Supplementary Materials for DiST

Zhiwu Qing[1] Shiwei Zhang[2*] Ziyuan Huang[3] Yingya Zhang[2]
Changxin Gao[1] Deli Zhao[2] Nong Sang[1*]
[1]Key Laboratory of Image Processing and Intelligent Control
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Alibaba Group [3]ARC, National University of Singapore
{qzw, cgao, nsang}@hust.edu.cn {zhangjin.zsw, yingya.zyy}@alibaba-inc.com
ziyuan.huang@u.nus.edu zhaodeli@gmail.com

Overview

In this supplementary material, we first provide more ablation studies, and comparison with state-of-the-art approaches on zero-shot task and video recognition task in Sec. 1. Then, the implementation details for SSV2, K400 and EK100 are presented in Sec. 2.

1. Additional Results

1.1. Ablation Studies

Method	Pretraining	Frozen	SSV2	K400
EVL [15]	ImageNet-21k	✓	N/A	75.4
ST-Adapter [18]	ImageNet-21k	✓	62.8	76.6
TimeSformer [2]	ImageNet-21k	✗	59.5	78.0
X-ViT [3]	ImageNet-21k	✗	64.4	78.5
DiST	ImageNet-21k	✓	**66.8**	**79.8**
EVL [15]	CLIP	✓	61.0	82.9
ST-Adapter [18]	CLIP	✓	66.3	82.0
DiST	CLIP	✓	**68.7**	**83.6**

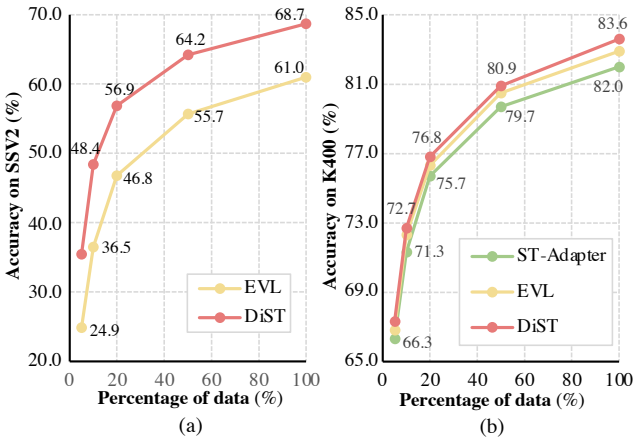
Table A1: Comparsion with state-of-the-art under different pre-trained image encoders.

Fine-tuning with ImageNet pretrained models. In fact, our DiST is not limited to CLIP pre-trained image models, and thus we have attempted to explore fine-tuning video models based on ImageNet supervised pre-training by following existing methods [15, 18]. As shown in Tab. A1, DiST outperforms the similar frozen CLIP-based fine-tuning method, *i.e.*, EVL [15] by 4.4 %. Compared to the adapter-based approach, *i.e.*, ST-Adapter [18], we exceed by 4.0% and 2.2% on SSV2 and K400 datasets, respectively. Besides, DiST also shows performance advantages over full fine-tuning methods, such as TimeSformer [2] and X-ViT [3]. These results indicate that DiST is a more general network for image-to-video transfer learning.

Frame	1-4	5-8	9-12	SSV2	K400
✗	✗	✗	✓	66.0	83.0
✗	✗	✓	✓	67.8	83.3
✗	✓	✓	✓	68.0	83.4
✓	✓	✓	✓	**68.7**	**83.6**

Table A2: Different layer of features.

Different depth features. As shown in Tab. A2, if only using the deep features (*i.e.*, 9-12) of ViT, its performance on the temporally heavy dataset (*i.e.*, SSV2) is weak. Nevertheless, the introduction of low-level features can bring up to 2.8% gains at most. This implies that the rich temporal details in the lower-level features are more beneficial for spatio-temporal learning.



(a) (b)

Figure A1: Performance comparison on varying training data scales.

Data efficiency. Data efficiency refers to the utilization efficiency of limited data by fine-tuning the models with only a portion of the training data. In Fig. A1, DiST and other clip-

