LAST: STABILIZING TEST-TIME ADAPTATION VIA LOGIT STANDARDIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) improves model performance under distribution shift by leveraging unlabeled test data during inference. However, existing TTA approaches often suffer from model collapse, where predictions progressively concentrate on a limited subset of classes. Despite its prevalence, the underlying mechanism of this failure mode remains insufficiently understood. In this work, we investigate the dynamics of TTA collapse and identify logit norm instability as a key factor driving prediction degeneration. We show that adaptation-induced logit norm drift disturbs the balance of prediction confidence across classes, amplifying inherent prediction biases through a self-supervised process. Specifically, excessively increased logit norms lead to overconfident predictions that strengthen erroneous class preferences, while reduced logit norms weaken class separability and impair stable adaptation. Based on this insight, we propose **L**ogit **A**daptation via **S**tandardization at **T**est time (**LAST**), a lightweight and plug-and-play module that performs class-wise logit standardization before the `softmax` layer. By normalizing the distribution of logits for each class across test samples, LAST suppresses instance-dependent logit norm fluctuations and regulates prediction confidence. Extensive experiments across disserve benchmarks show that LAST consistently improves existing TTA methods and has ability to alleviate collapse, achieving up to a 36.5% improvement in Top-1 accuracy over TENT baselines.

1 INTRODUCTION

Deep learning models are often trained under fixed data distributions but deployed in environments where the data distribution may continuously change. This discrepancy between training and deployment can lead to significant performance degradation when models encounter unseen distribution shifts Sun et al. (2022); Hendrycks et al. (2021). Test-Time Adaptation (TTA) Su et al. (2024); Niu et al. (2022); Wang et al. (2021) addresses this distribution mismatch by adapting models to the target distribution using only unlabeled test samples without requiring access to training data.

Existing TTA methods achieve effective adaptation through various strategies, such as filtering reliable samples, adjusting intermediate features, or utilizing auxiliary statistics. Based on whether they require test-time backpropagation (BP), they can be broadly categorized into two paradigms. The first paradigm adapts model parameters via gradient updates, with strategies ranging from tuning batch normalization layers Wang et al. (2022), optimizing learnable prompts Feng et al. (2023), and inserting additional adapters Lai et al. (2026). In contrast, BP-free methods avoid gradient optimization by leveraging source-domain statistics Ma et al. (2026); Lim et al. (2023), maintaining test-time prototypes Huang et al. (2026); Murphy et al. (2026), or or employing transductive strategies to estimate the target distribution Han et al. (2025b); Zhang et al. (2025).

Despite their promising performance, prior TTA methods can suffer severe degradation under long-term continuous data streams or encounter challenging distribution shifts Wang et al. (2022); Niu et al. (2023). As adaptation progresses, the model may accumulate errors, experience excessive entropy reduction, and eventually collapse into a single-class prediction (*i.e., mode collapse*). Existing literature has largely attributed this performance degradation to noisy pseudo-labels or feature drift, and proposes complex regularizers, confidence thresholding, or weight smoothing to mitigate it Hu et al. (2025); Lee et al. (2024); Niu et al. (2023). However, the underlying logit-space mechanisms driving prediction collapse remain largely unexplored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

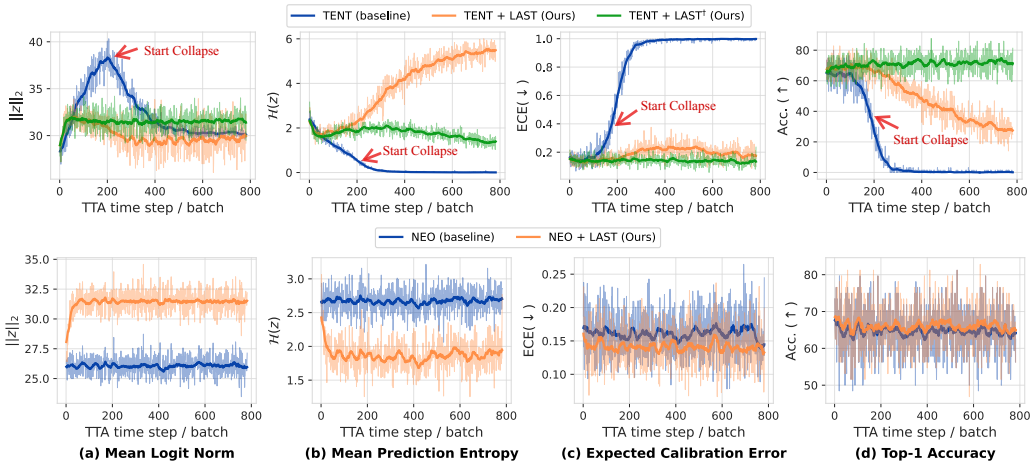


Figure 1: **Logit evolution of existing TTA paradigms on ImageNet-C (Snow, severity 5).** BP-based methods (TENT) exhibit unstable logit dynamics with excessive logit inflation before collapse. In contrast, BP-free methods maintain a consistently lower and stable logit norm, accompanied by higher prediction entropy. Integrating LAST effectively calibrates logit norm drift across both paradigms. Additional analyses of logit scaling are presented in Appendix E.

In this work, motivated by our analysis of logit evolution during adaptation, we identify *logit norm drift*, i.e., the uncontrolled evolution of logit scale (measured by logit norm) across adaptation steps, as a key factor contributing to prediction collapse in TTA. As shown in Figure 1, BP-based TENT Wang et al. (2021) exhibits severe logit norm drift before collapse, where the logit scale first increases rapidly and then eventually drops, accompanied by a continuous decrease in prediction entropy. Such uncontrolled logit drift can substantially alter prediction entropy without necessarily changing the relative ranking of class logits, making prediction confidence an unreliable indicator of correctness. Based on this observation, we propose **L**ogit **A**daptation via **S**tandardization at **T**est time (**LAST**), which mitigates collapse by stabilizing logit norm dynamics throughout adaptation.

The key idea behind LAST is to decouple optimization objectives such as entropy minimization from uncontrolled logit norm drift. Specifically, LAST leverages exponential moving averages to estimate evolving class-wise statistics from the test stream and applies adaptive class-wise $\mathcal{Z}$-score standardization to stabilize logit scales during adaptation. This standardization normalizes the logit distribution and prevents entropy minimization from inducing arbitrary logit scaling, thereby mitigating accumulated class-wise confidence biases. Since LAST operates solely in the logit space before the `softmax` operation and requires no additional backpropagation, it can be seamlessly integrated into existing TTA frameworks. Empirically, as shown in Figure 1(c) and (d), LAST consistently improves both calibration and predictive performance across existing TTA methods.

Extensive experiments demonstrate the effectiveness of LAST in mitigating model collapse and enhancing model performance. First, LAST itself is a BP-free logit standardization approach that achieves competitive performance compared with existing state-of-the-art TTA methods while introducing negligible computational and memory overhead. Second, LAST provides flexible deployment modes: it can operate as a standalone TTA method, serve as an online logit calibration module, or be integrated into existing entropy-based TTA frameworks with calibrated logits participating in subsequent backpropagation. Finally, LAST remains robust across continual adaptation and challenging test-time settings, including mini-batch adaptation and class-imbalanced scenarios, demonstrating its broad applicability across diverse TTA scenarios.

Our contributions are summarized as follows:

- We provide a logit-space analysis of existing TTA methods, revealing unstable logit-scale evolution during adaptation: excessive logit scales amplify prediction biases and may result in prediction collapse, whereas insufficient scales weaken class discrimination and reduce prediction reliability.

- We propose **LAST**, a plug-and-play logit calibration module that leverages class-wise temporal statistics to correct unstable logit-scale variations and achieve reliable confidence calibration without altering existing TTA objectives.
- Extensive experiments across continual, mini-batch, and class-imbalance TTA settings demonstrate that LAST effectively mitigates model collapse, improves calibration, and consistently enhances adaptation performance.

2 Related Work

Test-Time Adaptation (TTA) has emerged as an effective paradigm for addressing domain shifts that decrease model reliability in real-world deployment. The primary objective of TTA is to adapt a pretrained model online using only unlabeled test data, without requiring access to source data or ground-truth annotations Maharana et al. (2026); Liang et al. (2025). Based on whether backpropagation (BP) is involved during adaptation, existing TTA methods can be broadly categorized into two groups: BP-based and BP-free approaches.

Backpropagation-Based Test-Time Adaptation. Early TTA approaches mainly focus on statistical adaptation, where batch statistics are updated using statistics from intermediate target features to alleviate performance degradation caused by distribution shifts Benz et al. (2021). Subsequent methods introduce entropy minimization as a dominant adaptation objective, updating model parameters online to reduce prediction uncertainty on unlabeled test samples Wang et al. (2021); Liang et al. (2020). To improve adaptation robustness, following works propose mechanisms such as sample selection, adaptive reweighting, and optimization stabilization Ma et al. (2025); Lee et al. (2024); Niu et al. (2023; 2022). Together, these methods establish a general framework for unsupervised TTA, enabling models to adapt to unlabeled target data through self-supervised prediction feedback.

More studies have extended TTA to the continual setting, where the target distribution evolves over time. Most approaches rely on self-training or self-supervised objectives to continuously update the model using pseudo-labels or supervisory signals derived from incoming test data. Representative strategies include teacher–student frameworks Wang et al. (2022); Döbler et al. (2023), carefully designed entropy-minimization objectives Hu et al. (2025); Han et al. (2025a), and lightweight adaptation modules Lai et al. (2026), with the common goal of improving adaptation stability and mitigating catastrophic forgetting under continually changing distributions.

Backpropagation-Free Test-Time Adaptation. Recent advances in TTA have increasingly emphasized adaptation efficiency, motivating the development of BP-free approaches. One line of work performs adaptation through forward-only optimization to eliminate gradient computation. For example, LAME Boudiaf et al. (2022) adjusts classifier decision boundaries without gradient updates, and FOA Niu et al. (2024) explores derivative-free prompt optimization for Vision Transformers. While avoiding backpropagation, these methods typically require iterative forward evaluations, leading to additional inference latency.

Another line of work achieves adaptation analytically without iterative optimization. Representative approaches include recalibrating model statistics using source information Su et al. (2024); Schneider et al. (2020), aligning intermediate feature representations Murphy et al. (2026); Ma et al. (2026); Huang et al. (2026), and estimating target distributions through probabilistic modeling or Bayesian inference Zhang et al. (2025); Han et al. (2025b). Our proposed LAST falls into this category.

Overall, existing TTA methods, whether BP-based or BP-free, primarily focus on adapting models through either optimization-based updates driven by prediction objectives or depending on adjusting intermediate representations and statistics. In contrast, the final logit space before `softmax`, which directly controls prediction confidence and class probabilities, has been less explored as a target for stabilizing adaptation dynamics.

3 Methodology

3.1 Problem Statement

Let $f_\Theta : \mathcal{X} \to \mathbb{R}^K$ denote a K-class classifier trained on a labeled source distribution $P_s(x, y)$. During deployment, the model receives unlabeled samples from a target distribution $P_t(x, y)$ with

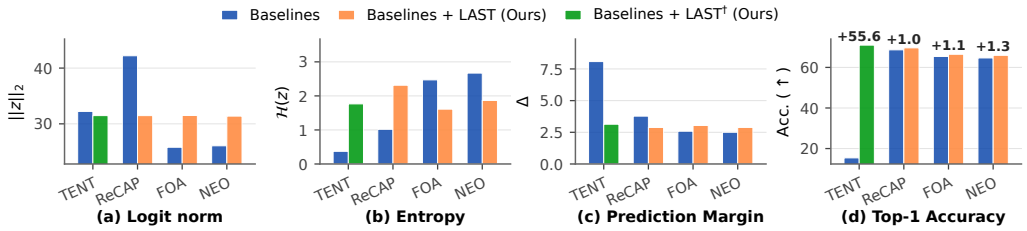


Figure 2: **Logit-scale behaviors and effect of LAST on ImageNet-C (Snow, severity 5).** (a) Logit norms reveal scale instability in BP-based methods and low-scale behavior in BP-free methods. (b) Prediction entropy, (c) margin, and (d) accuracy show that LAST improves confidence and performance. TENT+LAST† indicates that LAST is integrated into the BP loop.

distribution shift, i.e., $P_s(x) \neq P_t(x)$. This distribution shift creates a mismatch between the training and testing environments, resulting in degraded predictive performance.

