Supplementary Material: Inductive and Transductive Few-Shot Video Classification via Appearance and Temporal Alignments

Khoi D. Nguyen[1], Quoc-Huy Tran[2], Khoi Nguyen[1], Binh-Son Hua[1], and Rang Nguyen[1]⋆

[1] VinAI Research, Vietnam
[2] Retrocausal, Inc., USA

In this supplementary material, we first elaborate the difference between datasets that are sensitive or insensitive to action ordering in Sec. A. In Sec. B, we provide additional experiment results of our method on Something-Something V2 dataset. We then evaluate the performance of our method on two action recognition benchmarks, namely UCF-101 and HMDB-51 in Sec. C.

A Order-Sensitive Datasets vs Order-Insensitive Datasets

Here, we discuss the difference between datasets that are sensitive or insensitive to action ordering. For order-sensitive datasets, temporal cues (e.g., temporal order-preserving prior [16,2,3,9]) are essential in distinguishing between video categories. For example, in Something-Something V2, deciding whether a video belongs to "Pulling sth from left to right" or "Pushing sth from right to left" must consider the positional changes of an object presented in the video. Fig. S1 shows how videos from the two classes in Something-Something V2 can be aligned with each other. On the other hand, videos in order-insensitive datasets like Kinetics loosely rely on temporal cues. Their video classes can mostly be distinguished just with the appearance information.

B Additional Results

Running time. During training, we use the same ResNet-50 encoder with a linear classifier on top as the baseline [20], yielding the same number of parameters. At inference stage, we discard the linear classifier and use the trained ResNet-50 to extract frame features. In our implementation, our method has the same number of epochs (25 epochs) and roughly the same training (12 hours) and inference time (0.2 and 0.5 secs for a 1-shot and 5-shot episode) as the baseline.
Ablation results of α and ν. We perform ablation study on the two hyper-parameter α and ν. We first fix $\nu = 0.1$ and vary the value of α in the range $[0.01, 1.0]$. Results are shown in Fig. S2(a). As we can observe, increasing the

⋆ Corresponding author

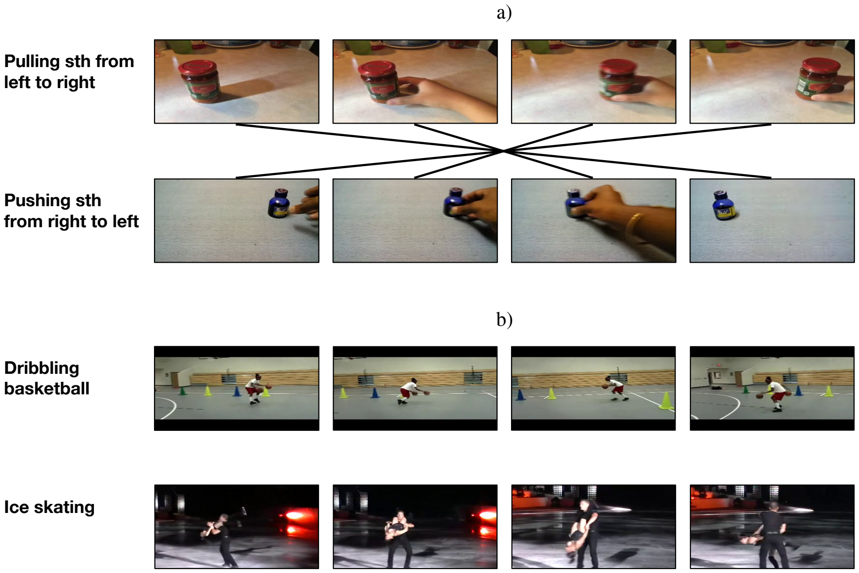
a)
Pulling sth from left to right
Pushing sth from right to left
b)
Dribbling basketball
Ice skating

Fig. S1. a) Example videos from Something-Something V2. The two videos can be aligned with each other in terms of appearance. To classify correctly, the model must consider the order of video frames. b) Example videos from Kinetics. Classes of Kinetics are mostly different in appearance and sometimes do not have a fixed frame order.

value of α improves the performance consistently. We then fix $\alpha = 0.05$ and vary the value of ν in the range $[0.01, 1.0]$. Results are shown in Fig. S2(b). $\nu = 0.1$ achieves the best results for both 1-shot and 5-shot settings.