Method	Pre-train	Architecture	Input Size	FLOPs×Cr.×Cl. (T)	Param (M)	Frozen	Top-1	Top-5
SlowFast [8]	ImageNet-21K	R101+NL	16×224^2	$0.1 \times 3 \times 1$	60	✗	63.1	87.6
ViViT FE [1]	IN21K+K400	ViT-L	16×224^2	$1.0 \times 3 \times 4$	612	✗	65.4	89.8
TAdaConvNeXt-T [10]	ImageNet-1K	ConvNeXt-T	32×224^2	$0.1 \times 3 \times 2$	38	✗	67.1	90.4
MTV-B(320p) [24]	IN21K+K400	-	32×224^2	$0.9 \times 3 \times 4$	310	✗	68.5	90.4
MViT [7]	Kinetics-600	MViT-B-24	32×224^2	$0.2 \times 3 \times 1$	53	✗	68.7	91.5
Video Swin [16]	IN21K+K400	Swin-B	32×224^2	$0.3 \times 3 \times 1$	60	✗	69.6	92.7
UnifromerV2 [14]	CLIP-400M	ViT-B	32×224^2	$0.37 \times 1 \times 3$	163	✗	70.7	93.2
EVL❄ [15]	CLIP-400M	ViT-B	32×224^2	$0.68 \times 1 \times 3$	175	✓	62.4	-
ST-Adapter❄ [18]	CLIP-400M	ViT-B	32×224^2	$0.61 \times 1 \times 3$	93	✓	69.5	**92.6**
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-B	8×224^2	$0.16 \times 1 \times 3$	105	✓	68.7	91.1
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-B	16×224^2	$0.32 \times 1 \times 3$	105	✓	70.2	92.0
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-B	32×224^2	$0.65 \times 1 \times 3$	105	✓	**70.9**	92.1
UnifromerV2 [14]	CLIP-400M	ViT-L	32×224^2	$1.73 \times 1 \times 3$	574	✗	73.0	94.5
TAdaFormer [11]	CLIP-400M	ViT-L	32×224^2	$1.70 \times 2 \times 3$	364	✗	73.6	-
EVL❄ [15]	CLIP-400M	ViT-L	32×224^2	$3.21 \times 1 \times 3$	654	✓	66.7	-
ST-Adapter❄ [18]	CLIP-400M	ViT-L	32×224^2	$2.75 \times 1 \times 3$	347	✓	72.3	**93.9**
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	8×224^2	$0.71 \times 1 \times 3$	336	✓	70.8	92.3
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	16×224^2	$1.42 \times 1 \times 3$	336	✓	72.5	93.0
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	32×224^2	$2.83 \times 1 \times 3$	336	✓	**73.1**	93.2
EVL❄ [15]	CLIP-400M	ViT-L	32×336^2	$8.08 \times 1 \times 3$	654	✓	68.0	-
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	8×336^2	$1.66 \times 1 \times 3$	336	✓	71.2	92.5
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	16×336^2	$3.32 \times 1 \times 3$	336	✓	72.6	93.0
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	32×336^2	$6.64 \times 1 \times 3$	336	✓	**73.3**	**93.5**

Table A3: Comparison with the state-of-the-art methods on Something-Something V2. "Cr." and "Cl." are the abbreviation for "spatial crops" and "temporal clips". "Frozen" indicates freezing the CLIP pre-trained parameters.

based pre-training methods [18, 15] are compared in terms of data efficiency on SSV2 and K400 datasets. As can be seen, our DiST exhibits consistent advantages over existing ST-Adapter [18] and EVL [15] across different proportions of training data. Especially on the temporally dependent dataset, *i.e.*, SSV2, DiST demonstrates improvements of approximately 10% over EVL [15] with fewer training scales. This suggests that DiST is easier to fine-tune and shows better generalization capability.

Method	Model	Frames	HMDB51	UCF101
ActionCLIP [21]	B/16	32×1×1	40.8±5.4	58.3±3.4
X-CLIP [17]	B/16	32×1×1	44.6±5.2	72.0±2.3
DiST$_{\gamma=2}$❄	B/16	32×1×1	55.4±1.2	72.3±0.6
†**DiST**$_{\gamma=2}$❄	B/16	32×1×1	57.4±0.9	73.2±0.6
DiST$_{\gamma=2}$❄	L/14	32×1×1	57.5±1.6	74.9±0.8
†**DiST**$_{\gamma=2}$❄	L/14	32×1×1	**61.8**±1.3	**75.8**±0.7

Table A4: Zero-shot performance on HMDB51 [12] and UCF101 [19] across three splits. † indicates Kinetics-710 [13, 4, 5] pre-trained models.

1.2. Comparison with the state-of-the-art methods

Zero-shot accuracy with Kinetics-710 pre-training. We further evaluate the zero-shot performance of DiST on HMDB51 [12] and UCF101 [19] with the large-scale video dataset, *i.e.*, Kinetics-710 [13, 4, 5], pre-trained models. As shown in Tab. A4, we can observe that regardless of the model size, remarkable improvements can be achieved by fine-tuning our lightweight temporal encoder and integration branch on Kinetics-710. Particularly, on HMDB51 that relies on temporal information, the gains of ViT-B and ViT-L can reach 2.0% and 4.3%, respectively. This further demonstrates the scalability of DiST on both data scale and model size.

