripts in the temporal dimension while failing to impose long-term temporal regularizations on the video encoder.

protocol. As reported in Table 5, our model achieves SOTA performance. The promising results show that our TVTS can also learn the association between video patterns and language semantics.

4.5 Ablation Study

Pre-training Objectives. In Table 6, we compare our proposed method with two variants on SSV2 and K400: (a) *scratch* that is initialized from ImageMAE and directly evaluated without pre-training. (b) *w/o sort* that is pre-trained by the contrastive objective only without TVTS. *w/o sort* outperforms *scratch*, which indicates that the natural language can be a promising supervision for video representation learning. Our method further improves the performance than *w/o sort*, which indicates the effectiveness of TVTS in learning spatiotemporal video representations.

Sort Proxy. Besides our method that casts the order prediction as a K-way classification task, we also tried three other strategies in modeling the ordering of the transcripts: (a) *pair sort* sorts the transcripts pairwisely by predicting the relative orders of the $K(K-1)/2$ transcript pairs. (b) *K! sort* predicts an overall ordering distribution by performing a $K!$-way classification ($K!$ possible orders given K transcripts), which adds an ordering token as the input of SortFormer. (c) *video sort* is similar to Merlot, which sorts the frames with ordered transcripts. The results are shown in Table 6. Both *pair sort* and *K! sort* drop performance, because the former ignores the overall relationship among the transcripts while the latter imposes the same penalty on the results with different number of transcripts that sorts incorrectly. But they still outperform *w/o sort*, indicating that sorting transcripts does help spatiotemporal representation learning. Our separate K-way classification modeling achieves the best performance. Furthermore, *video sort* lags far behind our method, since sorting video frames reduces the model's capability for long-term temporal reasoning.

Parameter Sensitivity. We further investigate the sensitivity of the masking ratio for TVTS and the temperature parameter τ in $\mathcal{L}_{\text{align}}$. The results are reported in Appendix A.1.

5 Conclusion and Discussion

In this work, we for the first time leverage script knowledge that is naturally tied with the video to benefit cognition-level spatiotemporal representation learning. We introduce a novel pretext task named Turning to Video for Transcript Sorting (TVTS), which regularizes the video encoder to learn transferable video representations for temporal commonsense reasoning. Extensive evaluations on downstream video tasks show the great superiority of our method.

Though helpful for long-term temporal understanding, we empirically observe the detriment of noisy ASR scripts to text encoder training as well as text-video alignment. Further studies are called for alleviating the noisy transcript problem. In future works, we would also like to evaluate the effectiveness of our method on other backbone architectures and scale-up datasets, and demonstrate the cognitive capabilities of our model in more applications.

REFERENCES

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, pp. 6644–6652, 2021.

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*, pp. 5803–5812, 2017.

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*, pp. 6836–6846, 2021.

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*, pp. 1728–1738, 2021.

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*, pp. 9244–9253, 2021.

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

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*, pp. 190–200, 2011.

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, pp. 1045–1053, 2021a.

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

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*, pp. 9640–9649, 2021b.

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.

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.

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*, pp. 1502–1512, 2021.

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*, pp. 4132–4141, 2022.

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.

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*, pp. 6824–6835, 2021.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

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.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pp. 214–229. Springer, 2020.

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*, pp. 16167–16176, 2022a.

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*, 2022b.

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*, pp. 5842–5850, 2017.

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*, pp. 9729–9738, 2020.

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*, pp. 16000–16009, 2022a.

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*, pp. 16000–16009, 2022b.

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

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*, pp. 7939–7949, 2021.

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*, pp. 8096–8105, 2021a.

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*, pp. 13886–13895, 2021b.

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

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*, pp. 4904–4916. PMLR, 2021.

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.

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.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pp. 2556–2563. IEEE, 2011.

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*, pp. 7331–7341, 2021.

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

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.

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*, pp. 7083–7093, 2019.

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*, pp. 8239–8249, 2021.

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.

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*, pp. 13708–13718, 2021.

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

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.

Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021.

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*, pp. 6964–6974, 2021.

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*, pp. 8748–8763. PMLR, 2021.

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*, pp. 2874–2884, 2022.

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*, pp. 3202–3212, 2015.

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.

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.

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.

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.

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*, pp. 11804–11813, 2021.

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.

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*, pp. 20–36. Springer, 2016.

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*, pp. 5288–5296, 2016.

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.

A Appendix

A.1 Additional Experiments

A.1.1 Downstream Action Recognition Datasets

The statistics of our downstream action recognition datasets are listed as follows: (a) **Something-Something V2** (SSV2) (Goyal et al., 2017) consists of 169K training videos and 20K validation videos belonging to 174 fine-grained action classes. (b) **Kinetics-400** (Kay et al., 2017) contains 240K training videos and 20K validations videos belonging to 400 classes. (c) **UCF-101** (Soomro et al., 2012) contains 9.5K/3.5K training and validation videos with 101 action classes. (d) **HMDB-51** (Kuehne et al., 2011) contains 3.5K/1.5K training and evaluation videos within 51 action classes.

A.1.2 Downstream Text-to-Video Retrieval Datasets

The statistics of our downstream text-to-video retrieval datasets are listed as follows: (a) **MSR-VTT** (Xu et al., 2016) contains 10K YouTube videos with 200K descriptions. 9K videos are used for training and the rest 1k videos are used for evaluation. (b) **DiDeMo** (Anne Hendricks et al., 2017) contains 10K Flickr videos with 40K descriptions. The training set has 9K videos, and the rest 1K videos are used for testing. (c) **MSVD** (Chen & Dolan, 2011) contains 1,970 YouTube videos with 80K descriptions. 1,200, 100, and 670 videos are used for training, validation, and testing respectively. (d) **LSMDC** (Rohrbach et al., 2015) consists of 118,081 video clips, in which 7,408 videos form the validation set and 1,000 videos form the test set.