Test-Time Adaptation (TTA) addresses this issue by adapting the source model using only unlabeled target samples during inference. In our work, we consider a realistic online TTA setting, where target samples arrive sequentially as a stream $\{x_t^1, x_t^2, \ldots\}$ and the model continuously updates its predictions under the evolving target environment. Such continual adaptation introduces evolving prediction statistics, making the stability of model outputs critical for reliable TTA.

3.2 Motivation

In the following, we analyze how logit norm evolves during adaptation and its impact on TTA stability. For BP-based TTA methods, such as TENT Wang et al. (2021) and DeYO Lee et al. (2024), which primarily rely on entropy for adaptation, the prediction entropy is formulated as:

$$H(z) = -\sum_k p_k \log p_k, \quad p = \mathrm{softmax}(z). \tag{1}$$

Subsequently, many methods adopt entropy minimization as the primary adaptation objective. However, reducing entropy does not necessarily improve class discrimination, as it can be achieved through logit-space transformations unrelated to class separability. For a test sample x, the corresponding logit vector $z = f_\Theta(x) \in \mathbb{R}^K$ has a squared ℓ_2 norm that can be decomposed as:

$$\|z\|_2^2 = K\,\mathrm{Var}(z) + K\bar{z}^2, \tag{2}$$

where $\bar{z} = \frac{1}{K}\sum_k z_k$, and $\mathrm{Var}(z) = \frac{1}{K}\sum_k (z_k - \bar{z})^2$. This decomposition separates logit scaling into a `softmax`-invariant mean component $\bar{z}$ and a variance component $\mathrm{Var}(z)$ that determines relative logit separation. Since larger logit scales generally reduce entropy, entropy minimization may implicitly encourage variance inflation without improving class separability. Formally,

Proposition 3.1. *Entropy minimization implicitly encourages three logit-space behaviors that increase predictive confidence without necessarily improving class separability.*

(1) Translation invariance. *For any $\beta \in \mathbb{R}$, $\mathrm{softmax}(z + \beta\mathbf{1}) = \mathrm{softmax}(z)$, and hence,*

$$H(z + \beta\mathbf{1}) = H(z).$$

Therefore, the entropy objective is invariant to global logit translation.

(2) Scale monotonicity. *For $z_\alpha = \alpha z$ with $\alpha > 0$, where $p(\alpha) = \mathrm{softmax}, (z_\alpha)$*

$$\frac{dH(z_\alpha)}{d\alpha} = -\alpha\,\mathrm{Var}_{p(\alpha)}(z) \leq 0,$$

Thus, entropy minimization can implicitly encourage larger logit scales, as increasing the scale reduces entropy while preserving the relative ordering of class logits.

(3) Probability sharpening. *The entropy gradient satisfies:*

$$-\frac{\partial H(z)}{\partial z_k} = p_k(\log p_k + H(z)).$$

The gradient reinforces existing probability preferences, causing entropy minimization to amplify prediction biases and produce increasingly concentrated predictions.

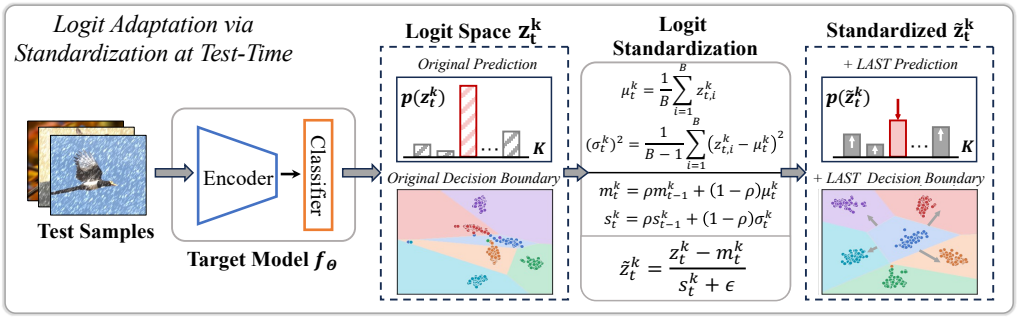


Figure 3: **Diagram of LAST.** LAST operates directly in the logit space by calibrating class-wise bias and scale, leading to more reliable decision boundaries.

Proposition 3.1 shows that entropy minimization reduces sample-wise uncertainty without explicitly controlling class-wise confidence balance. Consequently, logit scale amplification and probability sharpening may reinforce existing prediction biases, resulting in increasingly polarized predictions and potential model collapse.

Conversely, excessively reduced logit scales weaken class discrimination by pushing the output distribution toward uniformity, as observed in BP-free methods. Specifically, the first-order Taylor expansion of $p(\alpha) = \mathrm{softmax}(z_\alpha)$ around $\alpha = 0$ gives:

$$p(\alpha) = \frac{1}{K}\mathbf{1} + \frac{\alpha}{K}(z - \bar{z}\mathbf{1}) + \mathcal{O}(\alpha^2),$$

where K is the number of classes. This indicates that the deviation from the uniform distribution is first-order in α. Consequently, assuming a unique largest logit, the prediction margin satisfies:

$$p_{(1)} - p_{(2)} = \frac{\alpha}{K}(z_{(1)} - z_{(2)}) + \mathcal{O}(\alpha^2),$$

and therefore decreases linearly as $\alpha \to 0$. Hence, overly attenuated logit scales weaken class discrimination and degrade adaptation reliability. Figure 2 empirically demonstrates distinct logit-scale behaviors across existing TTA methods. These observations motivate us to explore a complementary strategy that stabilizes the logit space before `softmax` through logit standardization.

3.3 LAST: Logit Adaptation via Standardization at Test Time

Our analysis shows that existing adaptation objectives can reduce prediction uncertainty through logit-scale changes, which may not correspond to improved class discrimination and can cause unstable confidence evolution during TTA. Therefore, instead of further optimizing the entropy objective, we aim to explicitly regulate the evolution of logits to preserve a more stable predictions.

Specifically, we propose to regularize logit dynamics by performing temporal class-wise logit standardization before the `softmax` operation. Let $z_t \in \mathbb{R}^{B\times K}$ denote the logits generated by the model at adaptation step t, where B is the batch size. The class-wise logit statistics are computed as:

$$\mu_t^k = \frac{1}{B}\sum_{i=1}^{B} z_{t,i}^k, \qquad \sigma_t^k = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(z_{t,i}^k - \mu_t^k)^2}, \tag{3}$$

where $k \in \{1, \ldots, K\}$ denotes the class index. To capture temporal logit dynamics while reducing batch-wise fluctuations, LAST maintains exponential moving averages of class-wise logit statistics:

$$m_t^k = \rho m_{t-1}^k + (1-\rho)\mu_t^k, \qquad s_t^k = \rho s_{t-1}^k + (1-\rho)\sigma_t^k, \tag{4}$$

where $\rho \in [0, 1)$ controls the weight assigned to historical statistics.

In extreme online scenarios where samples arrive individually (i.e., $B = 1$), the batch-wise variance cannot be estimated reliably. To handle this case, LAST updates class-wise means using the incoming logits and maintains an exponentially weighted temporal scale estimate:

$$m_t^k = \rho m_{t-1}^k + (1-\rho)z_{t,1}^k, \quad s_t^k = \sqrt{\rho(s_{t-1}^k)^2 + (1-\rho)(z_{t,1}^k - m_{t-1}^k)^2}. \tag{5}$$

LAST standardizes the logits using estimated class-wise temporal statistics by:

$$\tilde{z}_{t,i}^k = \frac{z_{t,i}^k - m_t^k}{s_t^k + \epsilon}, \qquad \tilde{p}_{t,i}^k = \texttt{softmax}(\tilde{z}_{t,i})^k, \tag{6}$$

where ϵ is a small constant for numerical stability. This operation mitigates unstable affine variations in the logit space while preserving class-specific logit information, resulting in more stable class separation as illustrated in Figure 3. The pseudo-code is provided in Appendix B.1.

3.4 Theoretical Properties

As LAST operates directly on logit statistics, we analyze its robustness to affine logit distortions. The following proposition establishes that LAST is invariant to affine transformations of logits, reducing sensitivity to arbitrary shifts and positive scaling factors.

Proposition 3.2. *For logits transformed by a positive affine mapping $v_i^k = \alpha z_i^k + \beta$ with $\alpha > 0$, where μ^k and σ^k denote the mean and standard deviation computed over the sample dimension, the standardized logits satisfy:*

$$\frac{v_i^k - \mu_v^k}{\sigma_v^k} = \frac{z_i^k - \mu_z^k}{\sigma_z^k}.$$

Thus, logit standardization in LAST is invariant to global logit shifts and positive scaling.

4 Experiment

4.1 Experimental Setup

Datasets and Models. We evaluate our method on four benchmarks: CIFAR10-C and ImageNet-C Hendrycks & Dietterich (2019), ImageNet-Sketch Wang et al. (2019), and ImageNet-Rendition Hendrycks et al. (2021). CIFAR10-C contains 10 classes with 15 corruption types, while ImageNet-C contains 1,000 classe swith the same corruption types. ImageNet-S and ImageNet-R contain 1,000 and 200 ImageNet-derived classes, respectively, covering diverse real-world objects and visual concepts. We use three Vision Transformer (ViT) variants Dosovitskiy et al. (2021): ViT-S, ViT-B, and ViT-L, with 22M, 86M, and 307M parameters. Models are fine-tuned on ImageNet Deng et al. (2009) or CIFAR-10 Krizhevsky & Hinton (2009) based on the evaluation dataset.

Evaluation Metrics. We evaluate adaptation performance using Top-1 classification accuracy and Expected Calibration Error (ECE). Accuracy (%, ↑) measures the proportion of correctly classified samples, while ECE (%, ↓) quantifies the gap between predicted confidence and empirical accuracy Naeini et al. (2015), providing an additional measure of prediction reliability.

Baseline Methods. We compare LAST against representative test-time adaptation (TTA) methods from two categories. *Backpropagation(BP)-based methods* include TENT Wang et al. (2021), CoTTA Wang et al. (2022), SAR Niu et al. (2023), DeYO Lee et al. (2024), ReCAP Hu et al. (2025), and Surgeon Ma et al. (2025). *Backpropagation(BP)-free methods* include FOA Niu et al. (2024), ADAPT Zhang et al. (2025), PTA Huang et al. (2026), and NEO Murphy et al. (2026).

Implementation Details. Following previous work Murphy et al. (2026); Niu et al. (2024), we use a batch size of 64 throughout our experiments, except for methods evaluated under mini-batch settings. The momentum ρ in Eq. 4 is fixed to 0.9 in our experiment. We use the default hyperparameters reported in the original papers for all compared methods. All experiments are conducted on a single NVIDIA A800 GPU. Additional implementation details are provided in Appendix B.

4.2 Main Results

Comparison under Episodic Adaptation. We first evaluate LAST under the episodic setting, where the model is reset to the source parameters after each corruption type. Table 1 reports the average results on four datasets. LAST achieves competitive performance among BP-free methods and consistently improves existing adaptation methods, often leading to the best overall performance. For example, integrating LAST with TENT improves accuracy by +7.6% on ImageNet-C.

Table 1: **Episodic Adaptation.** Accuracy (%) comparison across four datasets with ViT-Base. For each baseline method, performance is reported without (w/o) and with (w/) LAST adaptation.

Datasets	*Backpropagation-Based*										*Backpropagation-Free*							
	LAST		TENT		SAR		DeYO		Surgeon		FOA		ADAPT		PTA		NEO	
	src.	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**
CIFAR10-C	84.4	85.7	87.8	89.0	85.7	86.0	89.1	**89.6**	87.5	88.4	85.0	85.7	85.1	85.1	84.5	**85.9**	85.3	85.8
ImageNet-C	55.5	60.7	58.3	**65.9**	61.9	63.8	61.6	62.6	46.9	54.4	60.5	**61.9**	59.9	60.3	56.3	58.7	59.2	60.7
ImageNet-S	44.9	50.5	50.8	54.4	48.8	52.7	50.4	53.4	52.8	**56.2**	46.1	49.9	49.4	50.0	46.5	50.1	46.8	**50.5**
ImageNet-R	59.5	61.3	66.4	68.1	63.4	64.9	68.1	69.3	68.3	**69.6**	60.7	**61.4**	59.2	59.4	58.9	60.9	60.5	61.3

Table 2: **Continual Adaptation.** Accuracy (%) comparison on ImageNet-C with ViT-Base. LAST† indicates that calibrated logits from LAST are used for backpropagation-based optimization.