Additional qualitative results. We provide additional qualitative results of inductive inferences on 2-way 1-shot tasks of Something-Something V2 in Fig. S3.

C Few-Shot Action Recognition Results on UCF-101 and HMDB-51

So far we have focused on the problem of few-shot video classification. We now consider a related problem of few-shot action recognition [17,10,19,1]. The main difference between video classification and action recognition is that, in video classification, videos/classes can describe general contents (e.g., “sled dog racing” in Kinetics), which are not limited to human actions as in action classification. In this section, we evaluate the performance of our method on two few-shot action recognition datasets, including UCF-101 [15] and HMDB-51 [8]. UCF-101 is an action recognition dataset consisting of 13,320 YouTube videos of 101 action classes. For HMDB-51, there are 6,849 videos collected from different sources, i.e., movies, Prelinger Archive, YouTube, and Google videos. For both datasets,

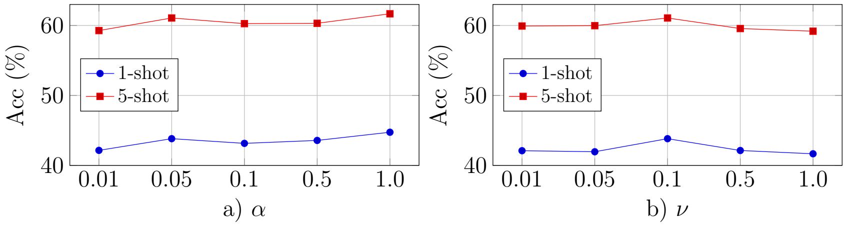

Fig. S2. Ablation results of a) α and b) ν.

we follow the splits from [19], which divide UCF-101 into 70/10/21 classes and HMDB-51 into 31/10/10 classes for training/validation/testing respectively.
Implementation Details. We apply the same preprocessing steps that we use for Kinetics and Something-Something V2 datasets to UCF-101 and HMDB-51 datasets. Specifically, a video is divided into $M = 8$ segments and a frame is sampled randomly from each segment. We use the SGD optimizer with an initial learning rate of 0.0005 for both datasets. For UCF-101, the model is trained for 30 epochs, and we reduce the learning rate to 10^{-4}, 10^{-5}, and 10^{-6} at epochs 15, 20, and 25 respectively. For HMDB-51, the model is trained for 25 epochs, and the learning rate is reduced to 10^{-4}, 10^{-5}, and 10^{-6} at epochs 10, 15, and 20 respectively.

C.1 The Relative Importance of Appearance and Temporal Cues on UCF-101 and HMDB-51

We investigate the effectiveness of the hyperparameter β (in Eq. 9 in the main paper) in the inductive and transductive settings on the UCF-101 and HMDB-51 datasets. Tabs. S1 and S2 provide the mean accuracy with 95% confidence interval on 10,000 episodes sampled from the validation set.

Tab. S1 shows that, in the inductive setting, appearance cues are more important than temporal cues for both datasets. In addition, utilizing both appearance and temporal cues ($\beta \in [0.2, 0.4]$) yields minor improvements over using appearance cues only ($\beta = 0$). Similarly, for the transductive results in Tab. S2, appearance cues play a more important role than temporal cues for both datasets. Moreover, leveraging both appearance and temporal cues ($\beta \in [0.2, 0.4]$) leads to significant improvements on HMDB-51 but marginal performance gains on UCF-101, as compared to using appearance cues only ($\beta = 0$).

C.2 Comparison with Previous Few-Shot Action Recognition Methods

We compare our method with previous works on the two benchmarks UCF-101 and HMDB-51.