More results on Something-Something V2 [9] and Kinetics-400 [13]. Here, we supplement more results with different frames and resolutions on two datasets in Tab. A3 and Tab. A5 for reference. From the two tables, we can draw the following conclusions: (i) The more temporal details brought by more frames is highly effective for both SSV2 and K400, regardless of model size and pre-training datasets used. For example, increasing the frames from 8 to 32 can result in consistent performance improvements of around 2.0% on SSV2 and around 1.5% on K400. This is

Method	Pre-train	Architecture	Input Size	TFLOPs×Cr.×Cl.	Param (M)	Frozen	Top-1	Top-5
TAda [10]	ImageNet-1K	ConvNeXt-T	32×224^2	$0.1 \times 3 \times 2$	38	✗	79.1	93.7
SlowFast [8]	-	R101+NL	16×224^2	$0.4 \times 3 \times 10$	60	✗	79.8	93.9
TimeSformer [2]	ImageNet-21K	ViT-L	96×224^2	$8.4 \times 3 \times 1$	430	✗	80.7	94.7
MViT [7]	-	MViT-B	64×224^2	$0.5 \times 1 \times 5$	37	✗	81.2	95.1
ViViT FE [1]	ImageNet-21K	ViT-L	128×224^2	$4.0 \times 3 \times 1$	N/A	✗	81.7	93.8
Video Swin [16]	ImageNet-21K	Swin-L	32×224^2	$0.6 \times 3 \times 4$	197	✗	83.1	95.9
UnifromerV2 [14]	CLIP-400M+K710	ViT-B	8×224^2	$0.13 \times 1 \times 3$	115	✗	85.2	96.7
ST-Adapter❄ [18]	CLIP-400M	ViT-B	32×224^2	$0.61 \times 1 \times 3$	93	✓	82.7	96.2
EVL❄ [15]	CLIP-400M	ViT-B	32×224^2	$0.59 \times 1 \times 3$	115	✓	84.2	-
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-B	8×224^2	$0.16 \times 1 \times 3$	112	✓	83.6	96.3
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-B	16×224^2	$0.32 \times 1 \times 3$	112	✓	84.4	96.7
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-B	32×224^2	$0.65 \times 1 \times 3$	112	✓	85.0	97.0
DiST$_{\gamma=2}$❄	CLIP-400M+K710	ViT-B	8×224^2	$0.16 \times 1 \times 3$	112	✓	85.1	96.8
DiST$_{\gamma=2}$❄	CLIP-400M+K710	ViT-B	16×224^2	$0.32 \times 1 \times 3$	112	✓	85.8	97.2
DiST$_{\gamma=2}$❄	CLIP-400M+K710	ViT-B	32×224^2	$0.65 \times 1 \times 3$	112	✓	**86.8**	**97.5**
UnifromerV2 [14]	CLIP-400M+K710	ViT-L	32×224^2	$2.66 \times 2 \times 3$	354	✗	89.3	98.2
TAdaFormer [11]	CLIP-400M+K710	ViT-L	32×224^2	$1.41 \times 4 \times 3$	364	✗	89.5	-
ST-Adapter❄ [18]	CLIP-400M	ViT-L	32×224^2	$2.75 \times 1 \times 3$	347	✓	87.2	97.6
EVL❄ [15]	CLIP-400M	ViT-L	32×224^2	$2.70 \times 1 \times 3$	363	✓	87.3	-
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	8×224^2	$0.71 \times 1 \times 3$	343	✓	86.9	97.6
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	16×224^2	$1.42 \times 1 \times 3$	343	✓	87.6	97.8
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	32×224^2	$2.83 \times 1 \times 3$	343	✓	88.0	97.9
DiST$_{\gamma=2}$❄	CLIP-400M+K710	ViT-L	8×224^2	$0.71 \times 1 \times 3$	343	✓	87.8	97.9
DiST$_{\gamma=2}$❄	CLIP-400M+K710	ViT-L	16×224^2	$1.42 \times 1 \times 3$	343	✓	88.6	98.2
DiST$_{\gamma=2}$❄	CLIP-400M+K710	ViT-L	32×224^2	$2.83 \times 1 \times 3$	343	✓	**89.5**	**98.4**
X-CLIP [17]	CLIP-400M	ViT-L	16×336^2	$3.09 \times 3 \times 4$	354	✗	87.7	97.4
BIKE [23]	CLIP-400M	ViT-L	32×336^2	$3.73 \times 3 \times 4$	230	✗	88.6	98.3
EVL❄ [15]	CLIP-400M	ViT-L	32×336^2	$6.07 \times 1 \times 3$	363	✓	87.7	-
Text4Vis❄ [22]	CLIP-400M	ViT-L	32×336^2	$3.83 \times 1 \times 3$	231	✓	87.8	97.6
UnifromerV2❄ [14]	CLIP-400M+K710	ViT-L	32×336^2	$6.27 \times 1 \times 3$	354	✓	88.8	98.1
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	8×336^2	$1.66 \times 1 \times 3$	343	✓	87.2	97.6
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	16×336^2	$3.32 \times 1 \times 3$	343	✓	87.9	98.0
DiST$_{\gamma=2}$❄	CLIP-400M	ViT-L	32×336^2	$6.64 \times 1 \times 3$	343	✓	88.5	98.2
DiST$_{\gamma=2}$❄	CLIP-400M+K710	ViT-L	8×336^2	$1.66 \times 1 \times 3$	343	✓	88.1	97.9
DiST$_{\gamma=2}$❄	CLIP-400M+K710	ViT-L	16×336^2	$3.32 \times 1 \times 3$	343	✓	88.9	98.2
DiST$_{\gamma=2}$❄	CLIP-400M+K710	ViT-L	32×336^2	$6.64 \times 1 \times 3$	343	✓	**89.7**	**98.5**