	Noise			**Blur**				**Weather**				**Digital**				
Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	**Avg.**
Source	56.8	56.8	57.5	46.9	35.6	53.1	44.8	62.2	62.6	65.7	77.7	32.6	46.0	67.0	67.6	55.5
LAST	57.9	58.4	58.8	51.4	41.9	57.0	51.2	65.9	65.3	70.5	78.5	58.5	55.6	68.8	70.3	60.7
TENT	61.8	64.9	65.0	53.2	55.9	59.8	54.0	32.4	1.3	0.1	0.2	0.1	0.1	0.1	0.1	29.9
+ LAST	**62.6**	**65.3**	**65.2**	54.7	57.8	61.2	57.5	53.8	15.0	7.0	7.4	0.2	1.2	1.8	2.4	34.2
+ LAST †	62.3	65.2	64.7	**58.5**	**60.2**	**63.1**	**62.6**	**67.4**	**66.9**	**71.5**	**77.8**	**66.0**	68.5	71.4	70.1	**66.4**
SAR	59.2	61.4	61.6	54.8	55.6	59.0	56.1	61.2	62.1	64.8	76.8	59.8	57.9	68.8	69.1	61.9
+ LAST	60.7	62.5	62.7	56.9	57.6	60.9	58.9	64.6	64.0	66.7	77.4	61.2	62.3	69.6	69.9	63.7
+ LAST †	60.7	62.2	61.3	54.0	57.6	61.5	59.1	64.9	64.8	67.1	77.2	63.4	63.2	**71.9**	**71.7**	64.0
DeYO	59.6	61.9	61.1	51.6	52.9	57.6	7.5	59.6	61.8	66.9	77.1	5.0	0.1	0.1	60.6	45.6
+ LAST	60.8	62.6	61.9	53.2	55.3	59.7	15.0	62.6	63.6	68.2	77.2	5.3	0.2	0.1	61.5	47.2
+ LAST †	57.4	60.6	60.9	53.5	58.6	62.1	62.1	66.4	65.5	70.4	76.2	63.1	**68.6**	71.4	69.5	64.4
ReCAP	58.7	61.8	61.4	53.2	56.2	59.4	56.5	64.1	63.5	68.8	77.0	57.1	65.3	71.1	69.2	62.9
+ LAST	60.1	62.6	62.1	54.5	57.8	60.4	58.6	65.1	64.5	69.4	77.0	58.4	66.5	71.2	69.3	63.7
+ LAST †	57.4	59.1	59.6	52.5	58.1	61.4	61.0	65.4	65.0	70.1	77.2	62.7	68.2	71.0	69.5	63.9
FOA	**60.6**	**61.8**	**63.0**	51.9	40.1	56.7	50.7	65.1	63.5	63.8	78.5	57.6	51.9	68.5	69.1	60.2
+ LAST	60.2	61.8	62.6	**53.1**	**43.6**	**58.3**	**53.8**	**66.0**	64.5	64.8	**78.7**	**58.6**	**57.1**	**69.3**	**70.5**	**61.5**
PTA	56.5	56.8	57.5	44.5	35.8	53.0	45.0	62.2	59.8	57.0	77.6	21.1	44.6	67.2	68.5	53.8
+ LAST	57.5	58.3	58.6	50.4	42.0	56.9	51.2	65.8	**65.4**	69.5	**78.7**	54.1	55.5	68.8	70.4	60.2
NEO	57.5	57.6	58.3	47.5	36.4	53.8	45.7	62.7	63.3	68.9	77.9	37.5	46.8	67.4	68.2	56.6
+ LAST	57.9	58.4	58.8	51.4	41.9	57.0	51.2	65.9	65.3	**70.4**	78.5	58.5	55.6	68.8	70.2	60.7

Comparison under Continual Adaptation. We then evaluate LAST in the continual adaptation setting, where model updates accumulate throughout the test stream, increasing the risk of model collapse. As shown in Table 2, standalone LAST achieves the best performance among BP-free methods, attaining an average accuracy of 60.7%. We then integrate LAST into existing adaptation methods to directly calibrate the logit space, and further investigating the impact of incorporating standardized logits into the optimization process of BP-based methods. Although LAST initially exhibits suboptimal performance, it progressively achieves the best results as adaptation proceeds, demonstrating its effectiveness in stabilizing long-term adaptation. Notably, LAST mitigates model collapse in BP-based methods such as TENT and DeYO especially on latter corruption types, improving continual adaptation stability and yielding gains of +36.5% and +18.8%, respectively.

Comparison on Data Scarcity. We further evaluate LAST under severe data scarcity with batch size 1. As shown in Table 3, LAST consistently improves performance, while most baselines degrade in this setting. Due to space limitations, we report one representative LAST-enhanced variant for each BP-based and BP-free method. The improvements across both categories demonstrate LAST's robustness under low-data conditions, achieving up to +56.4% gain over TENT.

Table 3: **Batch Size = 1.** Accuracy (%) comparison on Cifar10-C using ViT-Base.

	Noise			Blur				Weather				Digital				
Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	**Avg.**
Source	76.2	77.0	74.6	92.8	69.2	90.2	94.0	93.7	92.2	90.2	97.7	93.0	82.9	57.2	84.4	84.4
LAST	76.9	78.1	**77.2**	92.8	70.5	**90.3**	**94.0**	**93.8**	**92.4**	**90.4**	**97.7**	93.2	83.3	63.2	84.7	85.2
SAR	12.3	25.6	21.0	61.8	12.8	16.6	92.6	93.7	91.8	25.3	97.7	27.8	20.1	19.2	18.4	42.5
ReCAP	76.7	77.9	76.4	92.8	72.0	90.2	94.0	93.7	92.2	90.2	**97.7**	93.0	83.0	55.3	84.5	84.6
Surgeon	53.5	65.9	26.9	93.1	**77.3**	91.2	**94.4**	**94.1**	**94.1**	**91.5**	97.5	**94.8**	**86.2**	22.9	**85.3**	77.9
TENT	14.2	16.7	18.3	38.5	13.6	23.3	47.2	18.5	36.7	25.1	86.9	21.9	16.4	38.8	18.3	29.0
+ LAST	28.2	30.3	30.7	52.7	26.8	34.7	52.4	27.6	46.0	36.9	88.2	32.6	30.5	46.8	29.1	39.6
+ LAST †	**77.8**	**79.0**	**75.9**	**93.3**	70.1	**91.8**	90.8	93.3	90.9	91.0	97.3	94.0	85.5	**68.6**	81.6	**85.4**
FOA	11.5	11.4	11.2	14.9	11.9	14.2	15.4	15.7	15.2	16.2	15.8	15.0	13.2	11.4	13.0	13.9
ADAPT	76.9	78.0	76.3	**92.9**	70.1	90.2	**94.0**	**93.8**	**92.4**	90.3	**97.7**	**93.3**	83.3	62.4	84.5	85.1
NEO	76.9	78.1	77.2	92.8	70.5	**90.3**	**94.0**	**93.8**	**92.4**	**90.4**	**97.7**	93.2	83.3	63.2	84.7	85.2
PTA	76.3	77.2	74.6	92.8	69.6	90.1	**94.0**	93.7	92.3	90.2	**97.7**	93.0	83.0	58.0	84.5	84.5
+ LAST	**77.1**	**78.2**	77.1	92.7	**70.8**	**90.3**	**94.0**	**93.8**	**92.4**	**90.4**	**97.7**	93.2	**83.4**	**63.7**	**84.8**	**85.3**

Table 4: **Label Imbalance Ratio = ∞.** Accuracy(%) comparison on ImageNet-C using ViT-Small.

	Noise			Blur				Weather				Digital				
Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	**Avg.**
Source	33.5	31.9	32.7	31.5	22.9	41.2	32.2	43.2	43.3	46.7	70.1	16.1	37.1	54.5	55.5	39.5
LAST	36.5	35.4	36.0	36.8	28.0	44.6	37.7	49.5	49.1	52.8	71.5	21.4	47.4	56.6	58.2	44.1
CoTTA	47.3	47.1	47.9	36.4	29.2	50.5	37.9	53.6	51.1	57.1	72.4	25.7	46.7	62.6	60.3	48.4
DeYO	46.0	48.8	47.0	42.0	51.9	57.8	2.5	55.8	60.5	67.2	75.3	48.1	66.8	69.6	67.1	53.8
Surgeon	49.6	46.7	46.6	42.7	35.9	49.2	39.5	18.8	55.0	58.6	74.7	12.9	14.6	63.8	62.1	44.7
SAR	49.7	51.0	**51.4**	51.1	52.4	**57.0**	54.2	40.4	**60.7**	**66.2**	**75.4**	54.3	**64.2**	**68.6**	**66.9**	57.6
+ LAST	**49.8**	**51.1**	**51.4**	**51.4**	**53.1**	**57.0**	**55.3**	**52.6**	60.4	65.7	74.5	**54.4**	**64.2**	67.8	66.3	**58.3**
PTA	33.6	32.4	32.9	32.5	24.1	41.6	33.4	44.7	44.3	47.4	70.6	16.8	39.6	54.5	55.8	40.3
ADAPT	35.6	34.5	35.2	36.7	27.3	44.1	36.8	48.0	47.0	51.7	71.2	20.4	44.5	55.9	57.8	43.1
NEO	36.0	34.9	35.6	36.1	26.4	44.3	36.1	48.2	47.6	52.4	71.8	20.3	43.7	56.7	58.2	43.2
FOA	38.5	40.4	39.4	42.6	38.8	48.0	39.0	51.5	53.8	59.3	**73.3**	41.0	49.5	**59.1**	**63.4**	49.2
+ LAST	**39.0**	**40.8**	**39.9**	**43.1**	**40.5**	**48.4**	**41.0**	**52.2**	**54.3**	**59.4**	72.9	**42.3**	**51.9**	59.0	63.0	**49.9**

Comparison on Imbalanced Label Distribution Shifts. To evaluate the robustness of LAST under complex distribution shifts with imbalanced label distributions, we further investigate its performance in an extreme label shift scenario. Following Niu et al. (2023), we set the imbalance ratio to ∞, meaning that all incoming samples within each batch almost exclusively belong to a single class. Since all BP-based methods already outperform the source baseline, we do not further investigate their backpropagation optimization process. Instead, we apply the lightweight LAST module to calibrate the logit space for both BP-based and BP-free methods. As shown in Table 4, LAST consistently improves performance under severe label distribution shifts.

4.3 Ablation Studies

Logit vs. Feature Space. As shown in Table 5, applying LAST in the logit space outperforms feature-space calibration. Modifying intermediate features degrades performance, showing that feature modification may disrupt well-learned representations.

Mean vs. Standard Deviation. We then separately evaluate the contributions of the mean and standard deviation used for logit calibration in LAST. Both components improve accuracy individually, while their combination achieves further gains, demonstrating their complementary effects.

Table 5: **Ablation Study.** Evaluated with ViT-Base on four datasets.

Dataset	Reference		Feat. Layer		Z-Score		Standardization			Plug-and-Play		
	Src.	**LAST**	Mid	Final	Mean	Std	LN	MaxAbs	MinMax	TENT	+ **LAST**	+ **LAST** †
CIFAR10-C	84.3	**86.3**	20.3	85.2	85.2	85.0	84.3	84.4	84.6	87.8	89.0	**91.6** (+3.8)
ImageNet-C	55.5	**60.7**	8.6	59.2	59.1	59.7	55.5	59.8	60.4	58.3	65.9	**68.3** (+10.0)
ImageNet-S	44.9	**50.6**	1.6	46.7	46.8	49.7	44.9	50.2	50.1	50.8	54.4	**55.4** (+4.6)
ImageNet-R	59.4	**61.3**	2.3	60.4	60.4	60.8	59.4	59.8	59.9	66.4	68.1	**68.6** (+2.2)

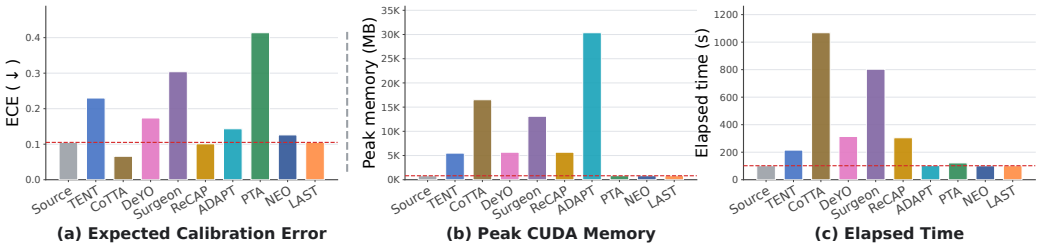


Figure 4: **Calibration and Efficiency.** Evaluated with ViT-Base on ImageNet-C.

