TAMT: Temporal-Aware Model Tuning for Cross-Domain Few-Shot Action Recognition

Supplementary Material

In the supplementary materials, we first explore the effect of hyper-parameters on the Hierarchical Temporal Tuning Network (HTTN), mainly including the number L and parameters γ & β for TAA blocks as well as the number of group G for ELSTC. Furthermore, to fully evaluate the generalization of TAMT, we set up a setting called generalization across datasets, and finally show its performance of varying shots and FSAR tasks. Lastly, we conduct visualization analyses to further validate the effectiveness of our TAMT.

1. Effect of Hyper-parameters on HTTN

Number L and Parameters γ & β in TAA block. In Tab. S1 and Tab. S2, we explore the optimal TAA transformer block number L and the parameter sharing strategy for parameters γ and β. Initially, L varies from 0 to 4, among which an L value of 0 implies a configuration without any adapters. Relative to this baseline ($L = 0$), the introduction of adapters yields a positive impact, enhancing performance by over 1.96% with only a minimal increase in training cost. Optimal performance is observed when L is set to 2 or 3. For higher efficiency and considering the performance-consumption balance, $L = 2$ is chosen as the default configuration. For the parameters γ and β in TAA blocks, they are partially shared. Only the $\mathbf{W}_{\downarrow}^{\gamma}$ and $\mathbf{W}_{\downarrow}^{\beta}$ in Eqn. (2) & Eqn. (3) are shared, resulting in an average gain of 0.48% across five datasets while reducing learnable parameters by 10%, as shown in Tab. S2.

Number of Group G for ELSTC. To evaluate the effect of different group numbers G in ELSTC, we consider values ranging from 1 to 8 for a sequential feature of length $T = 8$. We observe both performance and computational overhead (including feature dimension, number of parameters, GFLOPs, and inference latency), as shown in Tab. S3. The results show that grouping features ($G > 1$) effectively reduces the computational overhead compared to the original setting ($G = 1$). Moreover, increasing G further alleviates the overhead. Notably, optimal performance is achieved at $G = 4$, This improvement likely results from a balance between more effective optimization (compared with $G = 1, 2$) and better preservation of temporal interactions within each group (in contrast to $G = 8$).

2. Generalization Verification

To further validate the generalization of TAMT, we first conduct experiments under a setting of generalization across datasets. Furthermore, we validate the effect of our TAMT on more shot experiments, and finally demonstrate the generalization ability of our proposed HTTN in FSAR tasks.

L	SSV2	Diving	UCF	Average	Memory
0	53.41	42.87	94.97	63.58	1.2G
1	57.48	43.52	95.61	65.54	+0.0G
2	59.18	45.18	95.92	66.76	+0.7G
3	59.86	44.75	95.70	66.77	+1.8G
4	59.67	44.22	94.87	66.25	+2.9G

Table S1. Effect of the hyper-parameter L on HTTN, and the accuracy (%) of 5-way 5-shot is reported. Memory: GPU memory for training.

	P	HMDB	SSV2	Diving	UCF	RareAct	Average
S	2.8M	74.14	59.18	45.18	95.92	67.44	68.37
U	3.1M	73.94	58.34	44.12	95.58	67.47	67.89

Table S2. Comparison (%) of the shared parameters $\gamma\&\beta$ in $\mathbf{W}_{\downarrow}^{\gamma\&\beta}$. S: $\gamma\&\beta$ are Shared, U: $\gamma\&\beta$ are not Shared. P: Parameters.

Generalization Across Datasets. Here, we compare with the counterpart CDFSL-V on a challenging setting, where we pre-train the models on the K-400 dataset and fine-tune the models on on UCF or HMDB. Then, the fine-tuned models are directly adopted to four downstream datasets without any tuning. As shown in Tab. S4, our TAMT outperforms CDFSL-V by an average of 13.83% and 15.26% on four test datasets [2, 6, 8, 15], respectively. These results clearly demonstrate that our method can be well generalized across different datasets.

