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

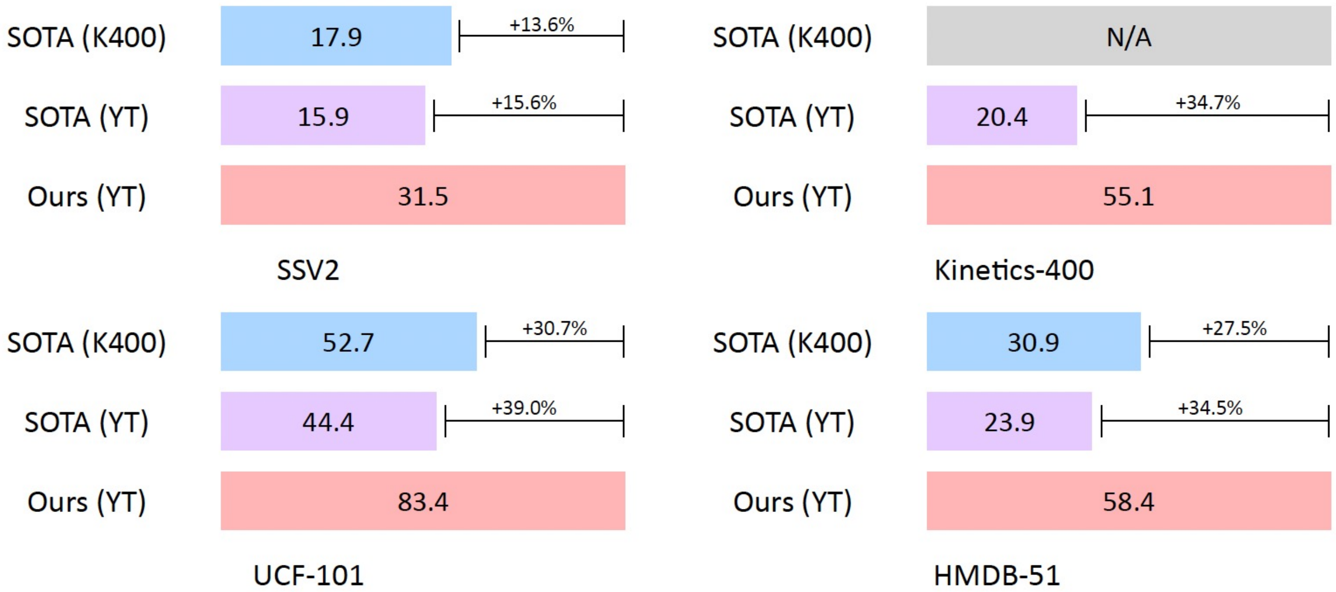
[*] equal contribution corresponding authors

https://github.com/TencentARC/TVTS

THU-PM-236

Existing Works

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA



Top-1 linear probe accuracy (%)

Data: highly curated videos Knowledge: pixel-level

Unsatisfactory out-of-the-box representations

How to use language semantics to boost transferable spatiotemporal representation learning?

