DECOMPOSING THE ENTROPY MINIMIZATION GRADIENT FOR CALIBRATED TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Entropy minimization (EM) is widely used for test-time adaptation (TTA), but its accuracy improvements under distribution shift often come at the cost of miscalibrated predictions. We study this behavior through the geometry of the EM gradient. After removing the softmax-invariant mean of the logits, we decompose the gradient into a radial component, which increases the centered-logit norm, and a tangential component, which changes relative class scores while preserving the norm to first order. Although these components are orthogonal in logit space, they induce conflicting parameter gradients, particularly in later network layers. Empirically, we find that the tangential component is the main driver of accuracy improvement, whereas the radial component is closely tied to confidence dynamics and becomes relatively stronger in later layers. These observations motivate two complementary interventions: Conflict-Aware Regularized Entropy Minimization (CAREM), which protects the tangential update by projecting conflicting radial gradients onto the tangential gradient's orthogonal complement, and freezing later layers during adaptation to improve calibration. CAREM modifies only the backward pass and introduces no additional hyperparameters or changes to the model architecture or adaptation objective. Across multiple entropy-based TTA methods, architectures, batch sizes, and distribution-shift benchmarks, CAREM improves classification performance, while its combination with late-layer freezing substantially improves calibration and retains strong accuracy. Our results show that accurate and calibrated entropy-based TTA can be achieved with simple changes to its optimization.

1 INTRODUCTION

Deep neural networks can become unreliable when test data differ from the distributions encountered during training, degrading both predictive performance and calibration (Hendrycks & Dietterich, 2019; Yang et al., 2022; Ovadia et al., 2019; Kirsch & Gal, 2022). This is particularly problematic for neural networks integrated into critical decision-making systems, such as self-driving cars (Bojarski et al., 2016) and healthcare (Jiang et al., 2012). In such systems, predictive accuracy alone is insufficient: confidence estimates should reflect the likelihood that a prediction is correct, so that uncertain outputs can appropriately trigger reliance on other sensors or human experts (Guo et al., 2017). Guo et al. (2017) showed that modern neural networks can be substantially miscalibrated even as their classification accuracy improves. Under distribution shift, calibration can deteriorate further (Ovadia et al., 2019; Kirsch & Gal, 2022).

Test-time adaptation (TTA) has gained interest as a possible solution to this problem. TTA aims to mitigate distribution shifts by updating the model with unlabeled samples at inference time (Liang et al., 2025). Many leading TTA approaches are based on entropy minimization (EM) (Wang et al., 2021; Niu et al., 2022; 2023; Lee et al., 2024; Zhang et al., 2025; Chen et al., 2026). EM optimizes the model to become more confident during inference, which has been shown to improve its accuracy (Wang et al., 2021). However, EM rewards every confident prediction, including confidently incorrect predictions and collapsed solutions, and has therefore been shown to harm the model's calibration, even when its accuracy improves (Yang et al., 2024; Tan et al., 2025). Consequently, a method intended to recover performance under distribution shifts can simultaneously make its errors more difficult to recognize.

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.

In this work, we study EM's accuracy-calibration tension through the geometry of its gradient. After removing the softmax-invariant mean of the logits, we decompose the logit gradient into two orthogonal components: a radial component parallel to the centered logits and a tangential component orthogonal to them. In logit space, the radial component increases the centered-logit norm while preserving class ordering, whereas the tangential component changes relative class scores while preserving the centered-logit norm to first order. Our experiments identify the tangential component as the principal driver of accuracy improvement, while the radial component is closely tied to confidence dynamics and is therefore central to calibration. The orthogonality of these components does not survive backpropagation through the network. We consistently observe large negative cosine similarities between their induced parameter gradients, especially in later layers, despite their exact orthogonality in logit space. The radial parameter gradients become particularly prominent in later layers, which we empirically identify as critical for calibration.

Motivated by these observations, we introduce *Conflict-Aware Regularized Entropy Minimization* (CAREM), a hyperparameter-free replacement for the backward pass of entropy-based TTA. CAREM treats the tangential component as the primary gradient and the radial gradient as auxiliary. When their parameter gradients conflict, CAREM projects the auxiliary gradient onto the orthogonal complement of the primary gradient; when they agree, both are retained unchanged. CAREM changes neither the adaptation objective nor the model architecture and can therefore be used with entropy-based TTA algorithms by replacing a standard backward pass. Our layerwise analysis further motivates freezing later layers during adaptation to control calibration. These two interventions play complementary roles: CAREM protects the tangential adaptation signal, while restricting late-layer adaptation substantially improves calibration. Together, they yield strong accuracy and calibration across distribution shifts, architectures, and TTA baselines.

Our contributions

- We decompose EM's gradient into radial and tangential components in centered-logit space and show that, despite their exact logit-space orthogonality, their induced parameter gradients become strongly conflicting after backpropagation, particularly in later layers.
- We identify the tangential component as the main driver of accuracy improvement, whereas the radial component becomes increasingly prominent in later layers and is closely associated with confidence dynamics.
- We use asymmetric gradient surgery to protect the tangential component (CAREM) and show that freezing late layers during adaptation is crucial for calibration. Their combination substantially improves accuracy and calibration across TTA methods and distribution shifts.

