Appendix for "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

A. Downstream Datasets

A.1. Action Recognition

The statistics of our downstream action recognition datasets are listed as follows: (a) **Something-Something V2** (SSV2) [8] 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** [10] contains 240K training videos and 20K validation videos belonging to 400 classes. (c) **UCF-101** [17] contains 9.5K/3.5K training and validation videos with 101 action classes. (d) **HMDB-51** [11] 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** [19] contains 10K YouTube videos with 200K descriptions. Following [3], we train on the training and validation set consisting of 9K videos and evaluate on the 1K-A test set. (b) **MSVD** [4] contains 1,970 YouTube videos with 80K descriptions, where each video has around 40 sentences. We adopt the official split [3], 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 [3, 6, 7] 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 [3, 6, 7]). (d) **LSMDC** [16] consists of 118,081 video clips harvested from 202 movies. We adopt the split of [3], where the validation and test set has 7,408 and 1,000 videos respectively.

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 1. 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 2. The linear probe and fine-tuning setup.

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 ex-

MSR-VTT					DiDeMo				
Method	R@1	R@5	R@10	MedR	Method	R@1	R@5	R@10	MedR
NoiseEst [1]	17.4	41.6	53.6	8.0	HERO [13]	2.1	-	11.4	-
MMT [5]	26.6	57.1	69.6	4.0	CE [14]	16.1	41.1	82.7	8.3
SupportSet [15]	30.1	58.5	69.3	3.0	ClipBert [12]	20.4	48.0	60.8	6.0
Frozen [3]	31.0	59.5	70.5	3.0	Frozen [3]	31.0	59.8	**72.4**	3.0
Ours	**34.6**	**61.5**	**72.2**	**3.0**	Ours	**32.4**	**59.8**	71.7	**3.0**
LSMDC					MSVD				
Method	R@1	R@5	R@10	MedR	Method	R@1	R@5	R@10	MedR
NoiseEst [1]	6.4	19.8	28.4	39.0	NoiseEst [1]	20.3	49.0	63.3	6.0
MMT [5]	12.9	29.9	40.1	19.3	SupportSet [15]	28.4	60.0	72.9	4.0
Frozen [3]	15.0	30.8	39.8	20.0	Frozen [3]	45.6	**79.8**	**88.2**	2.0
Ours	**17.2**	**32.8**	**41.7**	**17.0**	Ours	**45.9**	76.7	85.4	**2.0**

Table 3. 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 4. 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 5. 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 6. 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.

amining those with similarity above the threshold.

Our training hyper-parameters are listed in Table 1 and Table 2. We mostly follow the setting of [18] for convenience. Carefully tuning these parameters may yield better performance.

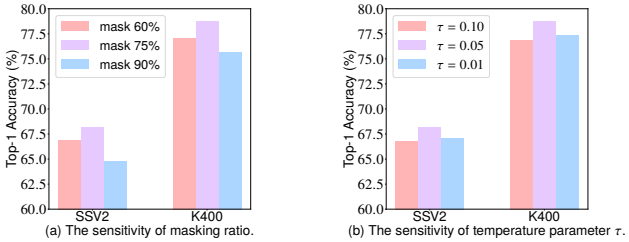

Figure 1. (a) The top-1 accuracy w.r.t. different masking ratio. (b) The top-1 accuracy w.r.t. different temperature parameter τ.

C. Additional Experiments

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

We compare our method with seven state-of-the-art methods [1, 3, 5, 12–15]. The full Recall@K and MedR results are reported in Table 3. 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 [9], 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