Standardization Strategies. We compare LAST with alternative standardization strategies, including sample-wise layer normalization. LAST consistently outperforms these variants, demonstrating the effectiveness of temporal logit statistics for stabilizing TTA. By reducing logit norm drift while preserving relative class discrepancy, LAST enables more reliable confidence estimation.

BP-Free vs. BP-Integrated. We further investigate the effect of integrating LAST into the optimization of BP-based methods. Plug-and-play LAST consistently improves TENT across all benchmarks, while incorporating standardized logits into BP optimization (+LAST†) yields further gains, particularly under model collapse, with improvements of up to +10.0% on ImageNet-C.

4.4 Calibration and Efficiency

Figure 4 compares the ECE of different TTA methods on ImageNet-C. LAST achieves lower ECE than most methods, indicating improved calibration and prediction reliability. In terms of adaptation efficiency, measured by peak memory and runtime, LAST introduces negligible overhead and performs comparably to no adaptation.

4.5 Supplementary Experiments

The appendix provides additional experiments to further validate LAST, including evaluations on ImageNet-C C.1, different corruption severities C.2, varying batch sizes C.3, and different backbones C.4. We further analyze calibration D.1, momentum sensitivity D.2, scale-factor robustness D.3, confusion matrices D.4, and t-SNE visualizations of class separability D.5.

5 Conclusion

In this paper, we investigate TTA from the perspective of logit-space dynamics, revealing distinct scaling behaviors in BP-based and BP-free methods and their impact on prediction confidence and discrimination. Motivated by this analysis, we introduce **LAST**, a plug-and-play module that stabilizes class-wise logits through temporal standardization. LAST consistently improves adaptation performance, calibration, and stability across TTA methods with negligible computational overhead.

Limitations. Although LAST is architecture-agnostic by design, our experiments focus on vision transformer architectures. Extending the evaluation to other model architectures remains an interesting direction for future work. We hope this study encourages further investigation into logit-space calibration and reliable adaptation strategies for TTA.

AI USE STATEMENT

AI tools were used to assist with literature search, manuscript polishing and proofreading, code development (e.g., plotting, Bash scripts, and output extraction), and assessment of the theoretical rigor of the manuscript.

ETHICS STATEMENT

This work does not involve human participants, sensitive personal information, or applications with direct societal impact. All experiments are conducted on publicly available benchmark datasets that are widely adopted by the research community.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility and promote transparency, we publicly release our code at `https://anonymous.4open.science/r/LAST-ICLR`. Additional details for reproducing our results, including pseudocode and hyperparameter settings, are provided in Appendix B.

REFERENCES

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Revisiting batch normalization for improving corruption robustness. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 494–503. IEEE, 2021.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8344–8353, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009.

Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7704–7714. IEEE, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *International Conference on Computer Vision (ICCV)*, pp. 2704–2714. IEEE, 2023.

Jisu Han, Jaemin Na, and Wonjun Hwang. Ranked entropy minimization for continual test-time adaptation. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21798–21813. PMLR, 2025a.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025b.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021.

Zixuan Hu, Yichun Hu, Xiaotong Li, Shixiang Tang, and Lingyu Duan. Beyond entropy: Region confidence proxy for wild test-time adaptation. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24371–24390. PMLR, 2025.

Zhaohong Huang, Yuxin Zhang, Wenjing Liu, Fei Chao, and Rongrong Ji. Prototype-based test-time adaptation of vision-language models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=ejmfnVdW2X.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Guannan Lai, Da-Wei Zhou, Zhenguo Li, and Han-Jia Ye. The golden subspace: Where efficiency meets generalization in continual test-time adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2026.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations (ICLR)*, 2024.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 6028–6039. PMLR, 2020.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision (IJCV)*, 133(1):31–64, 2025.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In *International Conference on Learning Representations (ICLR)*, 2023.

Ke Ma, Jiaqi Tang, Bin Guo, Fan Dang, Sicong Liu, Zhui Zhu, Lei Wu, Cheng Fang, Ying-Cong Chen, Zhiwen Yu, et al. Surgeon: Memory-adaptive fully test-time adaptation via dynamic activation sparsity. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30514–30523. IEEE, 2025.

Xiao Ma, Young D Kwon, Pan Zhou, and Dong Ma. Architecture-agnostic test-time adaptation via backprop-free embedding alignment. In *International Conference on Learning Representations (ICLR)*, 2026.

Sarthak Kumar Maharana, Shambhavi Mishra, Yunbei Zhang, Shuaicheng Niu, Taki Hasan Rafi, Jihun Hamm, Marco Pedersoli, Jose Dolz, and Yunhui Guo. Continual test-time adaptation in computer vision: Methods, benchmarks, and future directions. *Transactions on Machine Learning Research (TMLR)*, 2026. URL https://openreview.net/pdf?id=mM3r03Xw1V.

Alexander Murphy, Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. Neo—no-optimization test-time adaptation through latent re-centering. In *International Conference on Learning Representations (ICLR)*, 2026.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, pp. 2901–2907, 2015.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning (ICML)*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*, 2023.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *International Conference on Machine Learning (ICML)*, pp. 38298–38315. PMLR, 2024.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:11539–11551, 2020.

Zixian Su, Jingwei Guo, Kai Yao, Xi Yang, Qiufeng Wang, and Kaizhu Huang. Unraveling batch normalization for realistic test-time adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 15136–15144, 2024.

Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: A synthetic driving dataset for continuous multi-task domain adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21339–21350. IEEE, 2022.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7201–7211, 2022.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

The appendix is organized as follows:

- **Appendix A** provides theoretical proofs of the propositions presented in the main paper, along with additional analyses of the derived upper bound.
- **Appendix B** presents the pseudo-code, implementation details, and baseline method details used in our experiments.
- **Appendix C** reports additional experimental results, including episodic evaluation on ImageNet-C, results under different corruption severity levels, batch sizes, and ViT architectures on four benchmarks.
- **Appendix D** provides further experimental analysis, including ECE variations, hyperparameter sensitivity, the stability of the scale factor in LAST, confusion matrix comparisons, and additional t-SNE visualizations on decision boundaries.
- **Appendix E** provides additional analyses of logit scaling, including detailed evaluations across corruption types, baselines, and continual adaptation settings.

A Theoretical Analysis

A.1 Proof of Proposition 3.1

Proposition A.1. *Entropy minimization implicitly encourages three logit-space behaviors that increase predictive confidence without necessarily improving class separability.*

(1) Translation invariance. *For any $\beta \in \mathbb{R}$, $\mathrm{softmax}(z + \beta\mathbf{1}) = \mathrm{softmax}(z)$, and hence,*

$$H(z + \beta\mathbf{1}) = H(z).$$

Therefore, the entropy objective is invariant to global logit translation.

(2) Scale monotonicity. *For $z_\alpha = \alpha z$ with $\alpha > 0$, where $p(\alpha) = \mathrm{softmax}(z_\alpha)$,*

$$\frac{dH(z_\alpha)}{d\alpha} = -\alpha \operatorname{Var}_{p(\alpha)}(z) \leq 0,$$

Thus, entropy minimization can implicitly encourage larger logit scales, as increasing the scale reduces entropy while preserving the relative ordering of class logits.

(3) Probability sharpening. *The entropy gradient satisfies:*

$$-\frac{\partial H(z)}{\partial z_k} = p_k(\log p_k + H(z)).$$

The gradient reinforces existing probability preferences, causing entropy minimization to amplify prediction biases and produce increasingly concentrated predictions.

Proof. We prove each of the three properties in order.

(1) Translation invariance. After adding a constant offset β to all logits, the `softmax` probability of class k becomes:

$$\frac{\exp(z_k + \beta)}{\sum_j \exp(z_j + \beta)} = \frac{\exp(\beta)\exp(z_k)}{\exp(\beta)\sum_j \exp(z_j)} = \frac{\exp(z_k)}{\sum_j \exp(z_j)}.$$

Therefore,

$$\mathrm{softmax}(z + \beta\mathbf{1}) = \mathrm{softmax}(z).$$

Since the entropy is determined only by the softmax probabilities,

$$H(z + \beta\mathbf{1}) = -\sum_k p_k(z + \beta\mathbf{1}) \log p_k(z + \beta\mathbf{1}) = -\sum_k p_k(z) \log p_k(z) = H(z).$$

Hence, the entropy objective is invariant along the global translation direction.

(2) Scale monotonicity. Let the scaled logits be parameterized by a positive scalar α:

$$p_k(\alpha) = \frac{\exp(\alpha z_k)}{\sum_j \exp(\alpha z_j)}.$$

The corresponding entropy can be written as:

$$H(\alpha) = -\sum_k p_k(\alpha) \log p_k(\alpha) = -\alpha \mathbb{E}_{p(\alpha)}[z] + \log \sum_j \exp(\alpha z_j).$$

Differentiating with respect to α yields:

$$\frac{dH}{d\alpha} = -\alpha \frac{d\mathbb{E}_{p(\alpha)}[z]}{d\alpha},$$

where the first-order terms cancel due to the definition of the expectation. Furthermore,

$$\frac{d\mathbb{E}_{p(\alpha)}[z]}{d\alpha} = \sum_k z_k \frac{\partial p_k(\alpha)}{\partial \alpha}.$$

Using the derivative of the softmax function,

$$\frac{\partial p_k(\alpha)}{\partial \alpha} = p_k(\alpha) \left(z_k - \mathbb{E}_{p(\alpha)}[z] \right),$$

we obtain:

$$\frac{d\mathbb{E}_{p(\alpha)}[z]}{d\alpha} = \mathbb{E}_{p(\alpha)}[z^2] - \mathbb{E}_{p(\alpha)}[z]^2 = \mathrm{Var}_{p(\alpha)}(z).$$

Therefore,

$$\frac{dH}{d\alpha} = -\alpha \, \mathrm{Var}_{p(\alpha)}(z) \leq 0,$$

which is hold for $\alpha > 0$. Hence, increasing the logit scale monotonically decreases the predictive entropy. Moreover, positive scaling preserves the ordering of logits:

$$\arg\max_k \alpha z_k = \arg\max_k z_k,$$

which indicates that logit scaling changes prediction confidence without altering the predicted class.

(3) Probability sharpening. Using the softmax derivative,

$$\frac{\partial p_i}{\partial z_k} = p_i(\mathbf{1}_{i=k} - p_k),$$

the entropy gradient with respect to z_k is:

$$\frac{\partial H}{\partial z_k} = -\sum_i (1 + \log p_i) \frac{\partial p_i}{\partial z_k}.$$

Substituting the softmax derivative yields:

$$\frac{\partial H}{\partial z_k} = -p_k(\log p_k + H(z)).$$

Therefore, the gradient descent direction for entropy minimization is:

$$-\frac{\partial H}{\partial z_k} = p_k(\log p_k + H(z)).$$

This gradient is probability-dependent: the update magnitude in optimization is scaled by the current prediction probability p_k. Consequently, classes with larger probabilities receive stronger gradient feedback, which further sharpens the predicted distribution. Therefore, entropy minimization promotes confidence sharpening around existing predictions rather than encouraging class diversity. □

A.2 Proof of Proposition 3.2

A.3 Theoretical Properties

Proposition A.2. *For logits transformed by a positive affine mapping $v_i^k = \alpha z_i^k + \beta$ with $\alpha > 0$, where μ^k and σ^k denote the mean and standard deviation computed over the sample dimension, the standardized logits satisfy:*

$$\frac{v_i^k - \mu_v^k}{\sigma_v^k} = \frac{z_i^k - \mu_z^k}{\sigma_z^k}.$$

Thus, logit standardization in LAST is invariant to global logit shifts and positive scaling.

Proof. For the transformed logits, the batch mean becomes:

$$\mu_v^k = \alpha \mu_z^k + \beta.$$

Consequently, the centered logit satisfies: $v_i^k - \mu_v^k = \alpha(z_i^k - \mu_z^k)$. The corresponding standard deviation is:

$$\sigma_v^k = \sqrt{\frac{1}{B-1}\sum_i (v_i^k - \mu_v^k)^2} = \alpha \sigma_z^k,$$

where the last equality follows from $\alpha > 0$. Therefore,

$$\frac{v_i^k - \mu_v^k}{\sigma_v^k} = \frac{\alpha(z_i^k - \mu_z^k)}{\alpha \sigma_z^k} = \frac{z_i^k - \mu_z^k}{\sigma_z^k}.$$

Hence, standardization removes both the global shift and positive scaling of logits, proving the invariance property of LAST. □

A.4 Bounded Expected Magnitude in LAST

Beyond characterizing the effects of logit translation and scaling, we further establish an upper bound on the expected logit magnitude induced by LAST under accurate estimation of target-stream.

