Motion meets Attention: Video Motion Prompts

Qixiang Chen[1], Lei Wang[1,2], Piotr Koniusz[2,1], Tom Gedeon[3]

[1]Australian National University [2]Data61/CSIRO [3]Curtin University

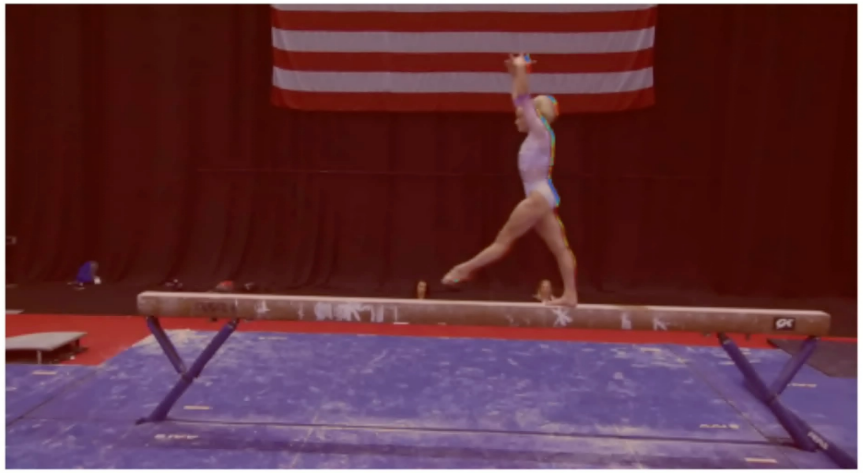
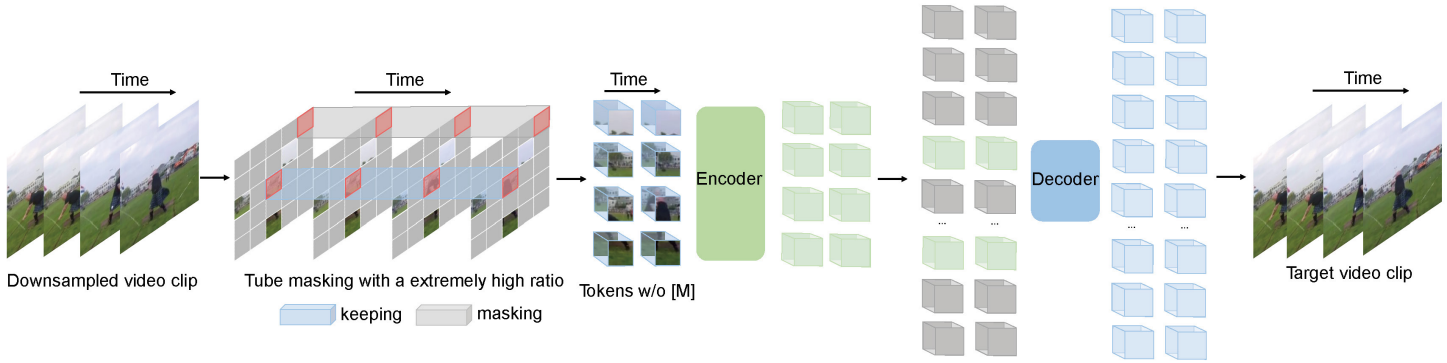
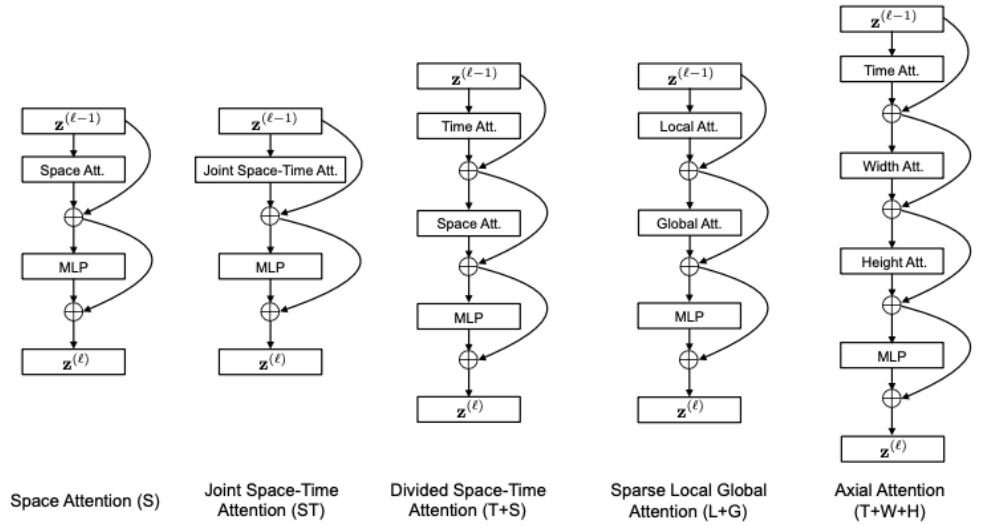


