MoETTA: Test-Time Adaptation Under Mixed Distribution Shifts with MoE-LayerNorm

Xiao Fan[1, 2*], Jingyan Jiang[3†], Zhaoru Chen[3], Fanding Huang[2], Xiao Chen[2], Qinting Jiang[2], Bowen Zhang[3], Xing Tang[3], Zhi Wang[2]

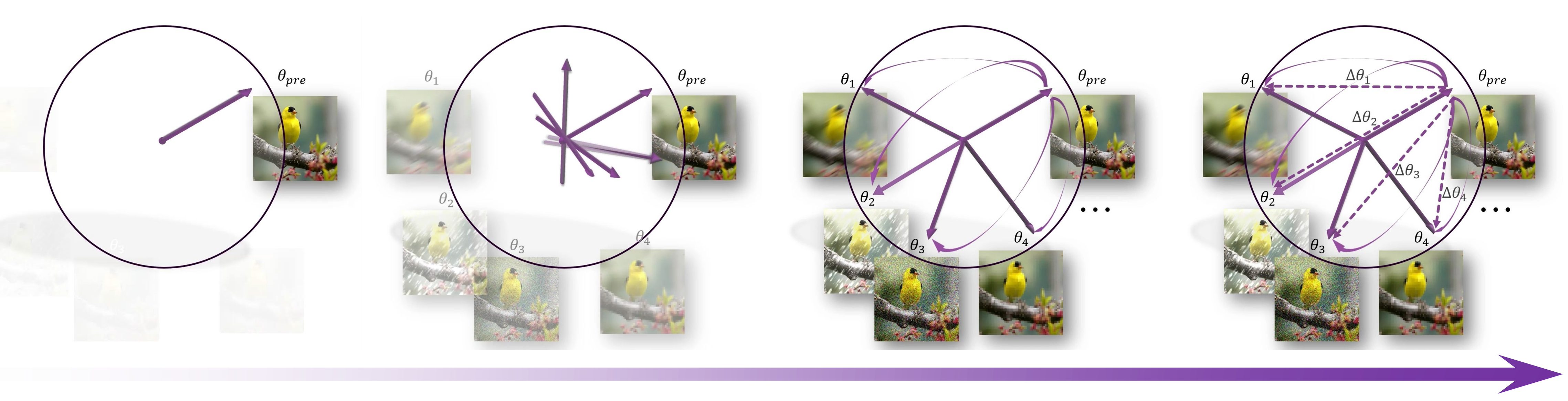
* Work completed during an internship at Tsinghua University. † Corresponding author.

WeChat

Homepage

Motivation

Illustration of Adaptation Process Under Mixed Distribution Shifts



Domain-Specific gradient signals can be inconsistent or even conflicting.

Theory

Expectation of cosine similarity between adaptation directions of different domains.

Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta} \in \mathbb{R}^d$ be i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, then