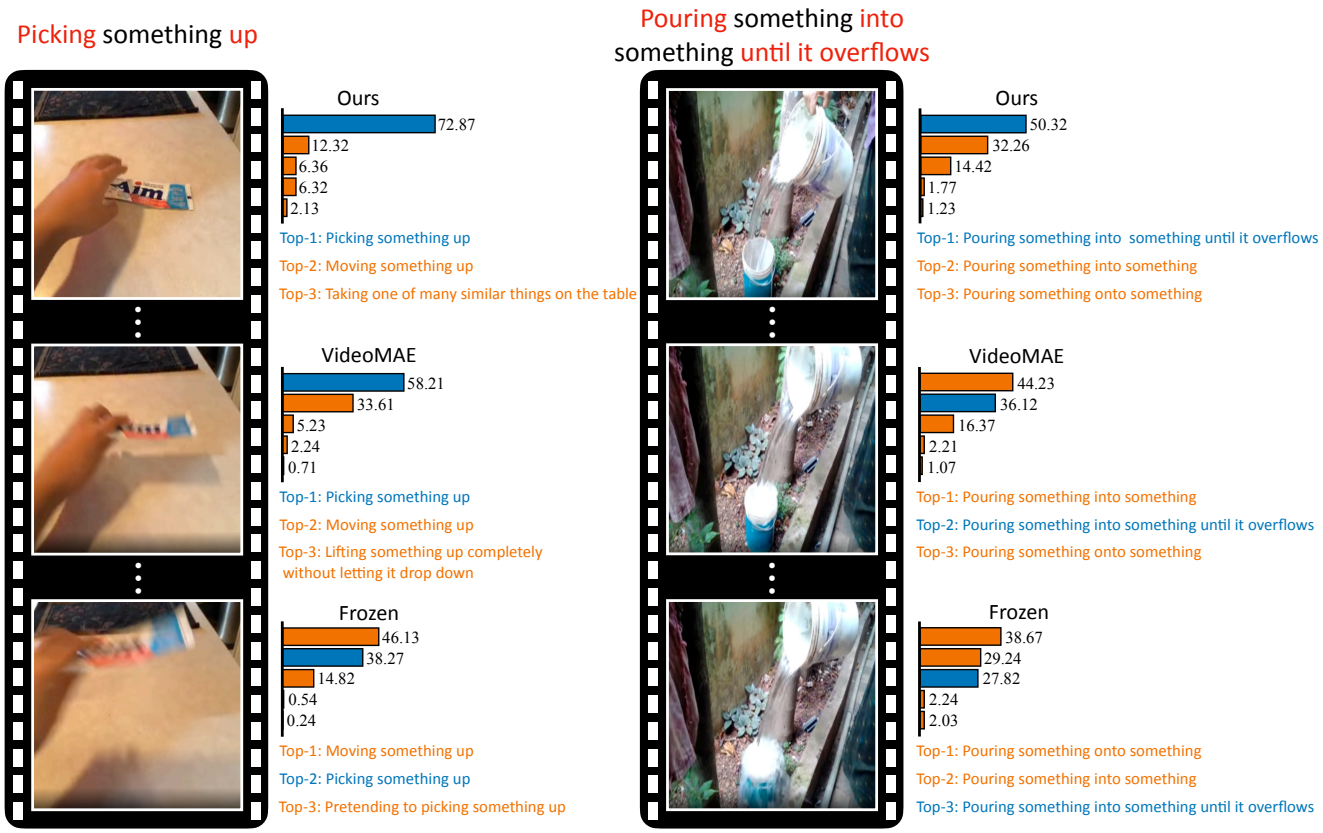

Figure 2. Visualization of the top-5 prediction scores on SSV2, we normalize the scores to make their summation 100%. The blue and orange rows denote the scores of the right and wrong classes, respectively.

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

C.3. Ablation Study (Cont.)

Contrastive Formulation. Since MERLOT [20] 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 5. 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 6, 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.

Masking Ratio. We compare different masking ratios for TVTS in Figure 1(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 1(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 2. 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.

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. 2

[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. 1

[3] 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. 1, 2

[4] 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. 1

[5] 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. 2

[6] 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. 1

[7] 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. 1

[8] 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

[9] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. 2

[10] 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

[11] 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

[12] 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. 2

[13] 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. 2

[14] 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. 2

[15] 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. 2

[16] 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. 1

[17] 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

[18] 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. 2

[19] 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. 1

[20] 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. 3