Table of content

Background & Motivation



VideoMAE[1]



TimeSformer[2]

[1]: https://arxiv.org/abs/2203.12602
[2]: https://arxiv.org/pdf/2102.05095

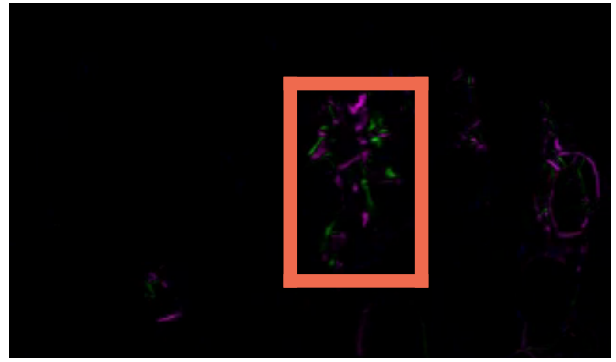
Background & Motivation

UCF-Crime: Fighting

Original video

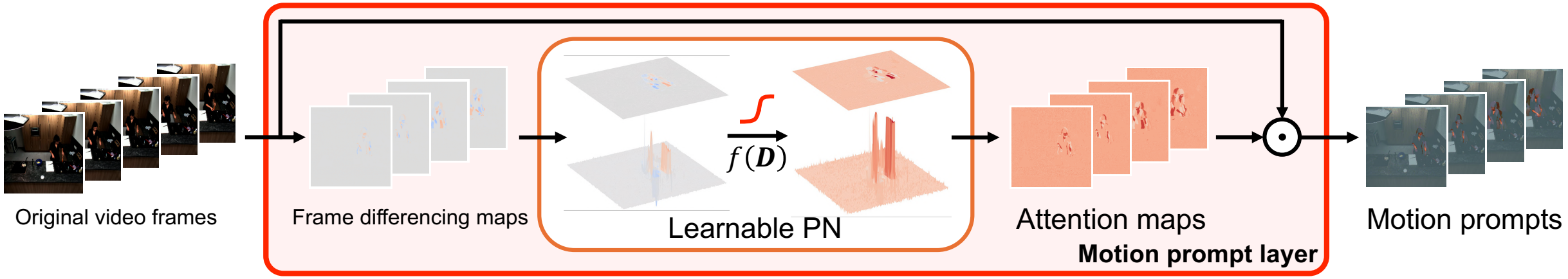
Normalized frame differencing map

Time-color reordering frame



Taylor video frame

Method