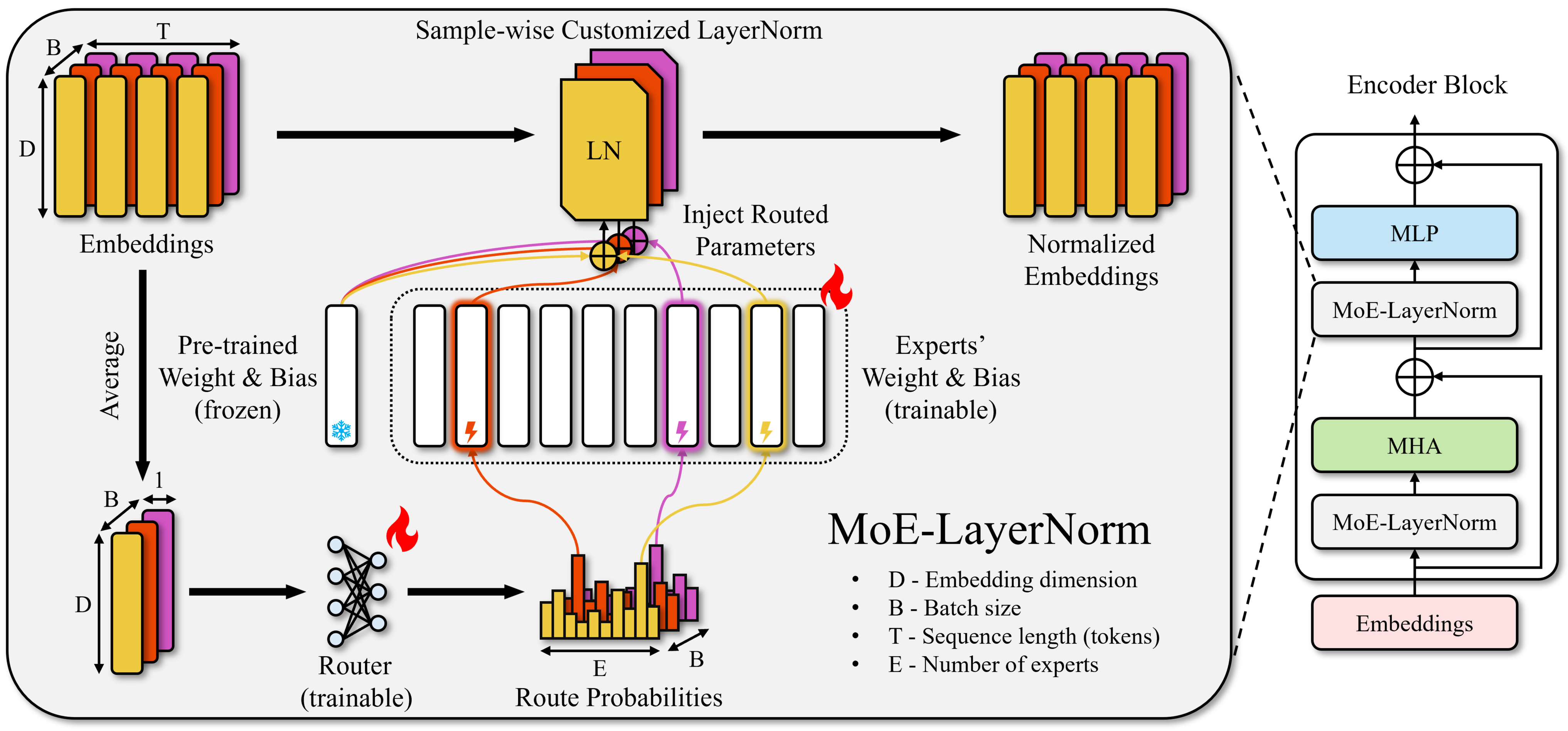
$$\lim_{d\to\infty} \mathbb{E}[\cos\langle\theta_1 - \theta, \theta_2 - \theta\rangle] = 0.5.$$

Empirical Evidence

Model	Tent	SAR	EATA	DeYO
ViT-B/16	0.69	0.64	0.66	0.71
ViT-L/16	0.34	0.37	0.28	0.26

Leveraging a diverse set of experts to represent multiple adaptation solutions within one model is particularly advantageous for mixed distribution shifts.

Methodology



Overall Loss Function

$$\frac{1}{|\mathcal{S}_t|}\sum_{\boldsymbol{x}\in\mathcal{S}_t}\exp[\mathrm{E}_0 - \mathrm{Ent}(\boldsymbol{x})]\mathrm{Ent}(\boldsymbol{x}) + \alpha_t \sum_{i=1}^{M}\mathcal{L}^i_{\text{load balancing}}$$

Sample selection

$$\mathcal{S}_t = \{x|\mathrm{Ent}(\boldsymbol{x}) < E^t_{\max} \wedge \boldsymbol{x} \in \mathcal{B}_t\}$$

Dynamic threshold

$$E^t_{\max} = \begin{cases} E^0_{\text{avg}}, & t = 0, \\ E^{t-1}_{\max} \times \dfrac{E^t_{\text{avg}}}{E^{t-1}_{\text{avg}}}, & t \geq 1. \end{cases}$$

Trade-off coefficient

$$\alpha_t = \begin{cases} \lambda \times E^0_{\text{avg}}, & t = 0, \\ \alpha_{t-1} \times \dfrac{E^t_{\text{avg}}}{E^t_{\text{avg}}}, & t \geq 1. \end{cases}$$

Differentiable load balancing loss

$$\mathcal{L}_{\text{load balancing}} = N \times \sum_{i=1}^{N} \boldsymbol{F}_i \times \boldsymbol{P}_i\,,$$