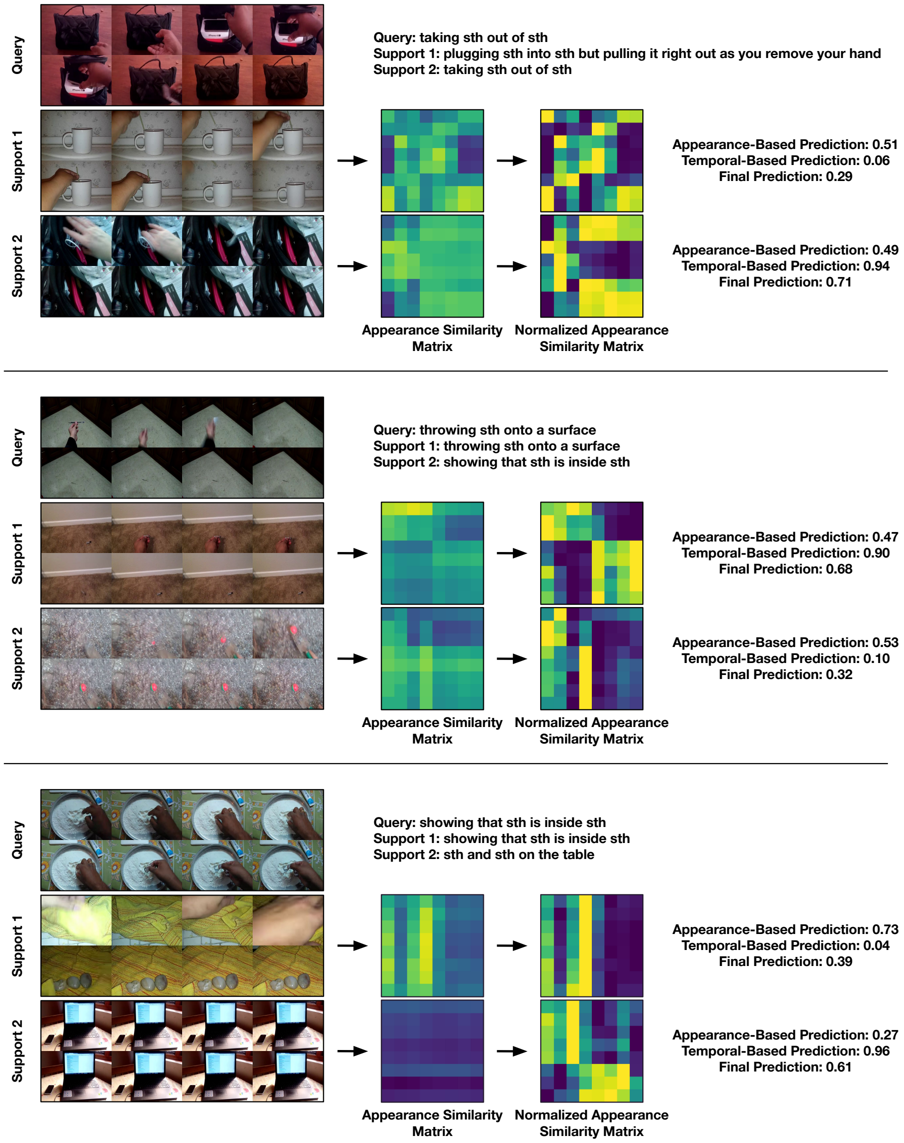
Query
Support 1
Support 2
Query: taking sth out of sth
Support 1: plugging sth into sth but pulling it right out as you remove your hand
Support 2: taking sth out of sth
Appearance-Based Prediction: 0.51
Temporal-Based Prediction: 0.06
Final Prediction: 0.29
Appearance-Based Prediction: 0.49
Temporal-Based Prediction: 0.94
Final Prediction: 0.71
Appearance Similarity Matrix
Normalized Appearance Similarity Matrix
Query
Support 1
Support 2
Query: throwing sth onto a surface
Support 1: throwing sth onto a surface
Support 2: showing that sth is inside sth
Appearance-Based Prediction: 0.47
Temporal-Based Prediction: 0.90
Final Prediction: 0.68
Appearance-Based Prediction: 0.53
Temporal-Based Prediction: 0.10
Final Prediction: 0.32
Appearance Similarity Matrix
Normalized Appearance Similarity Matrix
Query
Support 1
Support 2
Query: showing that sth is inside sth
Support 1: showing that sth is inside sth
Support 2: sth and sth on the table
Appearance-Based Prediction: 0.73
Temporal-Based Prediction: 0.04
Final Prediction: 0.39
Appearance-Based Prediction: 0.27
Temporal-Based Prediction: 0.96
Final Prediction: 0.61
Appearance Similarity Matrix
Normalized Appearance Similarity Matrix

Fig. S3. Qualitative Results of 2-way 1-shot Tasks on Something-Something V2. For each task, we present the appearance similarity matrix $\mathbf{D}$ between the query video and each support video in the second column. In the third column, we show the row-wise normalized version $\tilde{\mathbf{D}}$. Finally, we show the predictions of the two similarity scores and the final prediction. Ground truth class labels are shown at the top.

