Supplementary Material

A. Implementation Details

The training data is randomly cropped to 224 × 224 and we perform random flipping except for Something-Something datasets. At inference stage, all frames will be center-cropped to 224 × 224 except SlowFast [1] which adopts the resolution of 256 × 256 for evaluation. We use one-clip one-crop per video during evaluation except Uniformer [4] which utilizes one-clip three-crop evaluation protocol. We train all models on NVIDIA Tesla V100 GPUs and adopt the same training hyperparameters with the official implementations.

B. Results of Different Depths

Table 1. Experiments with different depths on Something-Something V1. The best results are bold-faced.

Method	Top-1 Acc.(%)		
	v^L	v^M	v^H
TSM(R18) [6]	16.82	33.12	42.95
TSM(R18)-ST	32.33	38.21	42.95
TSM(R18)-FFN	**36.83**(4.50↑)	**41.61**(3.40↑)	**43.57**(0.62↑)
TSM(R101) [6]	22.15	39.30	49.57
TSM(R101)-ST	40.76	46.96	49.57
TSM(R101)-FFN	**45.15**(4.39↑)	**50.24**(3.28↑)	**51.79**(2.22↑)

As we have shown in the main text, Temporal Frequency Deviation phenomenon exists in different depths of the network which means it has no relation to the representation ability. But whether FFN can address this issue at other depths remains a problem. As previous experiments are built on ResNet-50 [3], we conduct experiments on ResNet-18, ResNet-101 and include their results in Tab. 1. The results show that FFN outperforms Separated Training (ST) at different frame numbers which proves that FFN can effectively resolve Temporal Frequency Deviation problem regardless of the depths of the deep network.

C. Results of Different Middle Sequences

Another design choice in our method is the selection of middle sequence v^M, as v^L and v^H are usually set at first based on the range of the computations. Thus, we sample 8/10/12 frames for v^M respectively and evaluate them at various frame numbers in Tab. 2. When we sample 8 frames for v^M, FFN obtains the best performance at 8 Frame compared to the other two choices and the phenomenon is the same when sampling 10 or 12 frames for v^M. This meets our expectation as the specialized normalization for v^M learns its corresponding transformation. Overall, all three choices lead to consistent improvement over Separated Training (ST) at all frames.

D. Any Frame Inference of Input Sequences Combinations

In the main text, we have conducted the ablation of input sequences combinations. We further validate the three models at more fine-grained frame numbers with the proposed inference paradigm and the results are shown in Tab. 3. One can observe that FFN(2) obtains lower accuracy compared to ST at 6/8/10 Frame because of the missing middle sequence. While FFN(4) achieves the highest performance at 8/10/12 Frame as the introduced sequence at Frame 12 will alleviate the Temporal Frequency Deviation nearby.

E. Further Verification of Nearby Alleviation

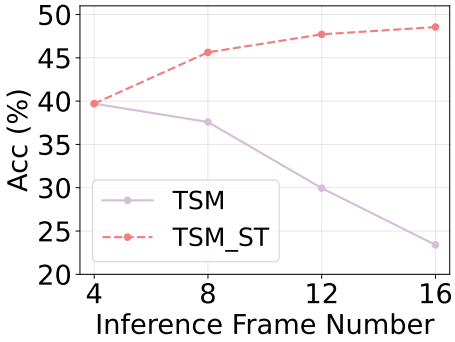
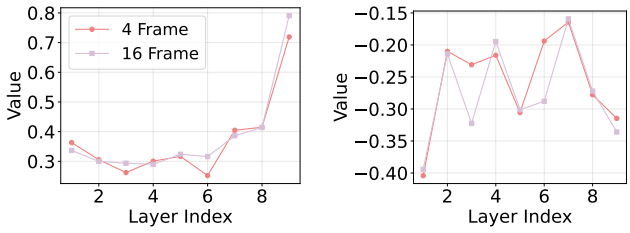


Figure 1. Validation results of TSM which is trained at 4 Frame on Something-Something V1 dataset.

In previous parts, we have conducted experiments which train the model at Frame 8/12/16 and evaluate their performance at different frames. Here we further train the model at 4 Frame and show the validation results in Fig. 1. Similarly, we can observe that frames close to 4 exhibit the slightest performance drop as their normalization statistics is more similar with frame 4 which further verifies the Nearby Alleviation phenomenon.

F. Statistics of Normalization Shifting



(a) Scale: γ (b) Bias: β

Figure 2. Batch Normalization statistics at various layers. TSM models are trained at 4 Frame and 16 Frame separately, and the statistics are calculated from the fourth stage of ResNet-50.

We have shown the calculated normalization statistics, Mean: μ and Variance: σ^2 in previous sections. In this part, we further include the calculated statistics of Scale: γ and Bias: β in Fig. 2. One can observe that the two curves are

Table 2. Experiments with different middle sequences on Something-Something V1. The best results are bold-faced.