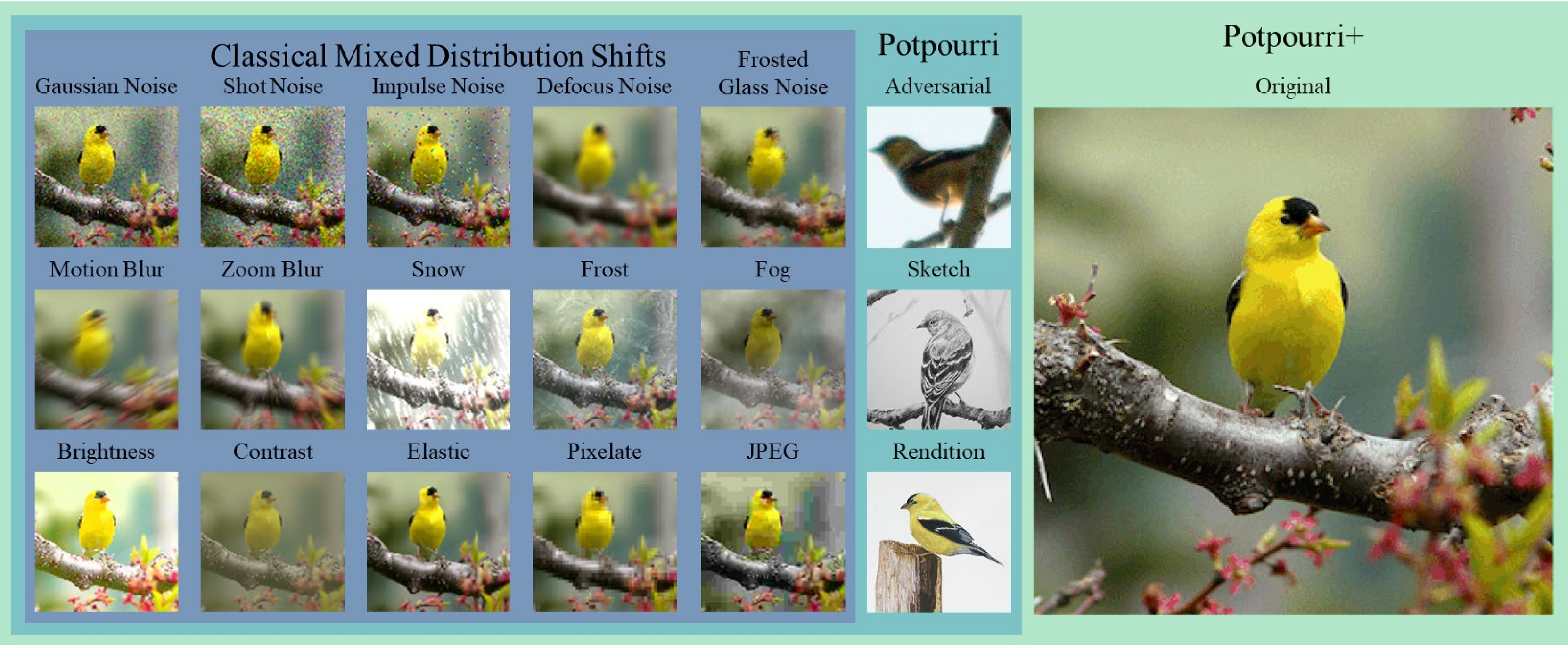
$$\boldsymbol{P}_i = \frac{1}{|\mathcal{B}_t|}\sum_{\boldsymbol{x}\in\mathcal{B}_t}\boldsymbol{p}_i(\boldsymbol{x})\,,$$

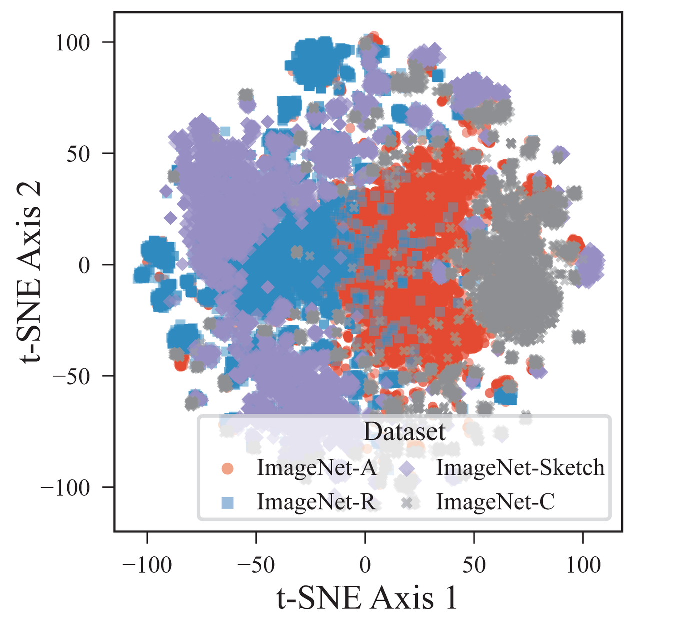
$$\boldsymbol{F}_i = \frac{1}{|\mathcal{B}_t|}\sum_{\boldsymbol{x}\in\mathcal{B}_t}\mathbb{1}_{\{\arg\max_k \boldsymbol{p}_k(\boldsymbol{x})=i\}}.$$