Inductive. Inductive results are presented in Tab. S3. The competing methods include from few-shot action recognition approaches, i.e., FAN [17], Proto-

Table S1. Ablation study on the relative importance of the appearance and temporal terms in computing the predictive distribution of the **inductive** inference on the UCF-101 and HMDB-51 datasets. Results are mean accuracy with 95% confidence interval on 10,000 episodes sampled from validation set.

β	**UCF-101**		**HMDB-51**	
	1-shot	**5-shot**	**1-shot**	**5-shot**
1.0	24.96 ± 0.37	30.17 ± 0.39	24.77 ± 0.37	32.24 ± 0.39
0.8	35.32 ± 0.41	80.24 ± 0.37	30.13 ± 0.39	53.14 ± 0.43
0.6	67.90 ± 0.41	94.47 ± 0.20	49.52 ± 0.42	79.58 ± 0.34
0.4	84.55 ± 0.31	**95.19 ± 0.19**	**65.78 ± 0.38**	**82.27 ± 0.32**
0.2	**84.63 ± 0.31**	95.16 ± 0.19	65.77 ± 0.38	82.26 ± 0.32
0.0	84.60 ± 0.31	95.15 ± 0.19	65.72 ± 0.38	82.21 ± 0.32

Table S2. Ablation study on the relative importance of the appearance and temporal terms in computing the assignment function and the predictive distribution of the **transductive** inference on the UCF-101 and HMDB-51 datasets. Results are mean accuracy with 95% confidence interval on 10,000 episodes sampled from validation set.

β	**UCF-101**		**HMDB-51**	
	1-shot	**5-shot**	**1-shot**	**5-shot**
1.0	83.46 ± 0.35	93.59 ± 0.23	62.01 ± 0.44	78.27 ± 0.39
0.8	88.95 ± 0.32	97.34 ± 0.17	70.26 ± 0.46	87.70 ± 0.35
0.6	91.73 ± 0.31	98.46 ± 0.14	74.11 ± 0.48	89.92 ± 0.34
0.4	92.72 ± 0.30	98.74 ± 0.13	75.08 ± 0.49	**90.23 ± 0.35**
0.2	**92.98 ± 0.31**	**98.75 ± 0.13**	**75.38 ± 0.50**	90.12 ± 0.35
0.0	92.91 ± 0.31	98.68 ± 0.14	75.02 ± 0.49	89.39 ± 0.36

GAN [10], ARN [19], and ITA [1], as well as a recent few-shot video classification approach, namely Baseline Plus [20]. ARN, ProtoGAN, and ITA use pretrained C3D network as their backbone, while FAN use a pretrained Dense-121 backbone network. Their results are taken from the original papers. The results of Baseline Plus are from our re-implementation. The datasets used for backbone network pretraining are presented in Tab. S3, except for FAN which does not mention that detail in their paper. Tab. S3 shows that CMOT performs the best, followed by ITA, which achieves the second best performance across all settings on both datasets. However, we note that CMOT and ITA use a C3D backbone, which is capable of extracting spatiotemporal information, and pretrained on a video dataset (i.e., Sports1M, Kinetics-400). In contrast, we use a 2D-based backbone (i.e., ResNet-50), and pretrain it on ImageNet dataset.

Transductive. Next, we consider the transductive setting. We re-implement three transductive techniques from few-shot image classification, i.e., Soft K-means [13], Bayes K-means [11], Mean-shift [11] as our competing methods.

Table S3. Comparison to the state-of-the-art methods in the **inductive** setting on the UCF-101 and HMDB-51 datasets. † denotes results from our re-implementation.