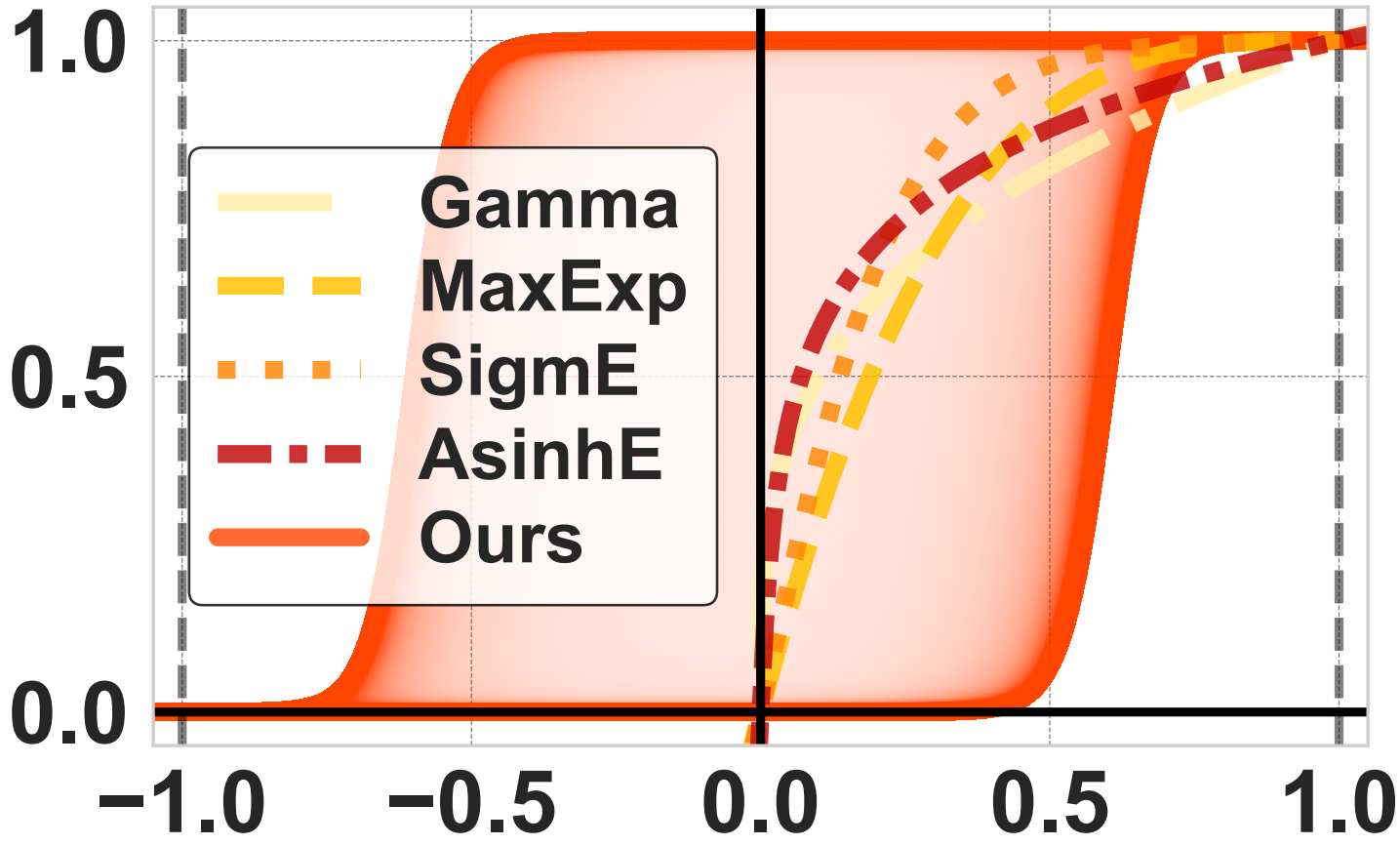

Original video frames
Frame differencing maps
$f(\boldsymbol{D})$
Learnable PN
Attention maps
Motion prompt layer
Motion prompts

Method

$$f(\boldsymbol{D}) = \frac{1}{1 + e^{-a(\boldsymbol{D}-b)}}$$

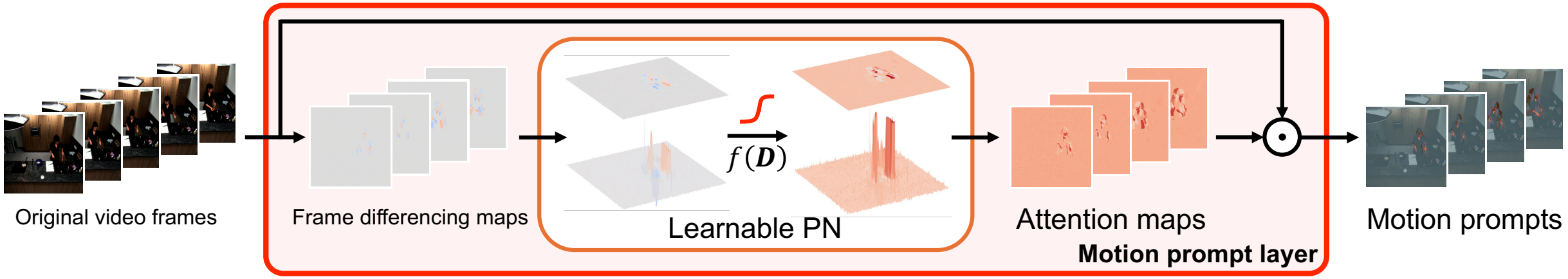
$$\begin{cases} a(m) = \dfrac{\alpha}{\beta\,|\tanh(m)| + \epsilon} \\ b(n) = \gamma \tanh(n) \end{cases}$$