Table A5: Comparison with state-of-the-arts on Kinetics-400.

because SSV2 relies more on temporal details, while K400 depends more on spatial information. (ii) Increasing the resolution from 224 to 336 has a relatively small impact on performance, typically around 0.3%, on both datasets, which is consistent with existing literature [15]. Interestingly, using higher resolutions is apparently less effective than increasing the size of the pre-training datasets or models. Therefore, in future research, we will further explore the potential of video models from both the model and data scale perspectives.

2. Implementation details

Tab. A6 summarizes the fine-tuning configurations across multiple datasets and models. In almost all the different datasets and models, the configurations are shared, which demonstrates the extensive adaptability of our proposed DiST.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. 2, 3

dataset	Kinetics-400		Something-Something V2		Epic-Kitchens 100	
backbone	ViT-B	ViT-L	ViT-B	ViT-L	ViT-B	ViT-L
α (dim. of integration branch)	1/2 (384)	3/8 (384)	1/2 (384)	3/8 (384)	1/2 (384)	3/8 (384)
β (dim. of temporal encoder)	1/8 (96)	3/32 (96)	1/8 (96)	3/32 (96)	1/8 (96)	3/32 (96)
γ (frames of temporal encoder)	2					
optimizer	AdamW, learning rate=3.2e-4, weight decay=1e-4, betas=[0.9, 0.999]					
batch size	256					
training epochs	36					
warmup epochs	6					
training crop scale	[0.4, 1.0]					
training crop size	224					
frame sampling rate	TSN [20] uniform sampling					
mirror	✓					
RandAugment [6]	✗					
MixUp [26]	0.8					
CutMix [25]	1.0					
testing views	3 temporal $\times$ 1 spatial					

Table A6: **Configurations for Kinetics-400, Something-Something V2 and Epic-Kitchens 100.**

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 1, 3

[3] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *NeurIPS*, 34:19594–19607, 2021. 1

[4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2

[5] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 2

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 702–703, 2020. 4

[7] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6824–6835, 2021. 2, 3

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 2, 3

[9] 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 *ICCV*, pages 5842–5850, 2017. 2

[10] Ziyuan Huang, Shiwei Zhang, Liang Pan, Zhiwu Qing, Mingqian Tang, Ziwei Liu, and Marcelo H Ang Jr. Tada! temporally-adaptive convolutions for video understanding. In *ICLR*, 2022. 2, 3

[11] Ziyuan Huang, Shiwei Zhang, Liang Pan, Zhiwu Qing, Yingya Zhang, Ziwei Liu, and Marcelo H Ang Jr. Temporally-adaptive models for efficient video understanding. *arXiv preprint arXiv:2308.05787*, 2023. 2, 3

[12] H Jhuang, H Garrote, E Poggio, T Serre, and T Hmdb. A large video database for human motion recognition. In *ICCV*, volume 4, page 6, 2011. 2

[13] 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. 2

[14] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 2, 3

[15] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *ECCV*, pages 388–404. Springer, 2022. 1, 2, 3

[16] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 2, 3

[17] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, pages 1–18. Springer, 2022. 2, 3

[18] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. Parameter-efficient image-to-video transfer learning. *arXiv e-prints*, pages arXiv–2206, 2022. 1, 2, 3

[19] 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. 2

[20] 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 *ECCV*, pages 20–36. Springer, 2016. 4

[21] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[22] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *AAAI*, volume 37, pages 2847–2855, 2023. 3

[23] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *CVPR*, pages 6620–6630, 2023. 3

[24] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, pages 3333–3343, 2022. 2

[25] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 4

[26] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 4