Turning to Video for Transcript Sorting

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Implementation

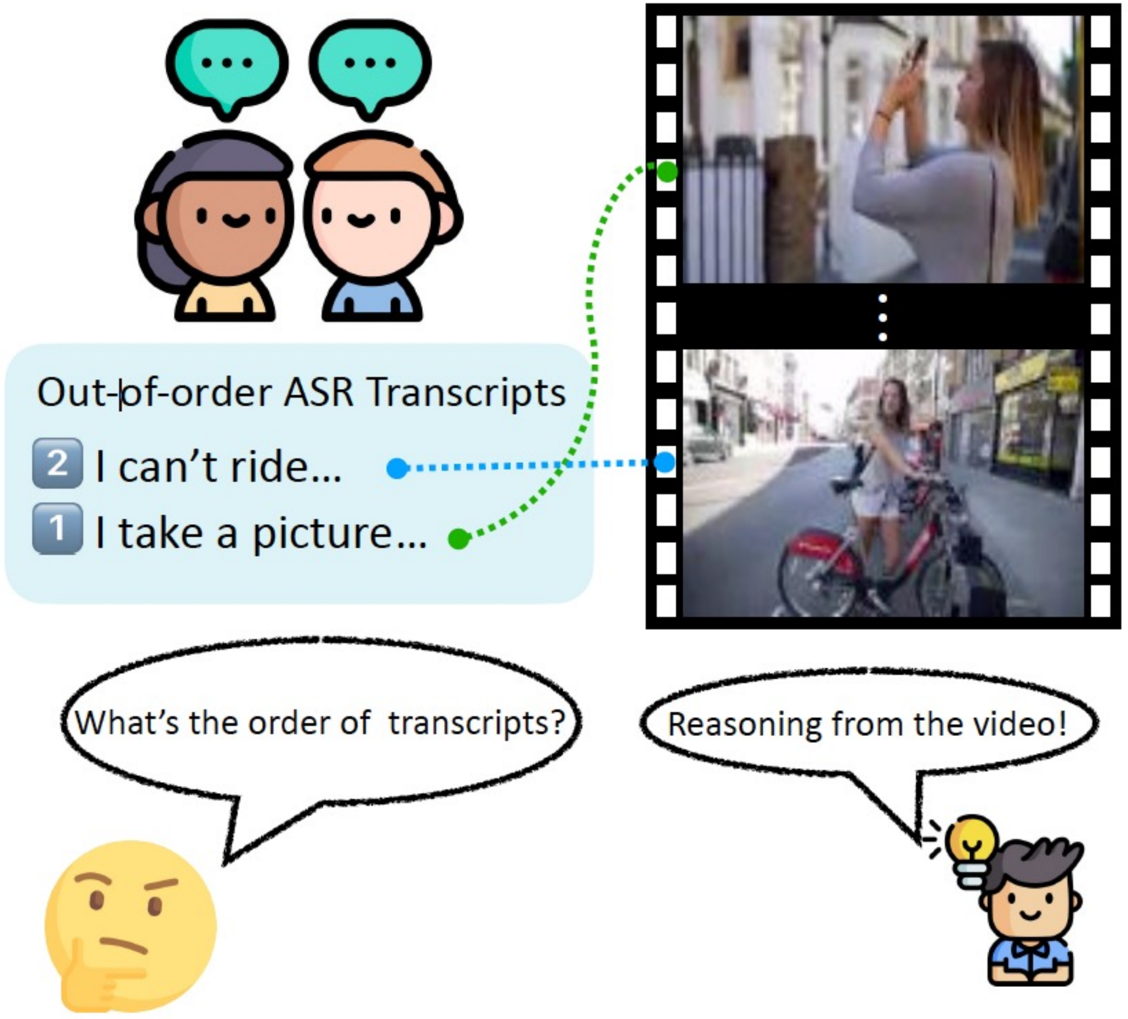
Sorting shuffled ASR transcripts by attending to learned video representations.

Motivation

Enforcing the model to contextualize what is happening over time so that it can re-organize the narrative transcripts.