Method



*Comparison of existing well-behaved Power Normalization (PN) functions Ko- niusz and Zhang (2021) and our learnable PN function

Method


Original video frames
Frame differencing maps
$f(\boldsymbol{D})$
Learnable PN
Attention maps
Motion prompt layer
Motion prompts

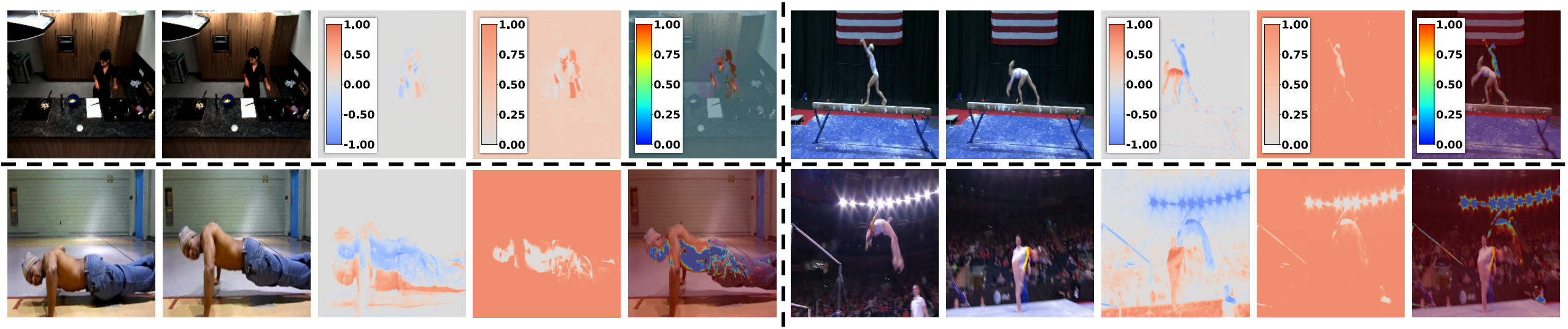
Method

Loss function: $$\mathcal{L} = \mathcal{L}_{\text{ori}} + \lambda \mathcal{V},$$

Temporal attention variation regularization:

$$\mathcal{V} = \frac{1}{T-2} \sum_{t=1}^{T-2} ||f(\boldsymbol{D}_{t+1}) - f(\boldsymbol{D}_t)||_F^2,$$