Results of Different Training Shots. To further assess the generalization of our TAMT method, we compare our TAMT on various 5-way K-shot ($K = 1, 5, 20$) settings, by using ViT-S with 112×112 input resolution. The performance of transferring from source dataset K-100 [20] to five target datasets [2, 6, 8, 10, 15] is presented in Tab. S5, in which the average Top-2 best performances are marked by **red** and **blue**, respectively. As shown in Tab. S5, our TAMT exhibits outstanding performance compared to the prime counterpart CDFSL-V [14] under 1-shot, 5-shot and 20-shot settings with a significant margin on average accuracy over 24.08%, 31.15% and 34.13%. Particularly, for the 5-way 1-shot setting, our TAMT is the only approach to achieve a significative performance (namely, above 20% for 5-way recognition) on HMDB, SSV2, Diving and UCF

Feature Splitting	G	Dim.	Params.	GFLOPs	Latency	SSv2	Diving	UCF	Avg.
×	1	262K	101M	10.5G	10.6ms	59.06	43.76	95.32	66.05
✓	2	65K	25M	4.2G	5.5ms	58.85	43.95	95.15	65.98
	4	4K	1.6M	2.2G	3.7ms	59.18	45.18	95.92	66.76
	8	1K	1.0M	2.1G	2.8ms	58.09	44.45	95.24	65.92

Table S3. Effect of the hyper-parameter G on HTTN, where 5-way 5-shot accuracy (%) and computation overhead are reported. Dim.: Dimension of $\mathbf{M}_2$. Params.: Training parameters. GFLOPs: GFLOPs of ELSTC. Latency: Inference latency of ELSTC.

Method	Pre-trained Dataset → Tuned Dataset →	Test Dataset				
		HMDB	SSV2	Diving	UCF	Average
CDFSL-V [14]	K400 → HMDB →	-	21.39	21.21	51.66	31.42
TAMT (Ours)		-	43.22	29.04	63.49	45.25$_{(+13.83)}$
CDFSL-V [14]	K400 → UCF →	51.97	24.36	22.62	-	32.98
TAMT (Ours)		72.50	43.02	29.21	-	48.24$_{(+15.26)}$

Table S4. Comparison (%) with CDFSL-V [14] on across datasets setting. All results are conducted on ViT-S network with 112 × 112 resolution, reported 5-way 5-shot accuracy on test dataset.

Method	K-shot	Target					
		HMDB	SSV2	Diving	UCF	RareAct	Average
STARTUP++ [12]	1-shot	16.66	14.17	13.13	24.48	17.21	17.13
DD++ [5]		17.44	14.96	13.73	26.04	19.02	18.24
CDFSL-V [14]		**18.59**	**16.01**	**14.11**	**27.78**	**20.06**	**19.31**
TAMT (Ours)		**47.02**	**34.45**	**27.04**	**72.38**	**36.04**	**43.39$_{(+24.08)}$**
STARTUP++ [12]	5-shot	24.97	15.16	14.55	32.20	31.77	23.73
DD++ [5]		25.99	16.00	16.24	34.10	31.20	24.71
SEEN*† [17]		52.80	31.20	40.90	79.60	50.20	50.94
CDFSL-V [14]		29.80	17.21	16.37	36.53	33.91	26.76
DMSD*† [4]		**54.90**	**32.10**	**42.28**	**81.90**	**53.30**	**52.90**
TAMT (Ours)		**61.76**	**48.90**	**38.33**	**87.76**	**52.81**	**57.91$_{(+31.15)}$**
STARTUP++ [12]	20-shot	30.48	17.15	17.30	34.02	38.45	27.48
DD++ [5]		33.09	17.56	17.33	36.72	39.97	28.93
CDFSL-V [14]		**36.89**	**18.72**	**17.81**	**39.92**	**42.51**	**31.17**
TAMT (Ours)		**73.71**	**55.45**	**42.68**	**91.38**	**63.27**	**65.30$_{(+34.13)}$**