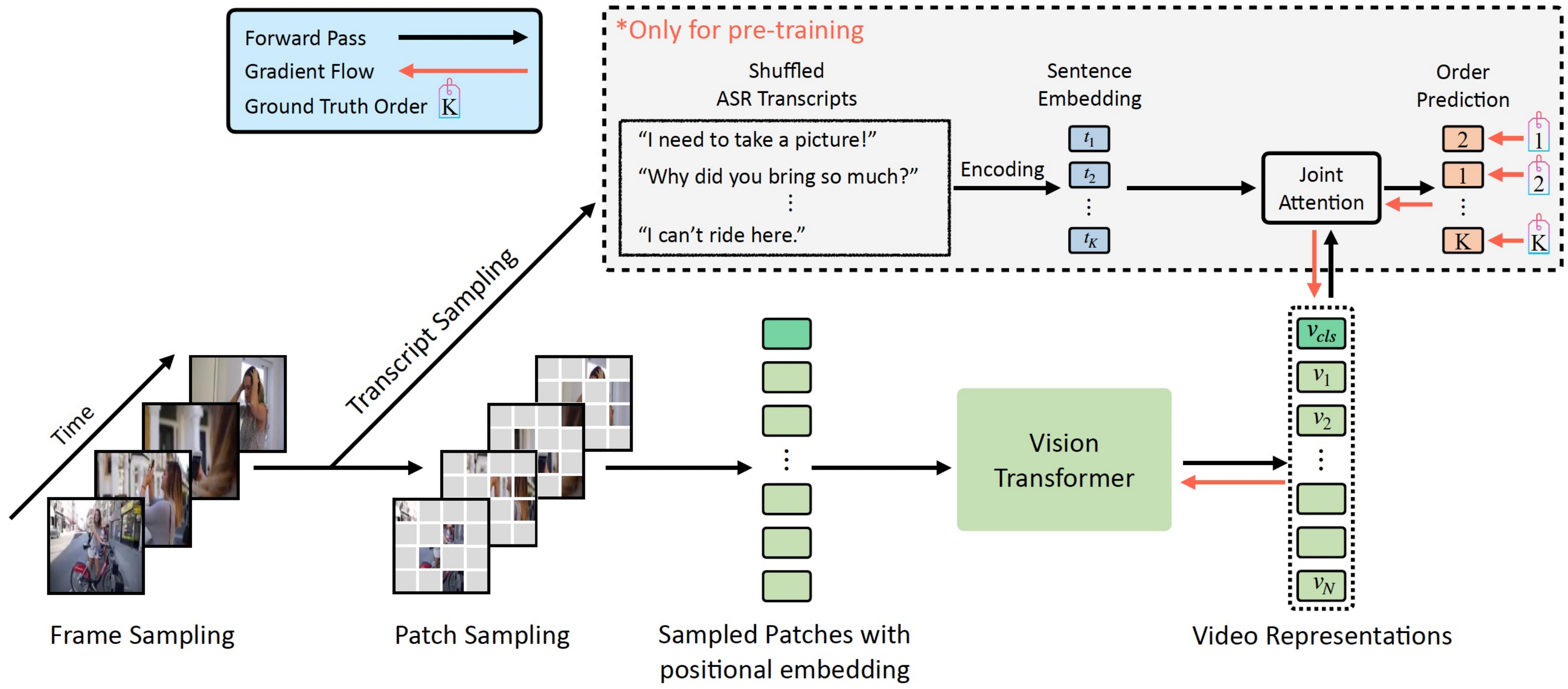
Merits

- ✓ Learn purely from video.
- ✓ Can seamlessly apply to large-scale *uncurated* video data in the real world.



TVTS illustration

Framework Overview



- Step1: Transcript and frame sampling
- Step2: Text and video feature extraction
- Step3: Turning to Video for Transcript Sorting

Transcript and Frame Sampling

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Sampled Frames

T_1: Wait a second, I can't ride in these sandals, oh that sucks.

T_2: God, this house looks so cute. Okay, I have to take a picture out.

T_3: I take like 200 pictures. Like how am I out of storage already? I don't even, I don't get it.

T_4: Okay, yeah that's not gonna fit. Why did I bring so much? I'm only here for five days.

Sampled Transcripts

Transcript Side: Consecutively sample K transcripts, each with a duration of l (in seconds), and an interval of 1s between adjacent transcripts.

Frame Side: Sample M frames between the beginning and ending time of all K transcripts.

Text and Video Feature Extraction

Text Side

- Encoder: DistilBERT-base.
- Pick the [CLS] tokens as the K unordered transcripts' representations $\{t_{o_i}\}_{i=1}^{K}$.

Video Side

- Encoder: ViT-base.
- Divide frames into patches, and randomly mask a large portion of them as the input.
- The video representations are denoted as $\{v_j\}_{j=0}^{N}$, where N denotes the number of unmasked video patches, and v_0 is the [CLS] token.
- We do not add the extra [MASK] token, and we have no explicit reconstruction target.