Experiments & Discussion

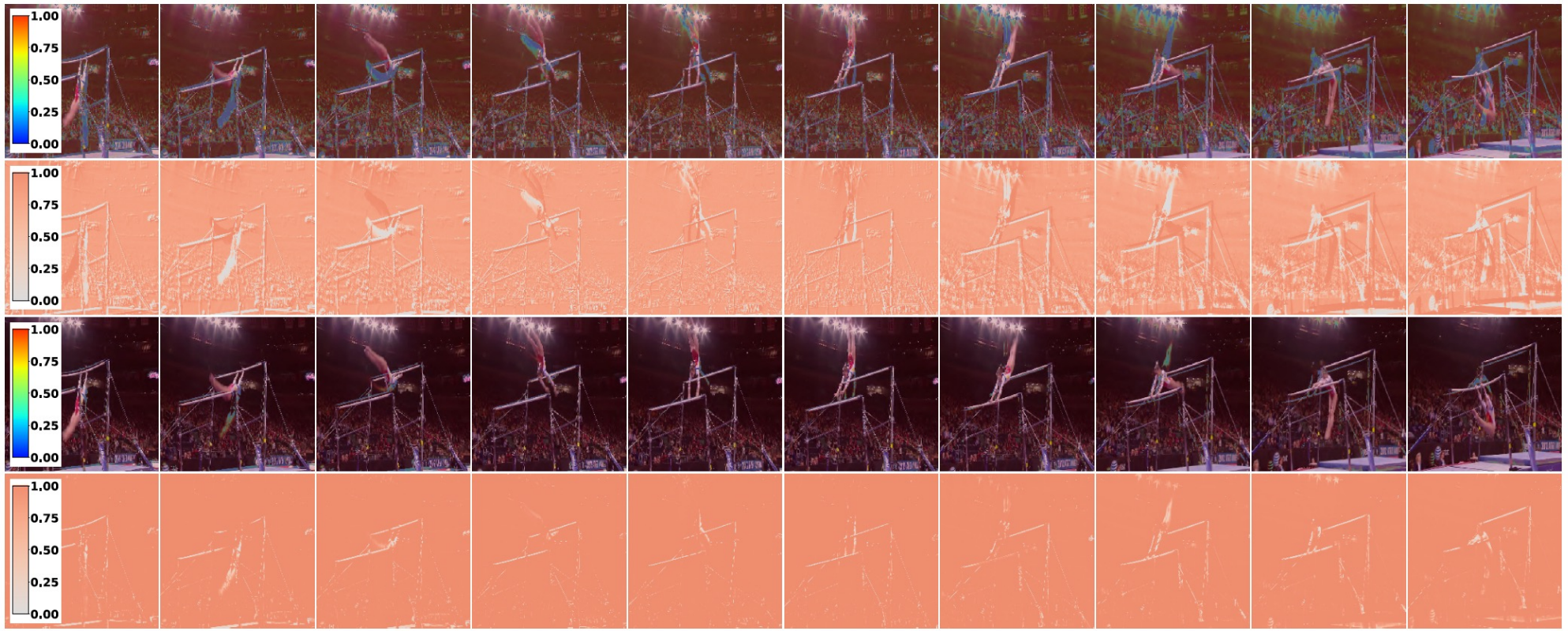


Experiments & Discussion

	[Pretrained❄]	[Pretrained❄ +**VMPs**✹]	Baseline ([Pretrained❄✹])	[Pretrained❄✹ +**VMPs**✹]	[Pretrained❄✹❄ +**VMPs**✹]	[Pretrained❄✹❄✹ +**VMPs**✹]
Top-1	36.6	**37.1**$^{\uparrow 0.5}$	50.6	**56.6**$^{\uparrow 6.0}$	**56.2**$^{\uparrow 5.6}$	**57.1**$^{\uparrow 6.5}$
Top-5	**66.9**	66.2 $^{\downarrow 0.7}$	81.8	**84.4**$^{\uparrow 2.6}$	**84.3**$^{\uparrow 2.5}$	**83.7**$^{\uparrow 1.9}$

Table1: Variant study of finetuning on MPII Cooking 2 using TimeSformer.