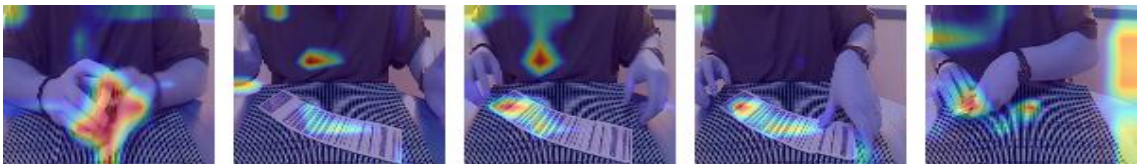
Table S5. Comparison (%) of state-of-the-arts on various 5-way K-shot settings ($K = 1, 5, 20$) of CDFSAR with employing K-100 as source dataset. All results are conducted with 112 × 112 resolution by using ViT-S backbone, except Method marked by * (224 × 224 resolution by using ResNet-18).

datasets. In addition, the performance of TAMT is boosted by 14.52% and 21.91%, when extending 1-shot to 5-shot and 20-shot settings, which is more remarkable than the 7.45% and 11.86% increase observed in CDFSL-V. All the above results reveal that our TAMT has a good ability to explore information lying in the annotated support set, effectively handling the challenging 1-shot setting and benefiting from the increase in support samples.

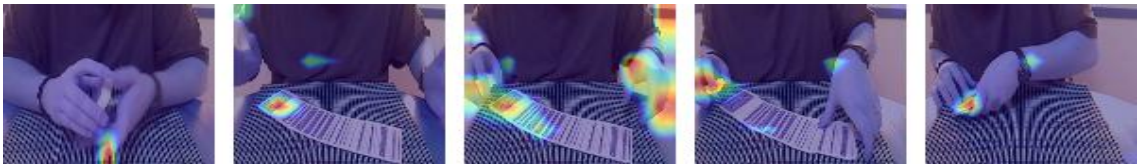
Generalization on FSAR Task. Our TAMT approach is also evaluated on the conventional FSAR problem, where we compare it alongside very recent Method based on large-scale models, such as CLIP-ViT-B and BLIP-ViT-B, as detailed in Tab. S6. Considering CLIP network is conducted pre-training using 400M data, our TAMT employs the Kinetics-710 [7] database for the pre-training phase with about 660K trainable instances. The results demonstrate that TAMT exhibits impressive performance superiority in dual modality settings. Specifically, TAMT outperforms comparable unimodal competitors by clear margin, which achieves about 5.0% and 3.0% on average across multiple datasets. Moreover, TAMT shows performance gain of 0.8% over huge-pretrained models within the CLIP family in terms of

Method	M.	Pre-training	Tuning	HMDB	SSV2	UCF	Average
CLIP* [13]	Multi-modal	CLIP-ViT-B	Frozen	58.2/77.0	30.0/42.4	89.7/95.7	59.3/71.7
CapFSAR [18]		BLIP-ViT-B	FFT	70.3/81.3	54.0/70.1	93.1/97.7	72.5/83.0
CLIP-CPM^{2}C [3]		CLIP-ViT-B	FFT	**75.9**/**88.0**	60.1/**72.8**	95.0/98.6	77.0/**86.5**
CLIP-FSAR [19]		CLIP-ViT-B	FFT	75.8/87.7	**61.9**/72.1	**96.6**/**99.0**	**78.1**/86.3
OTAM* [1]	Unimodal	CLIP-ViT-B(V)	FFT	72.5/83.9	50.2/68.6	95.8/98.8	72.8/83.8
TRX* [11]		BLIP-ViT-B(V)	FFT	58.9/79.9	45.1/68.5	90.9/97.4	65.0/81.9
HyRSM* [16]		BLIP-ViT-B(V)	FFT	69.8/80.6	52.1/69.5	91.6/96.9	71.2/82.3
MASTAF [9]		JFT-ViT-B	FFT	69.5/*N/A*	60.7/*N/A*	91.6/*N/A*	73.9/*N/A*
TAMT (Ours)		ViT-B	PEFT	**77.7**/**88.2**	**61.4**/**73.3**	**97.5**/**98.8**	**78.9**/**86.8**

Table S6. Comparison (%) of state-of-the-arts on FSAR setting in terms of 5-way 1-shot/5-shot accuracy. M.: Modality, (V): Only visual encoder of CLIP. *: from [18, 19].