Turning to Video for Transcript Sorting

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

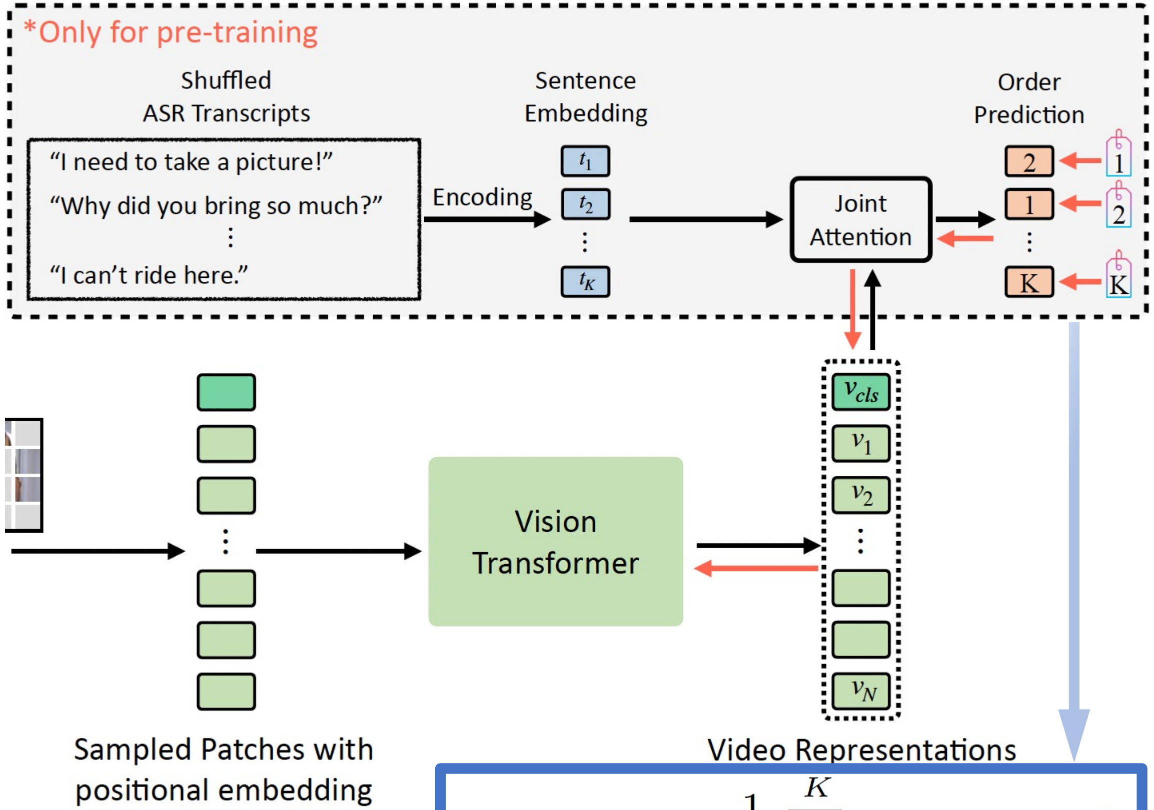
Our Expectation

Sorting the transcripts in the correct order by attending to the text and unmasked video tokens.

Implementation Details

Step1: Concatenate $\{t_{o_i}\}_{i=1}^{K}$ and $\{v_j\}_{j=0}^{N}$, and perform self-attention across all tokens.

Step2: The prediction of the transcript orders is modeled as a K-way classification task for each transcript, i.e., $\mathcal{L}_{\text{sort}}$.



$$\mathcal{L}_{\text{sort}} = -\frac{1}{K}\sum_{i=1}^{K}\log\text{softmax}(\hat{p}^i)$$

$$s.t. \quad \text{softmax}(\hat{p}^i) = \frac{\exp(p^i_{o_i})}{\sum_{j=1}^{K}\exp(p^i_j)},$$

Training Objectives

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- TVTS + Video-Text Contrastive ($\mathcal{L}_{\text{base}}$)

$$\mathcal{L}_{\text{sort}} = -\frac{1}{K}\sum_{i=1}^{K} \log \text{softmax}(\hat{p}^i) \quad + \quad \mathcal{L}_{\text{base}} = \text{NCE}(\hat{t}, \hat{v}) + \text{NCE}(\hat{v}, \hat{t})$$

$$s.t. \quad \text{softmax}(\hat{p}^i) = \frac{\exp(p^i_{o_i})}{\sum_{j=1}^{K} \exp(p^i_j)}, \qquad s.t. \quad \text{NCE}(q, k) = -\log \frac{\exp(q^\top k_+ / \tau)}{\sum_{i=1}^{B} \exp(q^\top k_i / \tau)},$$

- where $\hat{t} = \frac{1}{K}\sum_{i=1}^{K} t_i$, and $\hat{v}$ is the [CLS] token of the video, i.e., $\hat{v} \leftarrow v_0$
- Overall objective: $\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda \mathcal{L}_{\text{sort}}$ ($\lambda = 2$ in our practice)
- To prevent the model from learning shortcuts, we stop the gradients of $\mathcal{L}_{\text{sort}}$ from flowing toward encoding transcript features.

Sort Transcript or Video?

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. The model pre-trained only with $\mathcal{L}_{\text{base}}$ serves as the baseline.

➢ Sorting shuffled videos in pre-training is infeasible and counterintuitive for improving spatiotemporal representations.

Sort Implementation

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

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)
$\text{M}_{\text{sort}\backslash\text{sg}}$	✗	✓	✗	failed	failed
M_{sort}	✗	✓	✓	failed	failed
$\text{M}_{\text{ours}\backslash\text{sg}}$	✓	✓	✗	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 w.r.t. different objectives. “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 w.r.t. different ways to model TVTS. The model pre-trained with $\mathcal{L}_{\text{base}}$ serves as the baseline.

- TVTS can effectively regularize our model to learn transferable video representations.
- $\text{M}_{\text{ours}\backslash\text{sg}}$ drops performance, because the model learns from shortcuts.
- Both *Pairwise* and *Factorial* drop performance.

Transferability Evaluation

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

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.

Fine-tuning Performance

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**

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

Conclusion

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- 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.
- We introduce a novel pretext task for video pre-training, namely, *Turing to Video for Transcript Sorting (TVTS).* It promotes the capability of the model in learning transferable spatiotemporal video representations.
- We conduct comprehensive comparisons with advanced methods. Our pre-trained model exhibits strong out-of-the-box transferability on downstream tasks.

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

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

Code avaliable at

https://github.com/TencentARC/TVTS