Lemma A.3. *Assume that the moving statistics in LAST accurately estimate the target-stream statistics, where $m^k = \mathbb{E}[z^k]$ and $s^k = \sqrt{\mathrm{Var}(z^k)}$. Then each standardized logit dimension satisfies:*

$$\mathbb{E}[\tilde{z}^k] = 0, \qquad \mathrm{Var}(\tilde{z}^k) = 1.$$

For a K-dimensional standardized logit vector $\tilde{z}$, we have,

$$\mathbb{E}[\|\tilde{z}\|_2^2] = K, \qquad \mathbb{E}[\|\tilde{z}\|_2] \le \sqrt{K}.$$

Proof. By LAST standardization, $\tilde{z}^k = (z^k - m^k)/s^k$. Under accurate estimation of stream statistics, each normalized dimension has zero mean and unit variance.

Therefore, the expected squared norm is:

$$\mathbb{E}[\|\tilde{z}\|_2^2] = \sum_{k=1}^{K} \mathbb{E}[(\tilde{z}^k)^2] = \sum_{k=1}^{K} \left(\mathrm{Var}(\tilde{z}^k) + (\mathbb{E}[\tilde{z}^k])^2\right) = K.$$

Finally, since $\sqrt{x}$ is concave for $x \ge 0$, Jensen's inequality gives:

$$\mathbb{E}[\|\tilde{z}\|_2] = \mathbb{E}[\sqrt{\|\tilde{z}\|_2^2}] \le \sqrt{\mathbb{E}[\|\tilde{z}\|_2^2]} = \sqrt{K}.$$

Thus, LAST normalization constrains the expected magnitude of standardized logits under accurate target-stream statistics. □

This upper bound is consistent with the empirical results shown in Figure 1. Specifically, the orange curve converges to approximately 31.6 on ImageNet-C, which is close to $\sqrt{1000}$.

B Reproduction Details

B.1 Pseudo-code

Algorithm 1 presents the pseudo-code for the proposed LAST module. We additionally describe the case of batch size 1 (B=1) and detail how LAST can be integrated into existing backpropagation(BP)-based TTA methods.

Algorithm 1 Logit Adaptation via Standardization at Test-Time (**LAST**)

Require: Source model f_Θ, test stream $\{\mathbf{x}_t\}_{t\geq 1}$, momentum ρ

Initialize $\mathbf{m}_0 = \mathbf{0}$, $\mathbf{s}_0 = \mathbf{1}$
for $t = 1, 2, \ldots$ **do**
  Obtain logits $\mathbf{z}_t = f_\Theta(\mathbf{x}_t) \in \mathbb{R}^{B\times K}$
    ▷ From the source model or a base TTA method
  if $B > 1$ **then**
    Compute batch mean $\boldsymbol{\mu}_t$ and std $\boldsymbol{\sigma}_t$ ▷ Eq. (3)
    $\mathbf{m}_t \leftarrow \rho\mathbf{m}_{t-1} + (1-\rho)\boldsymbol{\mu}_t$; $\mathbf{s}_t \leftarrow \rho\mathbf{s}_{t-1} + (1-\rho)\boldsymbol{\sigma}_t$ ▷ Eq. (4)
  else
    $\mathbf{m}_t \leftarrow \rho\mathbf{m}_{t-1} + (1-\rho)\mathbf{z}_t$; $\mathbf{s}_t \leftarrow \sqrt{\rho\mathbf{s}_{t-1}^2 + (1-\rho)(\mathbf{z}_t - \mathbf{m}_{t-1})^2}$ ▷ Eq. (5)
  end if
  Standardize logits: $\tilde{\mathbf{z}}_t \leftarrow (\mathbf{z}_t - \mathbf{m}_t)/(\mathbf{s}_t + \epsilon)$
  Compute calibrated predictions: $\tilde{\mathbf{p}}_t \leftarrow \texttt{softmax}(\tilde{\mathbf{z}}_t)$ ▷ Eq. (6)
  // Parameter update for base TTA
  if the base TTA method requires BP **then**
    if LAST participates in BP **then**
      Update Θ using $\mathcal{L}_{TTA}(\tilde{\mathbf{z}}_t)$
    else
      Update Θ using the original objective $\mathcal{L}_{TTA}(\mathbf{z}_t)$
    end if
  else
    Keep Θ unchanged
  end if
end for

B.2 Implementation

To facilitate reproducibility, our source code is available at `https://anonymous.4open.science/r/LAST-ICLR`. A substantial portion of our implementation is built upon the codebases of NEO Murphy et al. (2026) and SAR Niu et al. (2023). For DeYO Lee et al. (2024), ReCAP Hu et al. (2025), PTA Huang et al. (2026), and ADAPT Zhang et al. (2025), we use the implementations provided in their official repositories.

Unless otherwise specified, we use the default hyperparameters reported in the respective papers. For Surgeon Ma et al. (2025), we use Adam with a learning rate of 10^{-5}. For TENT Wang et al. (2021), we set the learning rate to 2.5×10^{-4} across all datasets. For ADAPT, we adopt its transductive implementation, while for NEO, we use its parameter-free variant, which yields its best performance.

B.3 Baseline Methods

We compare LAST with the following state-of-the-art TTA methods. Below, we briefly introduce the baseline methods considered in our experiments.

The backpropagation (BP)-based methods include:

- **TENT** Wang et al. (2021) is the first TTA method to adapt a model by minimizing the entropy of its predictions, thereby encouraging more confident predictions on test samples.

- **CoTTA** Wang et al. (2022) considers the long-term continual adaptation setting at test time and mitigates error accumulation through weight-averaged models and augmentation-averaged predictions.
- **SAR** Niu et al. (2023) introduces a challenging wild TTA setting characterized by mixed distribution shifts, small batch sizes, and online imbalanced label distributions. It proposes sharpness-aware reliable entropy minimization to mitigate the adverse effects of noisy samples associated with large gradients, thereby preventing adaptation instability and potential model collapse.
- **DeYO** Lee et al. (2024) prioritizes samples containing dominant shape information and employs a dual-selection criterion based on entropy to identify high-quality samples for stable TTA.
- **ReCAP** Hu et al. (2025) identifies region confidence as an effective alternative to conventional entropy and leverages it to capture semantic variations in the representation space.
- **Surgeon** Ma et al. (2025) introduces a dynamic activation sparsity strategy that directly prunes activations to reduce the memory overhead of TTA.

The backpropagation (BP)-free methods include:

- **FOA** Niu et al. (2024) introduces forward-optimization adaptation for adapting models deployed on edge devices. It designs a fitness function based on the discrepancy between test-time and training-time statistics, together with model prediction entropy.
- **ADAPT** Zhang et al. (2025) reframes TTA as a Gaussian probabilistic inference problem by explicitly modeling class-conditional feature distributions and using the resulting class-conditional likelihoods for adaptation.
- **PTA** Huang et al. (2026) addresses inference latency and suboptimal adaptation performance by maintaining a set of class-specific knowledge prototypes that progressively accumulate knowledge from test samples.
- **NEO** Murphy et al. (2026) improves the alignment between source and distribution-shifted samples by re-centering target embeddings around the origin, thereby improving prediction accuracy under distribution shifts.

C Additional Experimental Results

C.1 Episodic Adaptation on ImageNet-C

We provide a detailed accuracy comparison on ImageNet-C across different corruption types in Table 6. We observe that LAST alone achieves competitive, and in some cases superior, performance compared with BP-free baselines. Furthermore, integrating LAST into existing TTA methods consistently improves performance and achieves state-of-the-art results.

C.2 Different Severity

We provide a more detailed analysis of different corruption severity levels for all compared methods on CIFAR10-C and ImageNet-C in Table 7. As the corruption severity increases, the performance of some methods degrades, indicating their limited robustness under more severe distribution shifts. Nevertheless, incorporating LAST consistently improves the performance of both backpropagation (BP)-based and BP-free methods across all severity levels. Moreover, the performance gains introduced by LAST become increasingly significant as the corruption severity increases, demonstrating its effectiveness in handling more challenging corruption scenarios.

C.3 Different Batch Size

We then investigate the impact of batch size on TTA methods and report detailed results under different batch size settings, ranging from [1, 4, 8, 16, 32, 64, 128], in Table 8. We observe that LAST remains robust across all batch size settings, which can be attributed to the adoption of EMA

Table 6: **Episodic Adaptation.** Accuracy (%) comparison on ImageNet-C with ViT-Base. For each baseline method, performance is reported without (w/o) and with (w/) LAST adaptation.

Corruption	LAST		*Backpropagation-Based*								*Backpropagation-Free*							
			CoTTA		SAR		DeYO		Surgeon		FOA		ADAPT		PTA		NEO	
	w/o	**w/**	src.	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**	w/o	**w/**
Noise																		
Gaussian	56.8	57.9	63.5	**63.8**	59.2	60.7	59.8	61.0	60.9	62.1	**60.6**	60.2	57.4	57.5	56.5	57.5	57.5	57.9
Shot	56.8	58.3	64.5	**64.8**	60.2	61.9	61.5	62.4	61.8	62.9	59.2	**59.4**	57.4	57.6	56.8	58.0	57.6	58.4
Impulse	57.5	58.8	64.0	**64.6**	60.7	61.9	60.3	61.3	61.0	62.6	60.6	**60.8**	57.9	58.0	57.4	58.5	58.3	58.8
Blur																		
Defocus	46.9	51.4	54.6	58.0	57.2	59.0	57.6	**59.2**	53.9	55.9	52.7	**54.2**	50.9	51.4	47.7	49.8	49.7	51.4
Glass	35.6	41.9	49.9	55.8	56.1	57.9	58.4	**59.8**	3.7	11.5	40.0	**43.5**	40.4	41.4	36.7	41.4	38.2	41.9
Motion	53.1	57.0	64.6	**66.3**	61.7	63.3	64.4	65.5	59.9	61.7	57.1	**58.5**	55.2	55.7	53.7	56.6	54.9	57.1
Zoom	44.8	51.2	57.7	62.3	57.8	60.7	61.9	**63.8**	53.3	57.2	49.3	**52.8**	48.9	50.0	45.9	50.2	47.6	51.2
Weather																		
Snow	62.2	65.9	68.9	**70.6**	65.8	68.2	69.0	70.0	69.0	70.1	65.6	**66.7**	64.7	65.0	63.0	65.5	64.7	65.9
Frost	62.6	65.3	69.2	**70.4**	63.3	65.2	66.3	67.1	26.0	38.8	65.5	**66.2**	65.9	66.0	61.5	64.5	65.0	65.3
Fog	65.7	70.5	73.7	**74.5**	66.6	68.4	71.0	71.5	2.5	37.7	70.8	71.2	71.6	**71.8**	67.8	65.7	71.3	70.5
Brightness	77.7	78.5	78.7	79.2	78.8	79.2	78.6	78.8	79.5	**79.6**	78.1	78.5	78.3	78.4	78.0	**78.8**	78.1	78.6
Digital																		
Contrast	32.6	58.5	3.7	7.9	36.0	**39.1**	0.2	0.3	3.8	28.2	57.7	59.6	60.1	**60.3**	35.5	40.4	58.6	58.5
Elastic	46.0	55.6	65.0	69.0	62.5	65.9	69.8	**70.8**	27.3	43.8	51.8	**57.2**	52.7	54.1	47.9	55.0	50.0	55.6
Pixelate	67.0	68.8	73.6	**74.8**	72.1	72.9	74.0	74.5	71.3	72.2	68.7	**69.5**	68.0	68.2	67.3	68.7	67.7	68.8
JPEG	67.6	70.2	71.5	**72.8**	70.1	72.0	71.8	72.5	69.7	71.8	69.6	**70.7**	69.5	69.7	68.6	70.1	68.8	70.2
Average	55.5	60.7	61.5	63.7	61.9	**63.8**	61.6	62.6	46.9	54.4	60.5	**61.9**	59.9	60.3	56.3	58.7	59.2	60.7

Table 7: **Different Severity Levels.** Accuracy (%) comparison under different corruption severity levels on CIFAR10-C and ImageNet-C. LAST consistently improves both BP-based and BP-free methods across all severity levels, with larger gains observed under more severe corruptions.