Method	Backbone	Pretrained Dataset	UCF-101		HMDB-51	
			1-shot	5-shot	1-shot	5-shot
ProtoGAN [10]	C3D [18]	Sports1M [6]	61.70 ± 1.60	79.70 ± 0.80	34.40 ± 1.30	50.90 ± 0.60
FAN [17]	Densenet-121 [5]	-	71.80 ± 0.10	86.50 ± 0.20	50.20 ± 0.20	67.60 ± 0.10
ARN [19]	C3D [18]	Sports1M [6]	62.10 ± 1.00	84.80 ± 0.80	44.60 ± 0.90	59.10 ± 0.80
Baseline Plus [20]†	Resnet-50 [4]	ImageNet [14]	81.06 ± 0.33	92.85 ± 0.22	57.56 ± 0.42	72.70 ± 0.37
ITA [1]	C3D [18]	Kinetics-400 [7]	88.71 ± 0.19	96.78 ± 0.08	63.43 ± 0.28	79.69 ± 0.20
CMOT [12]	C3D [18]	Sports1M [6]	**90.40 ± 0.40**	**95.70 ± 0.30**	**66.90 ± 0.50**	**81.50 ± 0.40**
Ours	Resnet-50 [4]	ImageNet [14]	84.93 ± 0.30	95.87 ± 0.17	59.57 ± 0.40	76.85 ± 0.36

Table S4. Comparison to the state-of-the-art methods in **transductive** setting on the UCF-101 and HMDB-51 datasets. Results of other methods are from our re-implementation on the trained feature extractor of [20].

Method	UCF-101		HMDB-51	
	1-shot	5-shot	1-shot	5-shot
Soft K-means [13]	90.26 ± 0.34	97.67 ± 0.17	64.44 ± 0.52	80.29 ± 0.45
Bayes K-means [11]	81.26 ± 0.33	93.13 ± 0.22	57.85 ± 0.42	73.44 ± 0.37
Mean-shift [11]	81.25 ± 0.33	89.35 ± 0.26	57.82 ± 0.42	67.96 ± 0.43
Ours	**94.18 ± 0.28**	**99.06 ± 0.11**	**68.07 ± 0.52**	**85.01 ± 0.42**

Results are shown in Tab. S4. To our best knowledge, we are the first to consider transductive inference in few-shot action recognition. As it is evident from the results, our method significantly outperforms the competing methods by large margins, i.e., $2-4\%$ on UCF-101 and $4-5\%$ on HMDB-51.

References

1. Cao, C., Li, Y., Lv, Q., Wang, P., Zhang, Y.: Few-shot action recognition with implicit temporal alignment and pair similarity optimization. CVIU (2021) 2, 5, 6
2. Cao, K., Ji, J., Cao, Z., Chang, C.Y., Niebles, J.C.: Few-shot video classification via temporal alignment. In: CVPR (2020) 1
3. Haresh, S., Kumar, S., Coskun, H., Syed, S.N., Konin, A., Zia, Z., Tran, Q.H.: Learning by aligning videos in time. In: CVPR (2021) 1
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 6
5. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017) 6
6. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014) 6
7. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 6
8. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011) 2
9. Kumar, S., Haresh, S., Ahmed, A., Konin, A., Zia, M.Z., Tran, Q.H.: Unsupervised activity segmentation by joint representation learning and online clustering. In: CVPR (2022) 1
10. Kumar Dwivedi, S., Gupta, V., Mitra, R., Ahmed, S., Jain, A.: Protogan: Towards few shot learning for action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 2, 5, 6
11. Lichtenstein, M., Sattigeri, P., Feris, R., Giryes, R., Karlinsky, L.: Tafssl: Task-adaptive feature sub-space learning for few-shot classification. In: ECCV (2020) 5, 6
12. Lu, S., Ye, H.J., Zhan, D.C.: Few-shot action recognition with compromised metric via optimal transport. arXiv preprint arXiv:2104.03737 (2021) 6
13. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. arXiv preprint arXiv:1803.00676 (2018) 5, 6
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV (2015) 6
15. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 2
16. Su, B., Hua, G.: Order-preserving wasserstein distance for sequence matching. In: CVPR (2017) 1
17. Tan, S., Yang, R.: Learning similarity: Feature-aligning network for few-shot action recognition. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2019) 2, 4, 6
18. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015) 6

19. Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P.H., Koniusz, P.: Few-shot action recognition with permutation-invariant attention. In: ECCV (2020) 2, 3, 5, 6
20. Zhu, Z., Wang, L., Guo, S., Wu, G.: A closer look at few-shot video classification: A new baseline and benchmark. arXiv preprint arXiv:2110.12358 (2021) 1, 5, 6