Benchmark

Propose two more benchmarks

- *Potpourri: More OOD samples*
- *Potpourri+: Include ID samples*



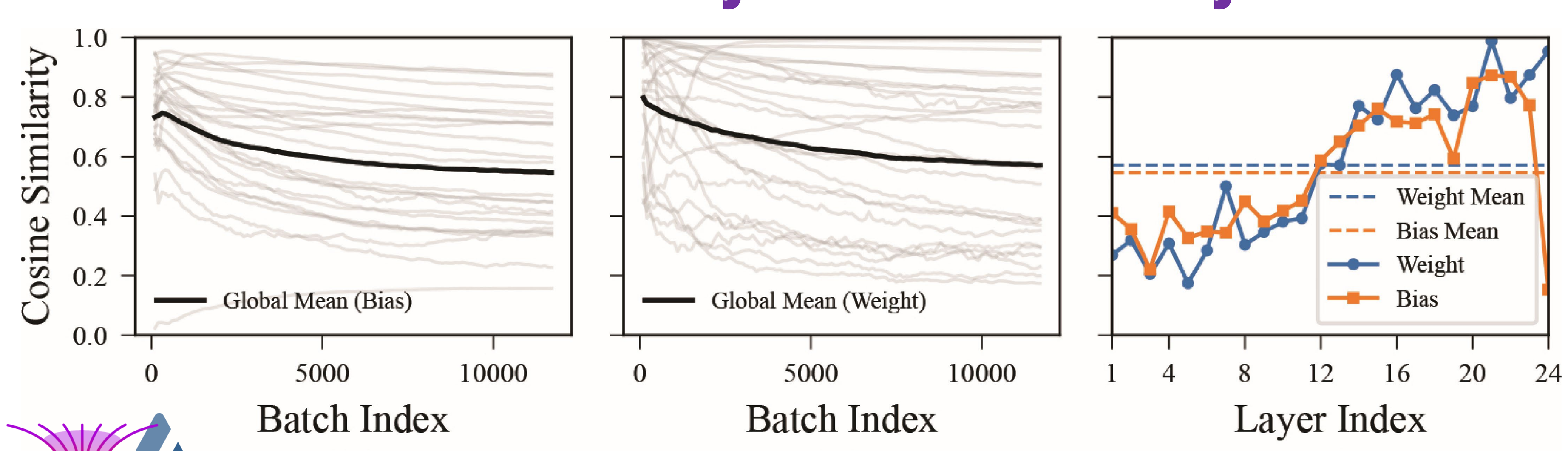


ViT CLS token t-SNE projection.