	Methods	CIFAR10-C					ImageNet-C				
		Level 1	Level 2	Level 3	Level 4	Level 5	Level 1	Level 2	Level 3	Level 4	Level 5
	Source	96.3	95.0	93.2	89.7	84.4	79.4	75.3	71.8	65.9	55.5
	LAST	**96.4**	**95.1**	93.4	**90.3**	85.7	**80.0**	**76.4**	**73.6**	**68.6**	**60.7**
BP-Based	SAR	96.3	95.0	93.2	89.8	85.7	79.7	76.5	74.2	69.5	61.9
	DeYO	96.8	95.8	94.0	92.5	**89.1**	79.4	76.6	74.7	71.1	61.6
	ReCAP	96.5	95.3	93.7	90.8	86.7	79.7	76.8	74.9	71.3	65.7
	Surgeon	96.7	95.7	94.2	91.5	87.5	79.7	76.4	74.0	69.6	46.9
	TENT	**96.9**	96.1	94.8	92.5	87.8	**81.2**	78.4	76.6	66.4	58.3
	+ LAST	**96.9**	**96.2**	**94.9**	**92.8**	89.0	**81.2**	**78.5**	**76.9**	**72.6**	**65.9**
BP-Free	FOA	96.3	94.9	93.2	89.9	85.0	80.0	76.2	73.7	68.4	60.5
	ADAPT	96.3	95.0	93.3	90.0	85.1	80.0	76.2	73.3	68.0	59.9
	PTA	96.3	95.0	93.2	89.8	84.5	79.8	75.7	72.4	66.5	56.3
	NEO	96.3	95.0	93.4	90.1	85.3	**80.0**	76.2	73.1	67.7	59.2
	+ LAST	**96.4**	**95.1**	**93.5**	**90.3**	**85.8**	**80.0**	**76.4**	**73.6**	**68.6**	**60.7**

Table 8: **Different Batch Sizes.** Accuracy (%) comparison under different batch size settings on TTA methods. LAST remains robust across various batch sizes, while BP-based methods exhibit higher sensitivity to batch size variations.

		ImageNet-S							ImageNet-R						
	Methods	B=1	B=4	B=8	B=16	B=32	B=64	B=128	B=1	B=4	B=8	B=16	B=32	B=64	B=128
	Source	44.9	44.9	44.9	44.9	44.9	44.9	44.9	59.5	59.5	59.5	59.5	59.5	59.5	59.5
	LAST	46.7	48.1	49.2	49.9	**50.3**	**50.5**	50.7	**60.2**	**60.5**	**60.5**	**61.0**	**61.3**	**61.3**	**61.3**
BP-Based	TENT	21.6	51.0	52.0	51.9	51.5	50.8	49.8	11.4	32.9	46.0	51.2	63.9	66.4	65.5
	CoTTA	0.2	0.5	1.4	34.7	48.6	42.5	47.0	0.5	1.5	15.1	58.3	65.4	63.9	60.9
	SAR	47.8	50.1	50.6	50.6	49.8	48.8	47.5	67.2	69.1	68.5	67.2	65.4	63.4	61.8
	DeYO	0.2	13.1	43.0	48.2	49.9	50.6	50.4	35.8	60.2	62.7	67.0	**68.9**	68.2	66.8
	Surgeon	0.3	52.7	53.0	53.1	52.9	52.8	52.9	26.3	68.6	68.8	**69.0**	68.8	68.4	68.4
	ReCAP	49.9	50.4	50.4	50.3	50.2	50.4	50.2	65.3	**67.5**	68.0	68.0	68.0	67.8	67.7
	+ LAST	**51.0**	**51.6**	**52.7**	**53.1**	**53.2**	**53.5**	**53.5**	**65.8**	**67.5**	**68.4**	68.8	68.8	**68.8**	**68.6**
BP-Free	FOA	0.1	45.4	47.0	46.9	46.2	46.1	46.7	0.4	54.8	59.3	60.7	60.7	60.7	60.6
	PTA	46.5	46.5	46.5	46.5	46.5	46.5	46.5	58.9	58.9	58.9	58.9	58.9	58.9	58.9
	NEO	46.8	46.8	46.8	46.8	46.8	46.8	46.8	60.5	60.5	60.5	60.5	60.5	60.5	60.5
	ADAPT	49.4	49.4	49.4	49.4	49.4	49.4	49.4	59.2	59.2	59.2	59.2	59.2	59.2	59.2
	+ LAST	**49.4**	**50.0**	**50.2**	**50.1**	50.1	50.0	50.0	59.2	59.6	59.5	59.4	59.4	59.4	59.3

for updating model statistics. Furthermore, the performance of LAST generally improves as the batch size increases, since larger batches provide more reliable and representative statistics.

Besides, most BP-free methods, such as PTA, NEO, and LAST, maintain consistent performance across different batch sizes, demonstrating their independence from batch size variations. In contrast, BP-based methods, such as CoTTA and Surgeon, are more sensitive to batch size changes, as their optimization procedures rely heavily on batch-level statistics.

C.4 Different Model Sizes

We further investigate the impact of model size on TTA methods and report results across four datasets in Table 9. We observe that methods such as TENT and ReCAP are more prone to model collapse when using ViT-Small, potentially because smaller models have limited capacity to capture sufficient and comprehensive representations under distribution shifts. As the model size increases, their performance becomes more stable. Besides, LAST consistently achieves similar performance gains across three different model sizes, demonstrating its robustness and scalability with respect to model capacity.

D Additional Experimental Analysis

D.1 LAST Improves Calibration of TTA Methods

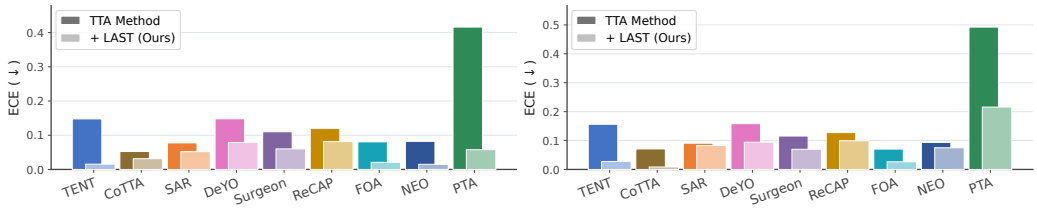
We evaluate model calibration to validate the effectiveness of LAST in calibrating output predictions by comparing the expected calibration error (ECE) before and after applying LAST. A lower ECE indicates better calibration, while a higher ECE suggests poor calibration, where the model may be overconfident in incorrect predictions or underconfident in correct predictions. As shown in Figure 5, all TTA methods equipped with LAST achieve lower ECE values, demonstrating the capability of LAST to improve prediction calibration and produce more reliable confidence estimates.

D.2 Hyper-parameter Sensitivity Analysis

Figure 6 evaluates the sensitivity of the momentum hyper-parameter ρ in Eq. 4, which controls the balance between historical statistics and current-batch statistics. We observe that, regardless of whether LAST is used independently (a), integrated into existing TTA methods (b), or combined

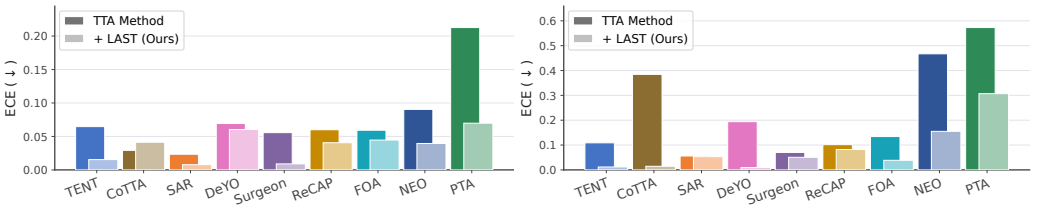
Table 9: **Different Model Sizes.** Accuracy (%) comparison using different model sizes across four datasets. LAST consistently provides stable performance improvements across various model capacities.

	Methods	CIFAR10-C			ImageNet-C			ImageNet-R			ImageNet-S		
		ViT-S	ViT-B	ViT-L	ViT-S	ViT-B	ViT-L	ViT-S	ViT-B	ViT-L	ViT-S	ViT-B	ViT-L
	Source	85.1	84.4	85.8	39.5	55.5	61.1	46.2	59.5	64.6	32.8	44.9	51.8
	LAST	87.5	85.7	87.8	44.1	**60.7**	63.4	**49.2**	**61.3**	66.7	**39.0**	**50.5**	56.7
BP-Based	TENT	91.0	87.8	89.9	49.0	58.3	71.2	15.3	66.4	59.3	9.8	50.8	57.5
	CoTTA	45.5	43.5	54.7	45.9	61.5	61.6	47.4	63.8	66.1	34.8	49.8	55.3
	SAR	86.3	85.7	86.3	56.9	61.9	67.4	55.3	63.4	66.8	42.4	48.8	54.6
	DeYO	**91.8**	**89.1**	**90.2**	**51.7**	**61.6**	49.5	50.7	68.1	69.4	41.9	50.4	55.6
	ReCAP	74.3	86.7	89.5	39.5	65.7	61.1	46.2	67.7	64.6	32.8	50.2	51.8
	Surgeon	90.9	87.5	88.6	46.6	46.9	68.8	54.6	68.3	72.7	42.3	52.8	58.8
	+ LAST	91.5	88.4	89.8	50.3	54.4	**69.0**	**56.6**	**69.6**	**73.3**	**46.4**	**56.2**	**61.0**
BP-Free	FOA	87.1	85.0	87.8	**47.3**	60.5	63.1	47.8	60.7	66.5	36.1	46.1	54.1
	ADAPT	86.4	85.1	86.4	43.3	59.9	**63.6**	47.4	59.2	65.6	38.6	49.4	57.1
	NEO	87.0	85.3	87.7	43.2	59.2	63.2	48.7	60.5	66.6	37.0	46.8	54.5
	PTA	85.9	84.5	86.0	40.3	56.3	61.4	47.1	58.9	65.4	35.2	46.5	54.4
	+ LAST	**87.8**	**85.9**	**88.0**	42.7	58.7	62.8	48.7	60.9	**67.0**	37.8	50.1	**57.4**



(a) Gaussian Noise, Level 5

(b) Defocus Blur, Level 5



(c) Brightness, Level 5

(d) Contrast, Level 5

Figure 5: ECE comparison among TTA methods.

with backpropagation-based adaptation (c), the performance remains relatively stable across different values of ρ. In particular, LAST consistently achieves strong performance with ρ around 0.9.

D.3 Robustness to Logit Norm Drift

To further verify the effectiveness of LAST in regulating logit scales, we artificially vary the scale factor α under logit standardization and examine the resulting changes in prediction accuracy and entropy. As illustrated in Figure 7, existing TTA methods achieve relatively stable accuracy under different logit scales; however, their prediction entropy varies considerably with changes in α. In contrast, incorporating LAST improves the robustness of these methods, leading to more consistent entropy behavior across different scale factors. This finding aligns with Proposition 3.1, confirming that LAST effectively constrains logit magnitudes and reduces the impact of scale variations.

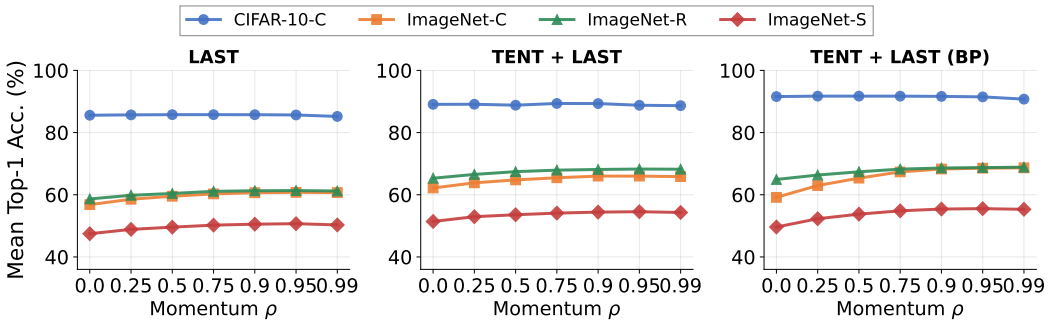
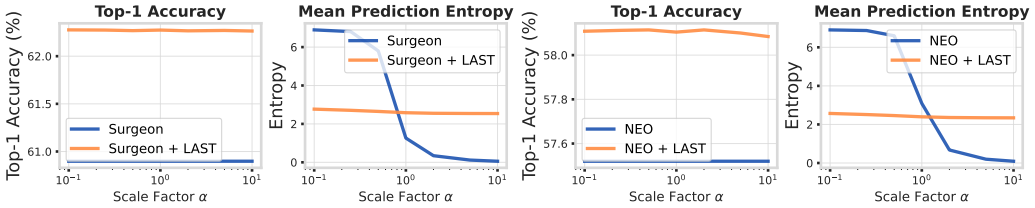


Figure 6: Sensitivity analysis of the momentum hyper-parameter ρ.