Experiments & Discussion



$\lambda = 0$

$\lambda = 2.5$

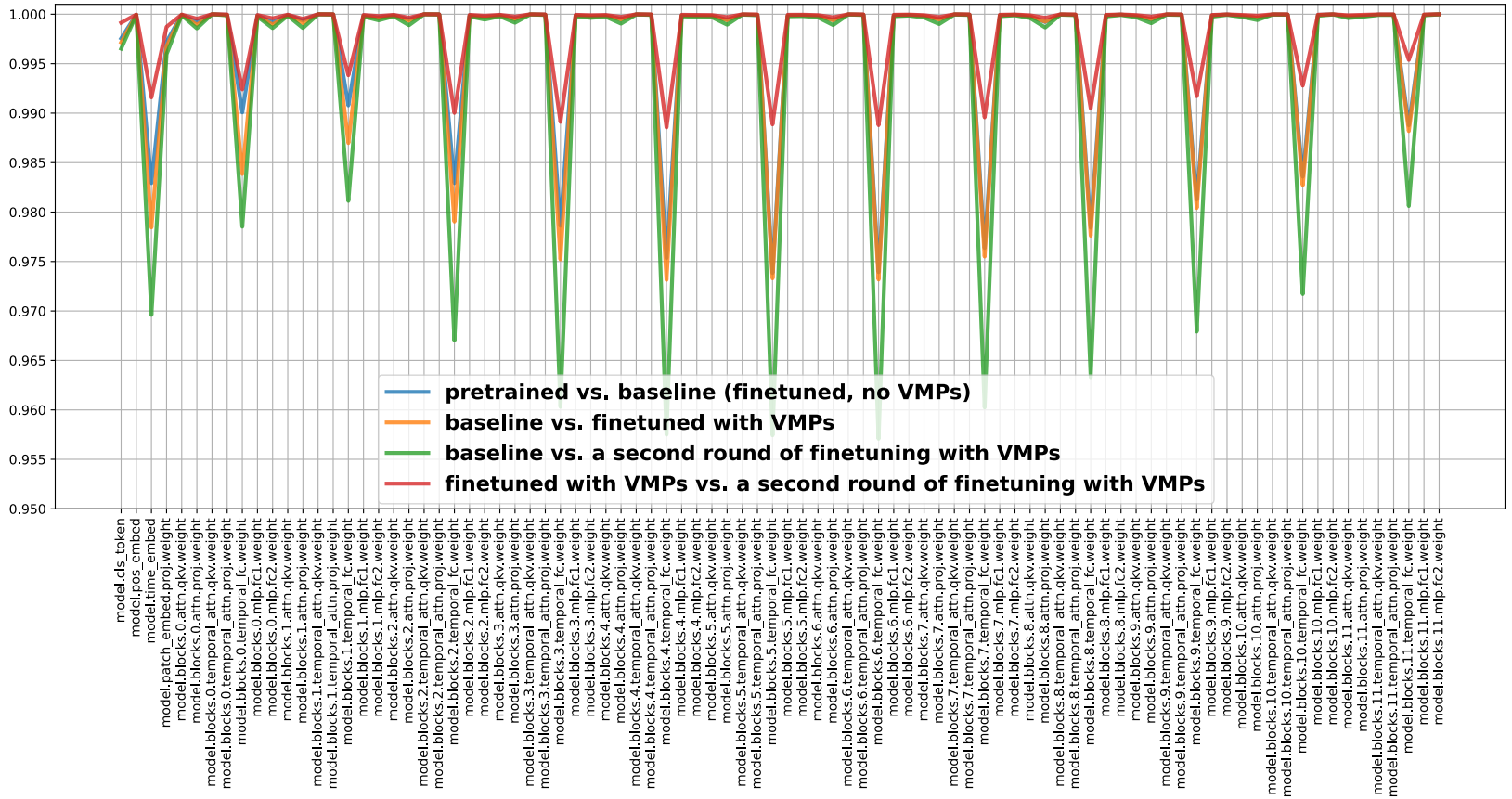
Experiments & Discussion

Table 2: Evaluations are conducted on (*left*) HMDB-51, and (*right*) FineGym, MPII Cooking 2, using SlowFast, X3D and TimeSformer as backbones. For SlowFast, we explore three variants by adding motion prompts into the slow-only stream, fast-only stream, and both slow and fast streams. We highlight improvements in red.

Model	HMDB-51			Mean
	Split 1	Split 2	Split 3	
SlowFast	75.4	76.2	76.9	76.2
+**VMPs** (slow-only)	**76.8**$^{\uparrow 1.4}$	**77.0**$^{\uparrow 0.8}$	**77.3**$^{\uparrow 0.4}$	**77.0**$^{\uparrow 0.8}$
+**VMPs** (fast-only)	**76.5**$^{\uparrow 1.1}$	**77.4**$^{\uparrow 1.2}$	**77.1**$^{\uparrow 0.2}$	**77.0**$^{\uparrow 0.8}$
+**VMPs** (slow&fast)	**76.2**$^{\uparrow 0.8}$	**76.7**$^{\uparrow 0.5}$	**77.1**$^{\uparrow 0.2}$	**76.6**$^{\uparrow 0.4}$
X3D	75.0	72.6	73.4	73.7
+**VMPs**	**75.8**$^{\uparrow 0.8}$	**73.2**$^{\uparrow 0.6}$	**73.6**$^{\uparrow 0.2}$	**74.2**$^{\uparrow 0.5}$
TimeSformer	72.7	73.1	72.2	72.7
+**VMPs**	**74.2**$^{\uparrow 1.5}$	**74.3**$^{\uparrow 1.2}$	**72.9**$^{\uparrow 0.7}$	**73.8**$^{\uparrow 1.1}$