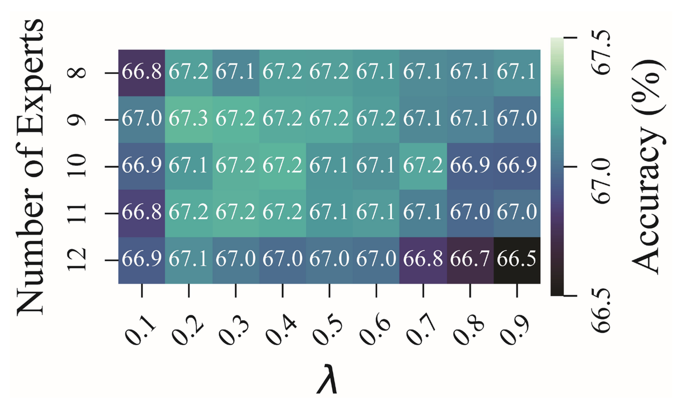
Robustness to Mixed Distribution Shifts

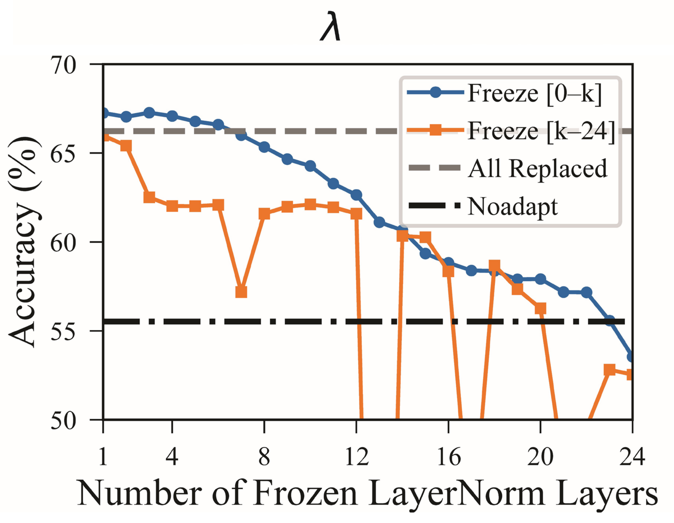
Model	Setting	Noadapt	Tent	EATA	CoTTA	SAR	DeYO	MGTTA	BECoTTA	Ours
ViT-B/16	Classical	55.52	$63.20_{0.08}$	$64.28_{0.09}$	$60.53_{0.57}$	$60.76_{0.04}$	$63.97_{0.04}$	$\underline{66.20}_{0.01}$	$61.57_{0.08}$	$\mathbf{67.20}_{0.03}$
	Potpourri	54.18	$60.99_{0.05}$	$61.99_{0.11}$	$59.67_{1.21}$	$58.71_{0.03}$	$61.66_{0.02}$	$\underline{62.98}_{0.26}$	$59.08_{0.86}$	$\mathbf{65.12}_{0.08}$
	Potpourri+	55.92	$62.28_{0.03}$	$63.17_{0.06}$	$59.26_{0.68}$	$59.99_{0.07}$	$62.90_{0.03}$	$\underline{64.35}_{0.07}$	$58.87_{3.23}$	$\mathbf{66.15}_{0.06}$
ConvNeXt-B (CNN Arch.)	Classical	54.81	$58.88_{0.06}$	$\underline{64.50}_{0.06}$	$59.65_{0.04}$	$61.67_{2.65}$	$64.32_{0.03}$	-	$50.16_{7.96}$	$\mathbf{67.40}_{0.02}$
	Potpourri	53.91	$58.23_{0.05}$	$\underline{62.69}_{0.07}$	$58.57_{0.00}$	$61.16_{0.41}$	$62.46_{0.07}$	-	$28.28_{20.54}$	$\mathbf{65.70}_{0.05}$
	Potpourri+	55.69	$59.69_{0.04}$	$\underline{63.94}_{0.07}$	$60.02_{0.02}$	$62.72_{0.10}$	$63.57_{0.06}$	-	$48.92_{9.35}$	$\mathbf{66.68}_{0.07}$

Evolution and Final-State Statistics of Expert Parameter Similarity Across MoE-LayerNorms



Hyper-Parameter Sensitivity





Experiment

Computation efficiency comparison

Method	#Act. params per sample	#Fwd	#Bwd	Used time
Noadapt	0	100%	0%	100%
Tent	0.04M	100%	100%	226%
EATA	0.04M	100%	80%	239%
SAR	0.03M	199%	175%	440%
DeYO	0.04M	196%	53%	317%
CoTTA	86.42M	199%	100%	798%
MGTTA	2.80M	100%	100%	227%
BECoTTA	0.13M	100%	86%	334%
Ours	0.23M	100%	76%	247%

Ablation analyses

	Classical	Pot.	Pot.+	Avg.
Full method	**67.25**	**65.14**	**66.21**	**66.20**
Loss Components				
w/o Sample selection	<u>67.04</u>	<u>64.01</u>	57.61	<u>62.89</u>
w/o Entropy re-weight	62.86	60.51	<u>61.79</u>	61.72
w/o $\mathcal{L}_{\text{load balancing}}$	26.27	16.27	21.29	21.28
MoE Architecture				
w/o Grad to router	<u>65.17</u>	<u>62.80</u>	<u>63.92</u>	<u>63.96</u>
w/o Sample-wise router	28.69	28.60	24.96	27.42
w/o MoE-LayerNorm	22.38	17.94	26.93	22.42
w/o Layer-wise router	17.40	27.09	15.18	19.89

AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20–27, 2026 | SINGAPORE