(a) BP-based TTA method (b) BP-free TTA methods

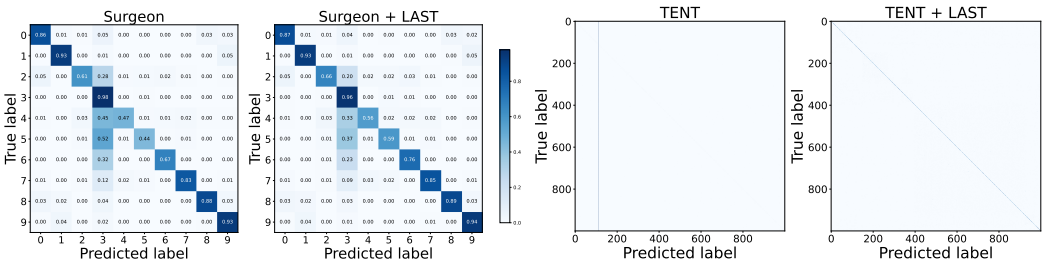
Figure 7: The influence of the logit scale factor α.

D.4 Confusion Matrix Visualization

To analyze the impact of LAST on prediction behavior, we visualize the confusion matrices before and after applying LAST in Figure 8. On CIFAR10-C, LAST reduces overconfident predictions and produces more balanced confident value across different classes. Moreover, on ImageNet-C, LAST alleviates the tendency of predictions to collapse toward dominant classes, resulting in more diverse and reliable predictions. These results demonstrate the effectiveness of LAST in improving the stability and robustness of TTA under distribution shifts.

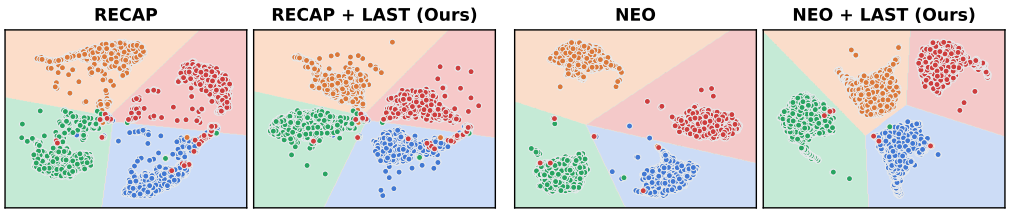
D.5 Decision Boundary Visualization

Figure 9 presents the t-SNE projections of the pre-`softmax` outputs on both CIFAR10-C and ImageNet-C. Existing TTA methods exhibit more dispersed logit distributions, with a considerable number of samples located near the decision boundaries, indicating uncertain predictions. In contrast, LAST produces more compact feature representations and effectively pulls ambiguous samples away from the decision boundaries toward their corresponding class clusters, resulting in improved class separability and more robust decision boundaries.

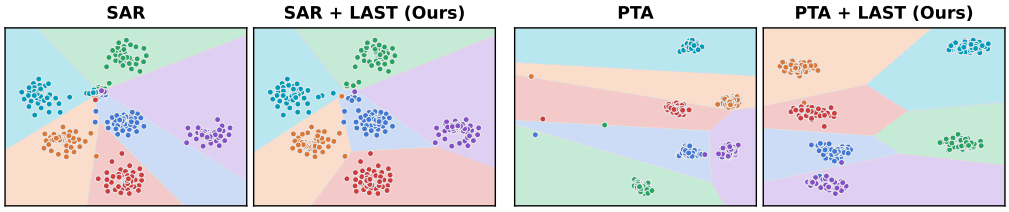


(a) Surgeon on CIFAR-10-C (Impulse Noise, Level 5) (b) TENT on ImageNet-C (Fog, Level 5)

Figure 8: Confusion matrices before and after integrating LAST.



(a) BP-based TTA on CIFAR10-C (Contrast, Level 5) (b) BP-free TTA on ImageNet-C (Bright., Level 5)



(c) BP-based TTA on ImageNet-C (Fog, Level 5) (d) BP-free TTA on ImageNet-C (Glass, Level 5)

Figure 9: t-SNE Visualization of decision boundaries. The colors indicate different classes.

E Additional Experimental Results on Logit Evolution

E.1 Episodic Logit Evolution

To further validate our analysis of logit dynamics, we provide additional results on episodic adaptation in Figure 10–Figure 12, with the corresponding Top-1 accuracy reported in Table 6. BP-based methods generally exhibit increasing logit norms with reduced entropy during adaptation, which is consistent with the scale amplification effect induced by entropy minimization, as discussed in Proposition 3.1. Since ADAPT adopts a transductive adaptation strategy, its episodic evolution is not included. For BP-free methods, we additionally report results on an additional corruption type due to the limited number of available baselines. These methods exhibit distinct logit norm evolution patterns, providing empirical evidence for the existence of logit norm drift during adaptation.

E.2 Continuous Logit Evolution

We further track the continuous logit evolution of TENT and DeYO across 15 corruption types, ranging from Gaussian Noise to JPEG. We compare the original baselines (blue), applying LAST only during forward inference (orange), and incorporating LAST into backpropagation (green). We choose TENT and DeYO because both methods suffer from model collapse during continuous adaptation, with the corresponding Top-1 accuracy reported in Table 2.

As shown in Figure 13, TENT exhibits similar behaviors among the three variants during the early corruption stages. However, as adaptation proceeds, the logit norm of the original TENT gradually increases, accompanied by deteriorated calibration. The model eventually collapses on the Snow corruption, where the logit norm increases to approximately 70, significantly deviating from the stable range maintained by LAST. By introducing LAST into BP, the logit magnitude is effectively regulated throughout adaptation, preventing excessive confidence amplification and avoiding model collapse. In contrast, applying LAST only during forward inference cannot recover the model after collapse, since the model parameters have already been updated toward a degraded state.

Similarly, Figure 14 shows the continuous logit evolution of DeYO. Although DeYO also exhibits an overall increasing trend in logit magnitude during adaptation, it encounters collapse under Zoom and Contrast corruptions. Interestingly, the collapse is accompanied by a local reduction of logit norm from approximately 60 to 50, suggesting that model degradation is not solely associated with logit explosion. Instead, abnormal deviations of logit statistics from a stable range can distort the confidence estimation and lead to unreliable predictions. With LAST incorporated into backpropagation, the logit norm remains consistently regulated across different corruptions, preventing unstable adaptation and improving robustness.

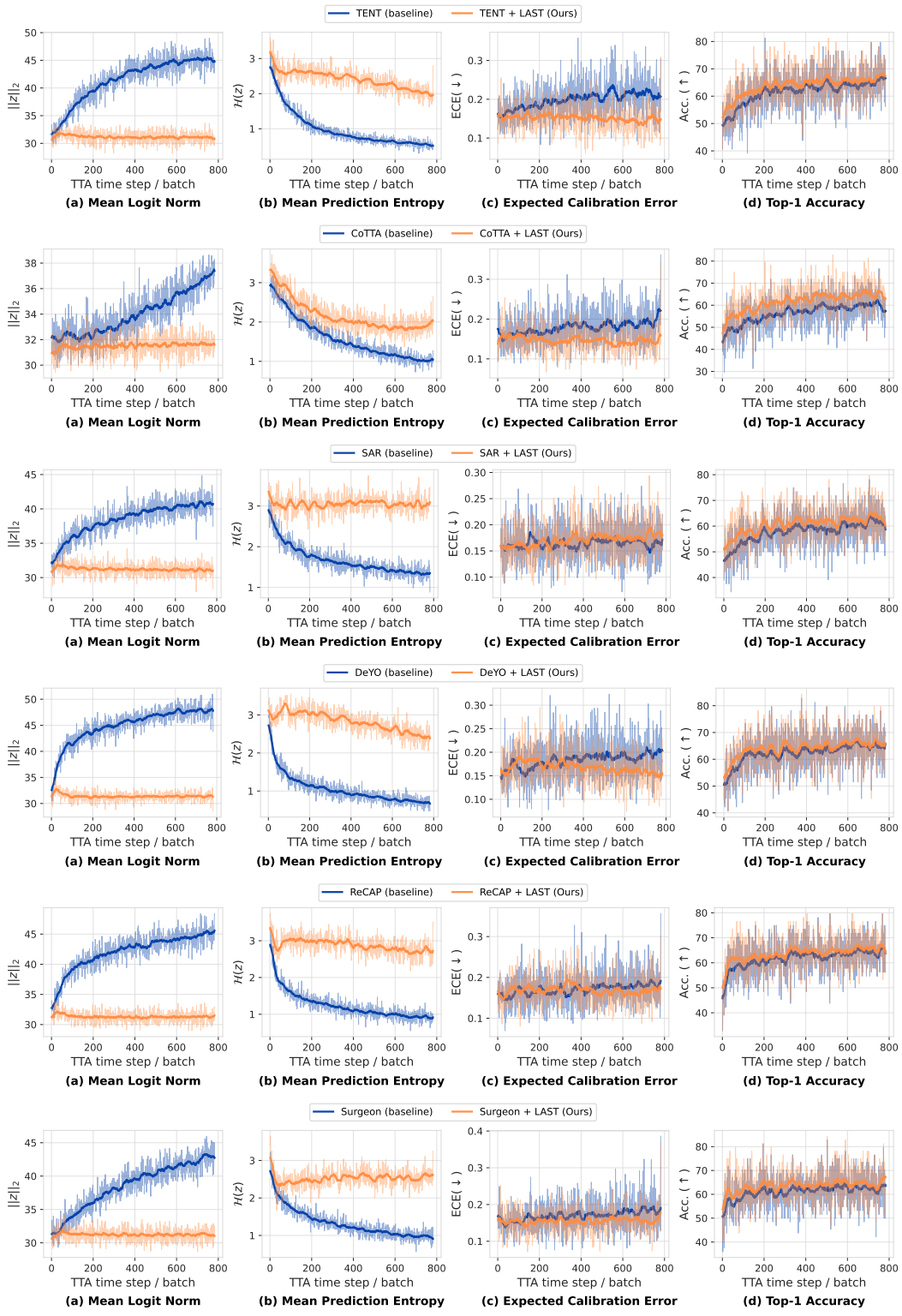


Figure 10: **Logit and prediction evolution of six BP-based TTA methods on ImageNet-C (Zoom Blur, Severity 5) with and without LAST.** LAST stabilizes logit evolution and improves calibration (lower ECE) and accuracy while moderating entropy minimization.