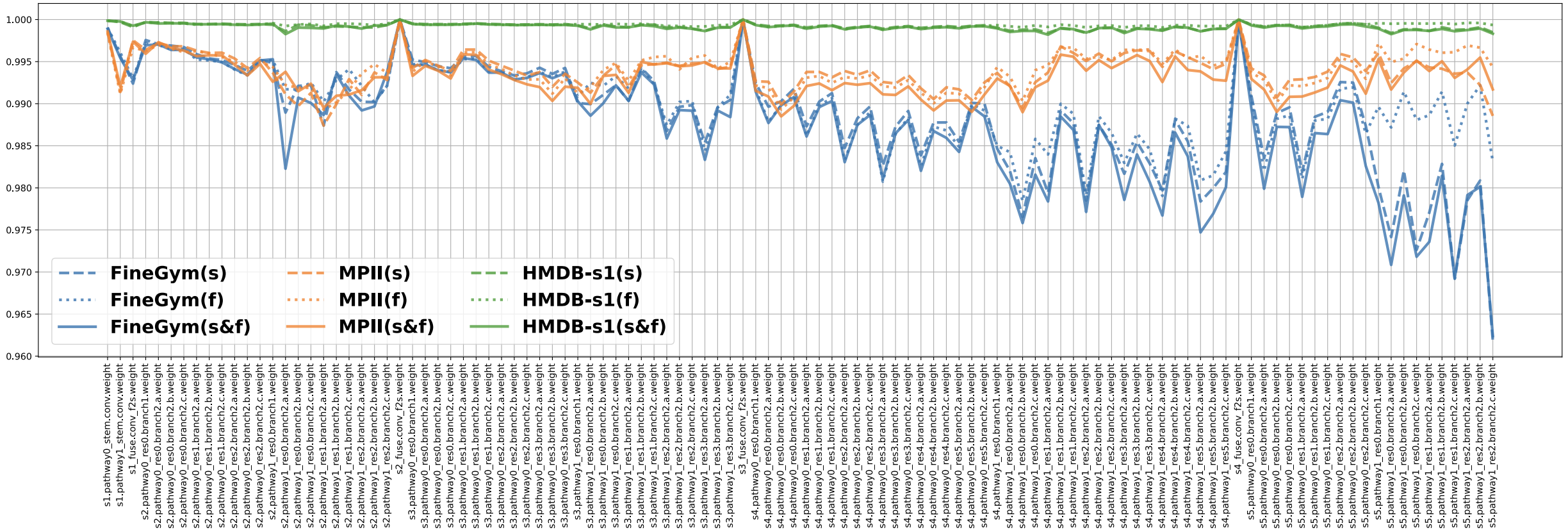
Model	FineGym		MPII Cooking 2	
	Top-1	Top-5	Top-1	Top-5
SlowFast	**89.8**	99.2	52.9	**86.1**
+**VMPs** (slow-only)	89.7$^{\downarrow 0.1}$	99.2	**55.5**$^{\uparrow 2.6}$	84.5$^{\downarrow 1.6}$
+**VMPs** (fast-only)	**90.3**$^{\uparrow 0.5}$	**99.3**$^{\uparrow 0.1}$	**55.2**$^{\uparrow 2.3}$	84.0$^{\downarrow 2.1}$
+**VMPs** (slow&fast)	**90.1**$^{\uparrow 0.3}$	**99.3**$^{\uparrow 0.1}$	**56.8**$^{\uparrow 3.9}$	**86.6**$^{\uparrow 0.5}$
X3D	83.0	98.4	48.4	**80.8**
+**VMPs**	**83.8**$^{\uparrow 0.8}$	**98.6**$^{\uparrow 0.2}$	**49.1**$^{\uparrow 0.7}$	80.6$^{\downarrow 0.2}$
TimeSformer	83.6	**98.7**	50.6	81.5
+**VMPs**	**84.4**$^{\uparrow 0.8}$	98.5$^{\downarrow 0.2}$	**56.6**$^{\uparrow 6.0}$	**84.4**$^{\uparrow 2.9}$