A.1.3 Full Results for Text-to-Video Retrieval

Method	R@1	R@5	R@10	MedR
NoiseEst	17.4	41.6	53.6	8.0
MMT	26.6	57.1	69.6	4.0
SupportSet	30.1	58.5	69.3	3.0
Frozen	31.0	59.5	70.5	3.0
Ours	**34.6**	**61.5**	**72.2**	**3.0**

Table 7: MSR-VTT retrieval results.

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

Table 8: DiDeMo retrieval results.

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

Table 9: MSVD retrieval results.

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

Table 10: LSMDC retrieval results.

We compare our method with seven state-of-the-art methods (Amrani et al., 2021; Gabeur et al., 2020; Patrick et al., 2020; Bain et al., 2021; Li et al., 2020; Liu et al., 2019; Lei et al., 2021). The full Recall@K and MedR results are reported in Table 7, 8, 9, and 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.

A.1.4 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 (Hendricks & Nematzadeh, 2021), a zero-shot test benchmark for *subject*, *verb*, and *object* understanding in image. 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 precision results in terms of different values of ρ, shown in Table 11. Our model

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-video retrieval.

reaches higher precision on all concepts, which implies our learned spatiotemporal representations have strong out-of-the-box capabilities.

A.1.5 Ablation Study (Cont.)

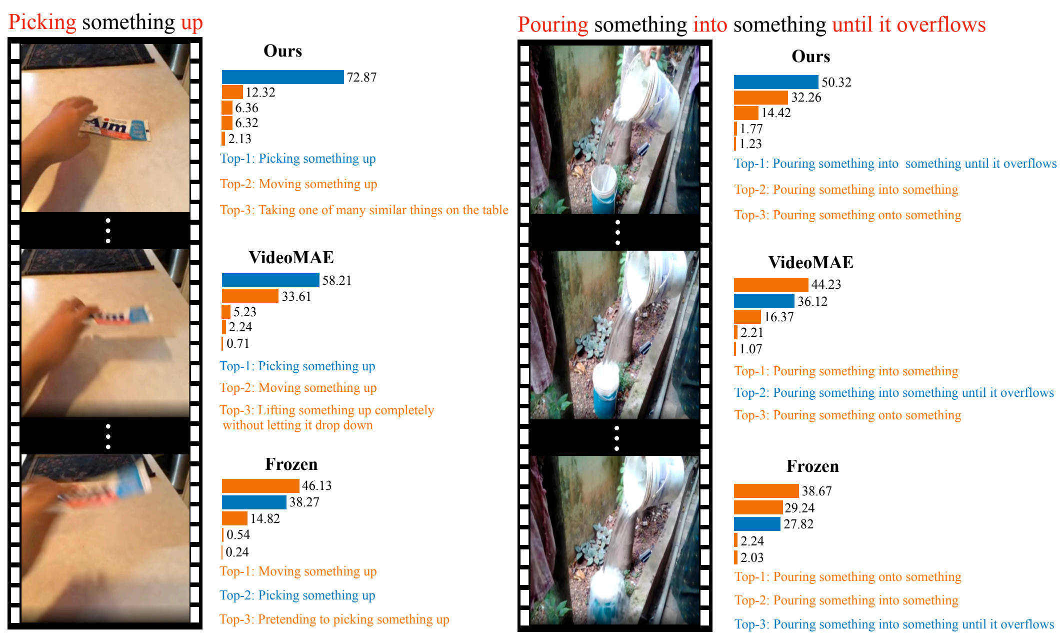
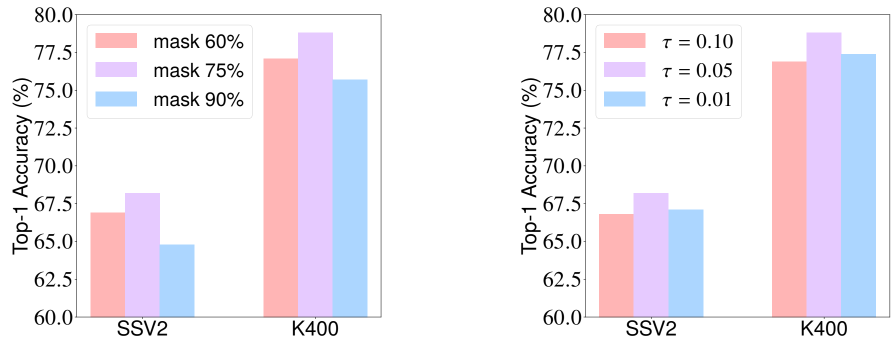


Figure 3: The visualization of top-5 prediction scores on SSV2, we normalize the scores to make their summation 100%. The blue and orange rows denote the scores of the right and wrong classes, respectively.

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 3. 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.

Masking Ratio. We compare different masking ratios for TVTS in Figure 4(a). Both lower (60%) and higher (90%) masking ratio drop performance than our method with 75% ratio. A lower masking ratio brings in temporal redundancy, while a higher ratio leads to the extremely limited knowledge for SortFormer to perform TVTS.



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

Figure 4: (a) The top-1 accuracy *w.r.t.* different masking ratio. (b) The top-1 accuracy *w.r.t.* different temperature parameter τ.

Temperature Parameter. We also investigate the influence of the temperature parameter τ in $\mathcal{L}_{\text{align}}$ in Figure 4(b). 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.

A.2 Hyper-parameters

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 12: 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 13: The linear probe and fine-tuning setup.

Our training hyper-parameters are listed in Tables 12 and 13. We mostly follow the setting of (Tong et al., 2022) for convenience. Carefully tuning these parameters may yield better performance.