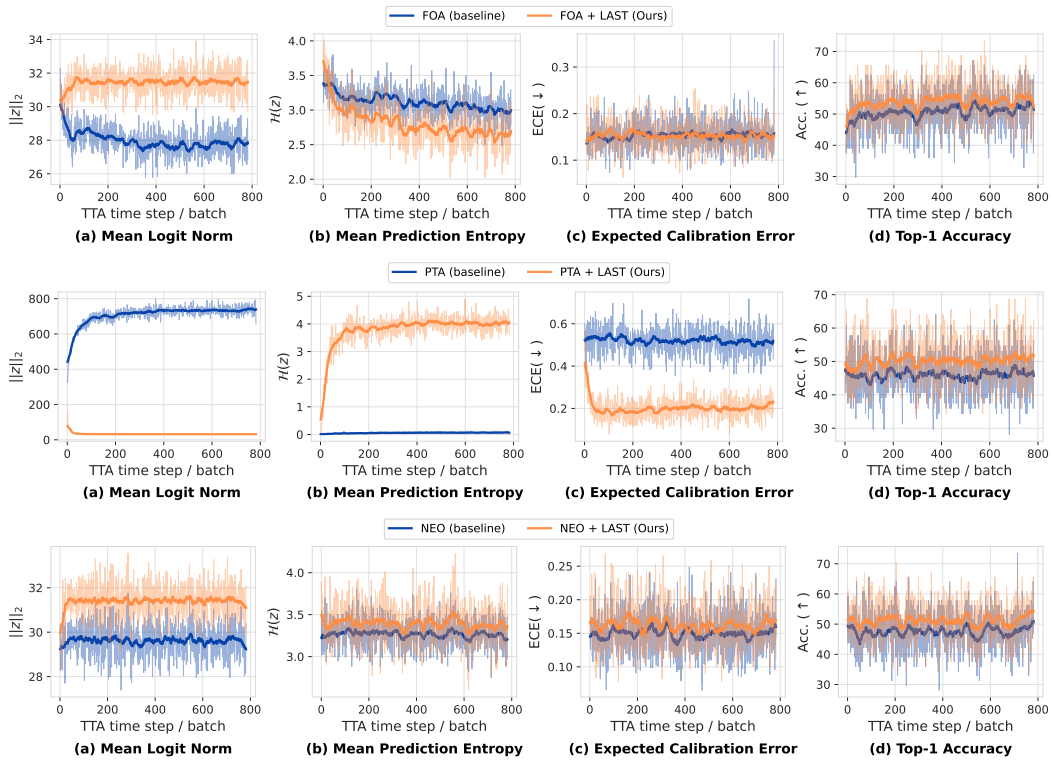


Figure 11: **Comparison of three BP-free TTA methods on ImageNet-C (Zoom Blur, Severity 5).**

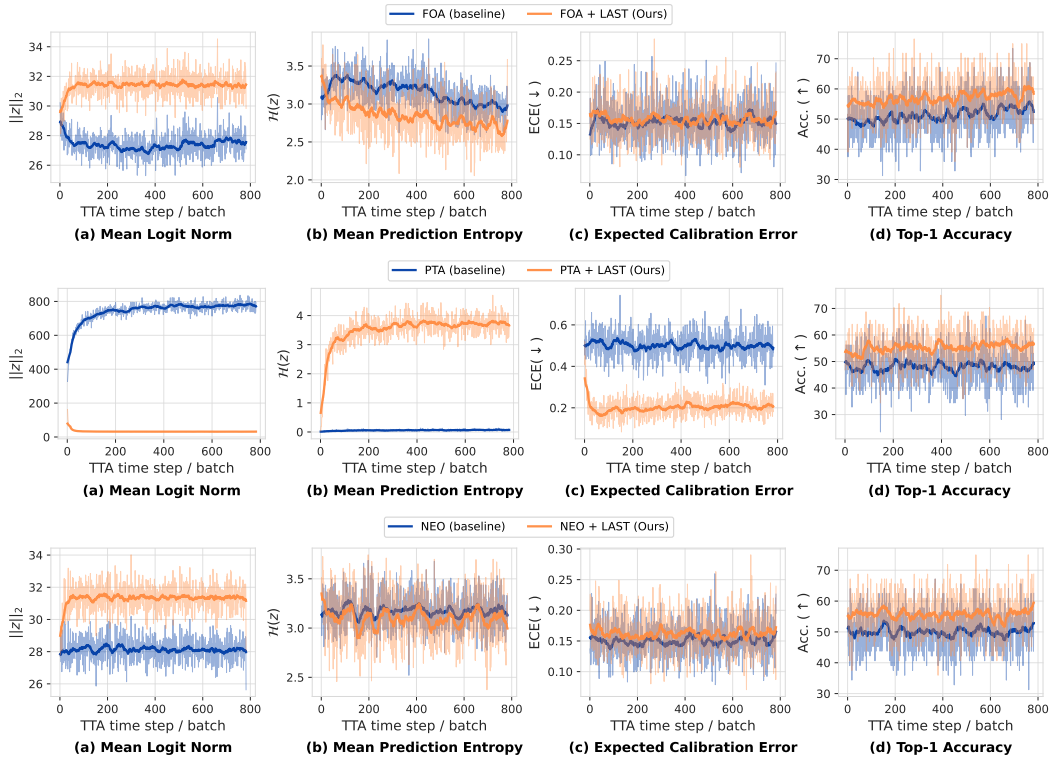


Figure 12: **Comparison of three BP-free TTA methods on ImageNet-C (Elastic Transform, Severity 5) with and without LAST.** LAST stabilizes logit evolution and improves calibration (lower ECE) and accuracy, while entropy remains largely unchanged.

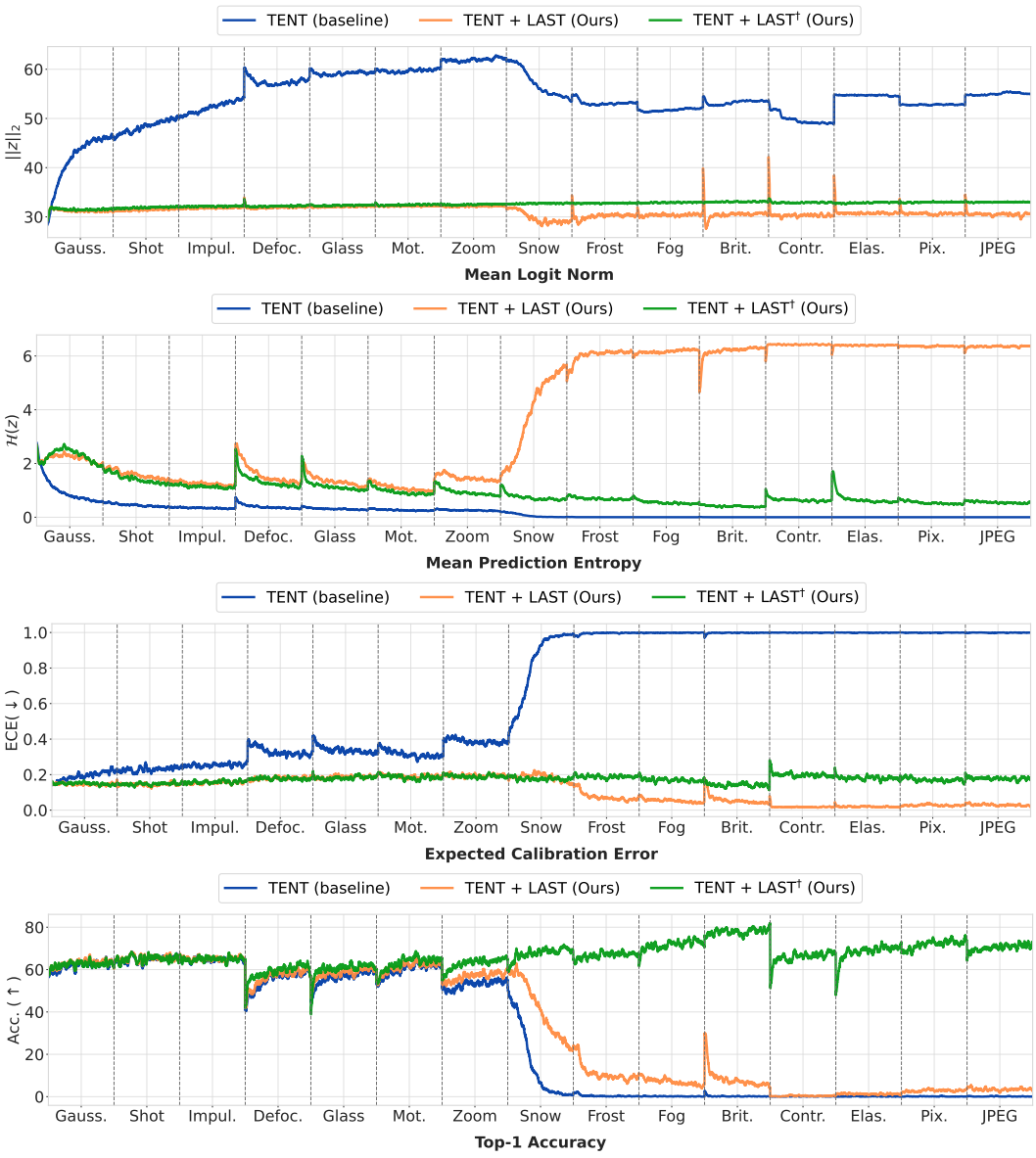


Figure 13: **Continuous logit evolution of TENT on ImageNet-C with and without LAST.** The original TENT suffers from increasing logit norms and calibration degradation during continuous adaptation, eventually leading to model collapse. LAST stabilizes logit magnitude throughout adaptation and prevents excessive confidence amplification, while applying LAST only during inference cannot recover the collapsed model.

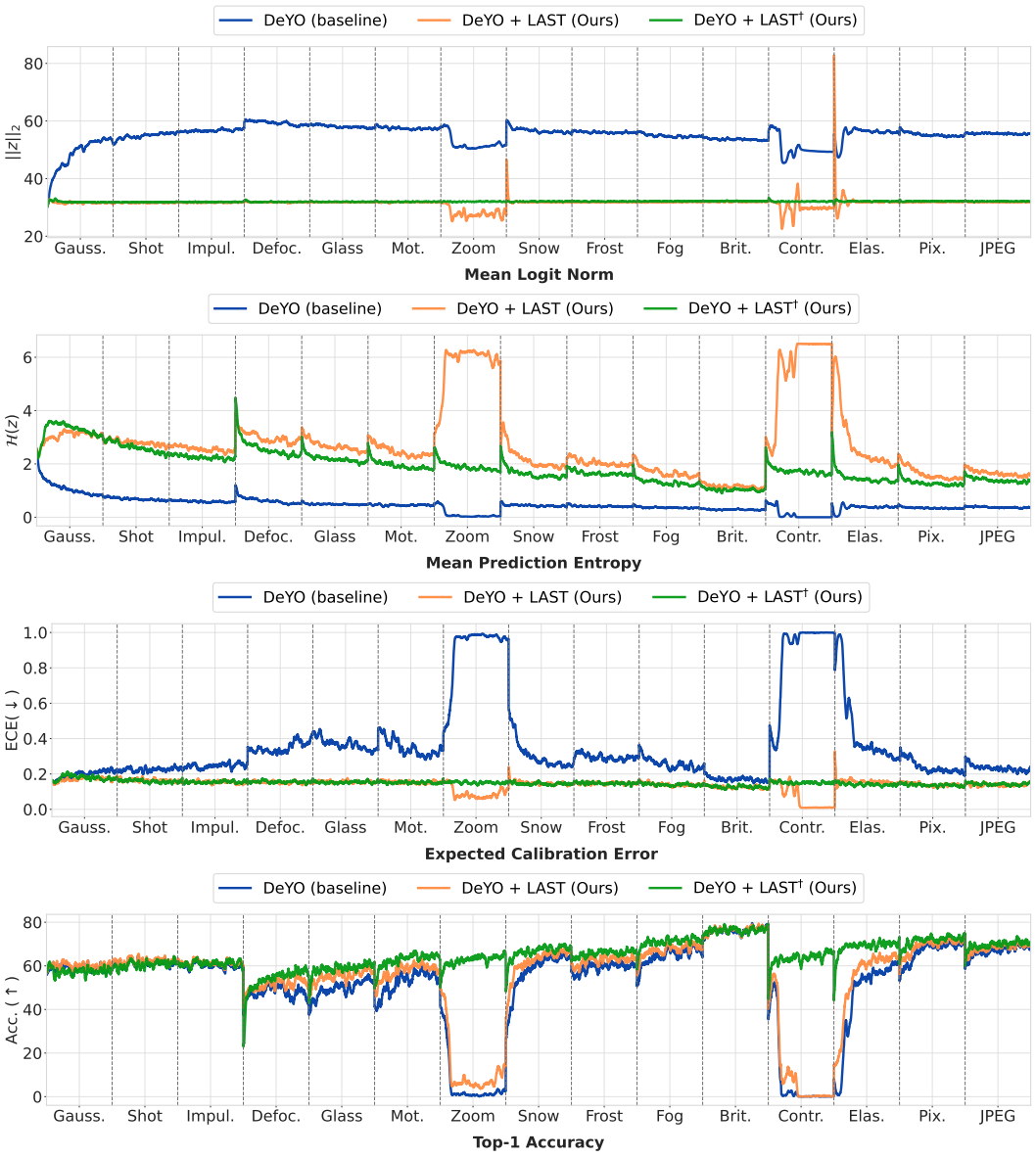


Figure 14: **Continuous logit evolution of DeYO on ImageNet-C with and without LAST.** DeYO exhibits unstable logit evolution across corruptions and suffers model collapse under Zoom and Contrast. Although its overall logit norm increases, local logit attenuation appears during collapse, indicating that abnormal logit dynamics rather than a single direction of change causes adaptation failure. LAST maintains stable logit statistics and improves adaptation robustness.