Experiments & Discussion



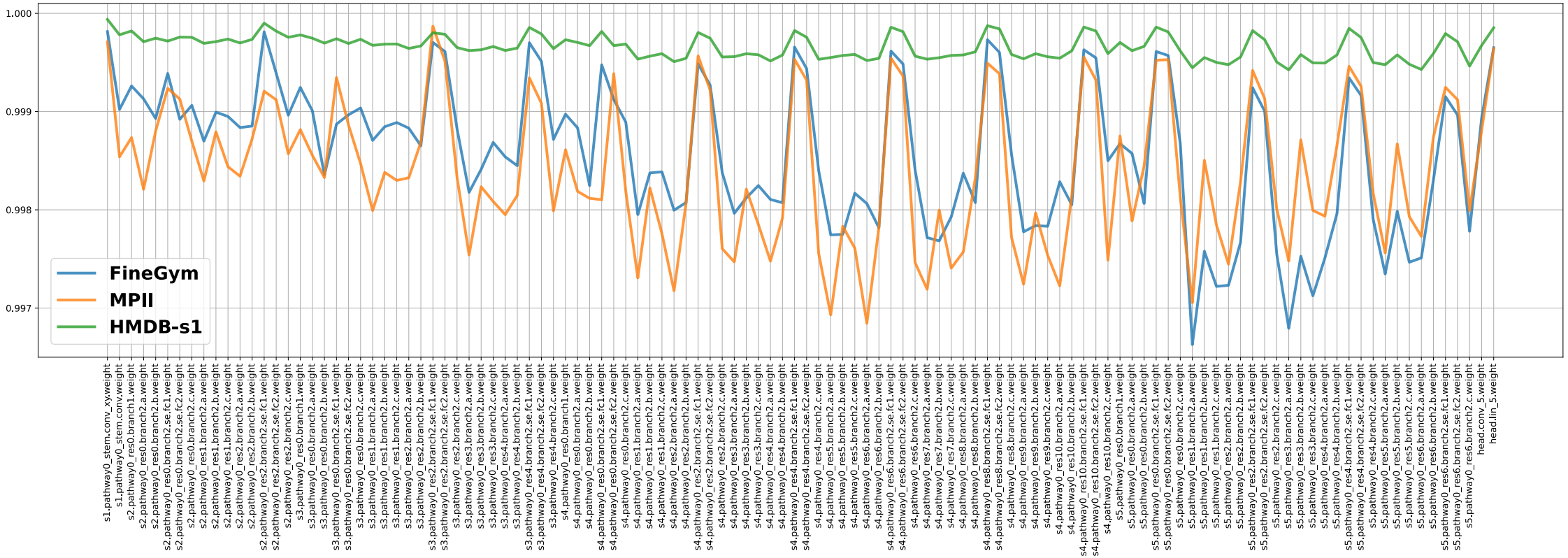
*Roles of VMPs in model finetuning via per-layer weight similarity comparison. We use TimeSformer pretrained on Kinetics-600 as the backbone, and finetuned on MPII Cooking 2 with or without VMPs.

Experiments & Discussion



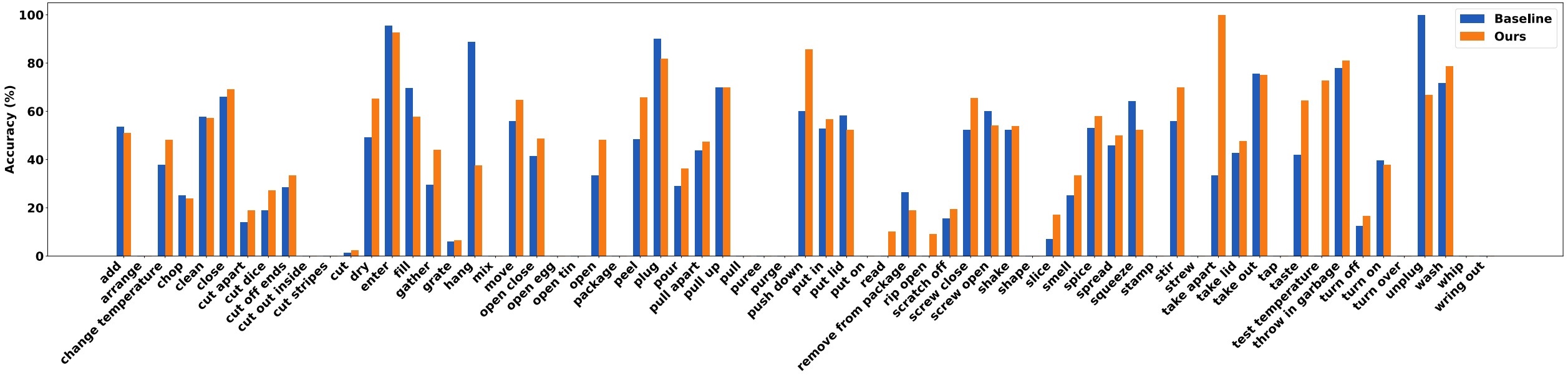
*Roles of VMPs in model finetuning via per-layer weight similarity comparison. We use SlowFast pretrained on Kinetics-600 as the backbone and finetune it on FineGym, MPII Cooking 2 (MPII), and HMDB-51 split 1 (HMDB-s1).

Experiments & Discussion



*Roles of VMPs in model finetuning via per-layer weight similarity comparison. We use X3D pretrained on Kinetics-600 as the backbone and finetune it on FineGym, MPII Cooking 2 (MPII), and HMDB-51 split 1 (HMDB-s1).

Experiments & Discussion



*Per-class accuracy comparison is conducted between the baseline model (pretrained on Kinetics-600 and then finetuned on MPII Cooking 2, without VMPs) and our VMP-enhanced model on MPII Cooking 2, using TimeSformer as the backbone.

Conclusion & Future Work