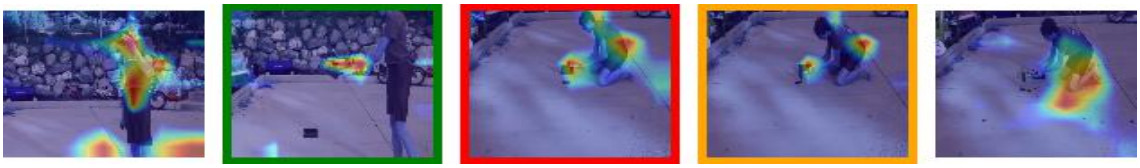


(1a) K-400 dataset, *"Playing cards"*, CDFSL-V [14]



(1b) K-400 dataset, *"Playing cards"*, TAMT (Ours)

(2a) K-400 → RareAct, *"Hammer phone"*, CDFSL-V [14]



(2b) K-400 → RareAct, *"Hammer phone"*, TAMT (Ours)

Figure S1. Feature visualization on setting of K-400 → RareAct.

1-shot accuracy, despite the absence of auxiliary text modality. Notably, by using the PEFT training protocol, TAMT theoretically benefits from a lower training complexity than these full fine-tuning (FFT) approaches. These results show that TAMT generalizes well to the FSAR setting, providing an efficient and effective alternative.

3. Visualization Analyses

To further validate the effectiveness of our TAMT method for addressing the problem of domain gap, we visualize feature heatmaps (the last layer of the backbone) of different models pre-trained on the source dataset (K-400) and those after tuning on the target dataset (RareAct) in Fig. S1. It can be observed that, on the source dataset, both CDFSL-V [14] and our TAMT focus on discriminative regions. After tuning on the target dataset, TAMT captures more semantic features for better recognition (e.g., the human body and phone in class "*Hammer phone*"), indicating superiority to address the problem of domain gap on downstream tasks.

References

[1] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10615–10624, 2020. 3

[2] 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, 2017. 1

[3] Fei Guo, YiKang Wang, Han Qi, Li Zhu, and Jing Sun. Consistency prototype module and motion compensation for few-shot action recognition (CLIP-CPM^{2}C). *Neurocomputing*, 611:128649, 2025. 3

[4] Fei Guo, Yi Kang Wang, Han Qi, Li Zhu, and Jing Sun. DMSD-CDFSAR: Distillation from mixed-source domain for cross-domain few-shot action recognition. *Expert Systems With Applications*, 270, 2025. 2

[5] Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, and Richard J Radke. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:3584–3595, 2021. 2

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *International Conference on Computer Vision (ICCV)*, pages 2556–2563, 2011. 1

[7] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. UniFormerV2: Unlocking the potential of image vits for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1632–1643, 2023. 2

[8] Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 1

[9] Xin Liu, Huanle Zhang, Hamed Pirsiavash, and Xin Liu. MASTAF: A model-agnostic spatio-temporal attention fusion network for few-shot video classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2508–2517, 2023. 3

[10] Antoine Miech, Jean Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. RareAct: A video dataset of unusual interactions. *arXiv preprint arXiv:2008.01018*, 2020. 1

[11] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational cross Transformers for few-shot action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 475–484, 2021. 3

[12] Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 3

[14] Sarinda Samarasinghe, Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. CDFSL-V: Cross-domain few-shot learning for videos. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 11643–11652, 2023. 1, 2, 3

[15] 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

[16] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19916–19925, 2022. 3

[17] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yiliang Lv, Changxin Gao, and Nong Sang. Cross-domain few-shot action recognition with unlabeled videos. *Computer Vision and Image Understanding (CVIU)*, 233:103737, 2023. 2

[18] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Few-shot action recognition with captioning foundation models. *arXiv preprint arXiv:2310.10125*, 2023. 3

[19] Xiang Wang, Shiwei Zhang, Jun Cen, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. CLIP-guided prototype modulating for few-shot action recognition. *International Journal of Computer Vision*, 132(6):1899–1912, 2024. 3

[20] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 751–766, 2018. 1