Method	v^M	Top-1 Acc.(%)						
		4 Frame	6 Frame	8 Frame	10 Frame	12 Frame	14 Frame	16 Frame
TSM [6]	-	20.60	30.23	37.36	42.72	45.97	47.49	48.55
TSM-ST	-	39.71	43.73	45.63	47.31	47.71	48.01	48.55
TSM-FFN	8F	42.85	**46.57**	**48.20**	48.81	48.90	50.47	50.79
TSM-FFN	10F	**43.10**	44.77	47.81	**49.26**	49.63	**50.67**	**51.12**
TSM-FFN	12F	42.92	43.57	46.82	48.85	**49.73**	50.40	50.79

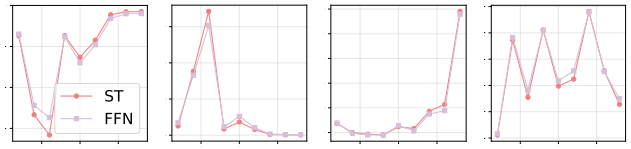
Table 3. Any frame inference results of input sequences combinations on Something-Something V1. The best results are bold-faced.

Method	Sequences	Top-1 Acc.(%)						
		4 Frame	6 Frame	8 Frame	10 Frame	12 Frame	14 Frame	16 Frame
TSM [6]	-	20.60	30.23	37.36	42.72	45.97	47.49	48.55
TSM-ST	-	39.71	43.73	45.63	47.31	47.71	48.01	48.55
TSM-FFN(2)	4/16	41.69	42.07	37.93	46.11	48.10	49.37	49.79
TSM-FFN(3)	4/8/16	42.85	**46.57**	48.20	48.81	48.90	**50.47**	**50.79**
TSM-FFN(4)	4/8/12/16	**43.40**	46.51	**48.66**	**48.92**	**49.77**	50.11	50.63

not aligned with each other which further demonstrates that the discrepancy of BN statistics is an important reason for Temporal Frequency Deviation phenomenon and specializing normalization operations in deep networks is an intuitive way to resolve normalization shifting.

G. Validation of Normalization Shifting

To further prove that our method can mitigate the normalization shifting problem, we compare the BN statistics of ST (16F) and FFN (16F) which is trained with TSM [6] on Something-Something V1 [2] dataset. As is shown in Fig. 3, one can observe that the two curves are well-aligned with each other which demonstrates that the calculated statistics are very similar and the normalization shifting problem can be alleviated by FFN.



(a) Mean: μ (b) Variance: σ^2 (c) Scale: γ (d) Bias: β

Figure 3. Batch Normalization statistics at various layers. TSM-ST is trained at 16 Frame and both models are evaluated at 16 Frame as well. The statistics are calculated from the fourth stage of ResNet-50.

H. Quantitative Results

In the Experiments section, we show performance analysis of FFN across architectures and datasets in the figure and we also provide the corresponding quantitative results in Tab. 4 and Tab. 5 for reference.

Table 4. Quantitative results of different architectures experiments on Something-Something V1. The best results are bold-faced.

Method	Top-1 Acc.(%)		
	v_L	v_M	v_H
TSM [6]	20.60	37.36	48.55
TSM-ST	39.71	45.63	48.55
TSM-FFN	**42.85**(3.14↑)	**48.20**(2.57↑)	**50.79**(2.24↑)
TEA [5]	21.78	41.49	51.23
TEA-ST	41.36	48.37	51.23
TEA-FFN	**44.97**(3.61↑)	**51.61**(3.24↑)	**54.04**(2.81↑)
SlowFast [1]	15.08	35.08	45.88
SlowFast-ST	39.91	44.12	45.88
SlowFast-FFN	**43.90**(3.99↑)	**47.11**(2.99↑)	**47.27**(1.39↑)
Uniformer [4]	22.38	47.98	56.71
Uniformer-ST	44.33	51.49	56.71
Uniformer-FFN	**51.41**(7.08↑)	**56.64**(5.15↑)	**58.88**(2.17↑)

Table 5. Quantitative results of different datasets experiments on TSM. The best results are bold-faced.

Method	Dataset	Top-1 Acc.(%)		
		v_L	v_M	v_H
TSM [6]	Sth-Sth V2	31.52	51.55	61.02
TSM-ST		53.38	59.29	61.02
TSM-FFN		**56.07**(2.69↑)	**61.86**(2.57↑)	**63.61**(2.59↑)
TSM [6]	Kinetics400	64.10	69.77	73.16
TSM-ST		66.25	70.38	73.16
TSM-FFN		**68.96**(2.71↑)	**72.33**(1.95↑)	**74.35**(1.19↑)
TSM [6]	HMDB51	42.16	46.38	48.30
TSM-ST		44.74	46.77	48.30
TSM-FFN		**45.67**(0.93↑)	**47.67**(0.90↑)	**48.80**(0.50↑)

References

[1] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 1, 2

[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 *ICCV*, 2017. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022. 1, 2

[5] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *CVPR*, 2020. 2

[6] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 1, 2