2 Related Work

Test-time adaptation updates a trained model using unlabeled samples available at inference time (Liang et al., 2025). **Entropy minimization** has become the dominant approach for TTA. It requires neither source data nor changes to source training. Tent established this paradigm by minimizing prediction entropy while updating only the affine parameters of normalization layers (Wang et al., 2021). Subsequent methods have primarily addressed the susceptibility of EM to model collapse in challenging scenarios. EATA filters high-entropy and redundant samples and introduces a Fisher regularizer for important parameters (Niu et al., 2022), while SAR combines sharpness-aware minimization (Foret et al., 2021) with model recovery to reduce collapse under wild TTA (Niu et al., 2023). DeYO argues that entropy alone is insufficient for identifying useful adaptation samples and incorporates additional sample selection and weighting based on the difference between predictions before and after applying an object-destructive transformation (Lee et al., 2024). Calibration-aware variants further show that reducing entropy can improve accuracy while producing overconfident predictions, and modify the EM objective to improve calibration (Yang et al., 2024; Tan et al., 2025). Most closely related to our work, COME replaces EM with an uncertainty-aware objective and stops gradients through the logit norm, thereby mitigating overconfidence (Zhang et al., 2025). ZeroSiam likewise aims to prevent logit norm inflation and convergence to a dominant class (Chen et al., 2026). It introduces a lightweight predictor and a stop-gradient target branch before the classifier, then aligns the entropy-minimizing online branch with the target branch to prevent collapsed solutions (Chen et al., 2026). Our work instead preserves the original model architecture, entropy

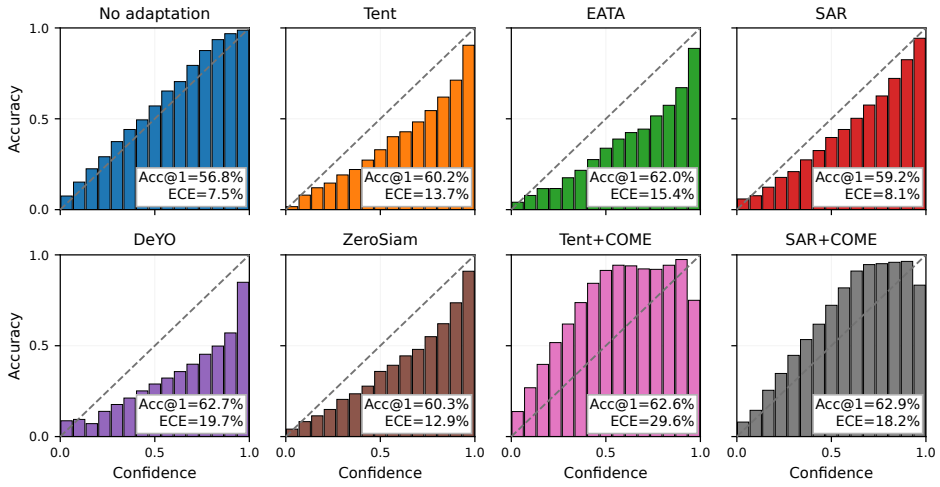


Figure 1: Calibration of entropy-based TTA methods on ImageNet-C Gaussian noise (severity 5), using ViT-Base and batch size 64. Top-1 accuracy and ECE (15 bins) are reported.

objective, and forward computation. We analyze EM's gradient geometry and use it to motivate complementary interventions for adaptation and calibration.

Gradient surgery has been studied extensively in multi-task learning, where several objectives are treated as jointly important. MGDA rescales the task gradients to converge towards a Pareto-optimal solution (Sener & Koltun, 2018), PCGrad projects mutually conflicting task gradients (Yu et al., 2020), and CAGrad balances average loss improvement against the worst local improvement among tasks (Liu et al., 2021). These methods seek a compromise across tasks rather than protecting a designated primary signal. Auxiliary-task learning considers the asymmetric setting that is closer to ours: improvement of an auxiliary objective is valuable only insofar as it benefits a main objective. Du et al. (2018) gate or weight an auxiliary gradient according to its cosine similarity with the main gradient. Bloop formulates auxiliary learning as minimizing an auxiliary objective over the solution set of the primary objective and combines the primary gradient with an orthogonalized auxiliary gradient and an exponential moving average to stabilize this projection under stochastic optimization (Hsieh et al., 2024). Our work differs from both settings because its two gradients do not arise from separate tasks or losses: they are orthogonal components of a single EM gradient in logit space that are respectively treated as primary and auxiliary.

3 Entropy Minimization's Miscalibration Problem

EM can initially improve accuracy by moving test samples into tighter feature clusters that remain close to the class structure learned from the training distribution (Press et al., 2024). With continued optimization, however, these test clusters can drift away from the training representations. The resulting benefit then saturates and accuracy may begin to decline (Press et al., 2024). EM continues to sharpen the model's predictions after this transition has occurred. Consequently, confidence can keep increasing after accuracy has stopped improving, causing the model to become increasingly overconfident and miscalibrated (Tan et al., 2025).

Figure 1 compares the accuracy and expected calibration error (ECE) (Naeini et al., 2015) of five entropy-based TTA methods: Tent (Wang et al., 2021), EATA (Niu et al., 2022), SAR (Niu et al., 2023), DeYO (Lee et al., 2024), and ZeroSiam (Chen et al., 2026). Although these methods differ in their parameter selection, sample filtering, optimization, and mechanisms for avoiding collapse, they all improve the original model's accuracy and degrade its calibration due to overconfidence in this setting. The best-performing methods, such as EATA (Niu et al., 2022) and DeYO (Lee et al., 2024), lead to the largest degradation in calibration. COME (Zhang et al., 2025), which explicitly constrains the logit-norm to remain constant, exhibits the opposite behavior and becomes underconfident despite accuracy improvements.

The reasons for these differences in calibration remain unclear. For example, the objectives alone do not explain why SAR is better calibrated than the other EM methods, or why COME becomes underconfident despite explicitly constraining the logit norm to remain constant. Importantly, COME can be reinterpreted as a manipulation of the EM gradient: stopping gradients through the logit norm suppresses its norm-inflating (radial) component in logit space. This points to gradient manipulation as a possible way to address EM's miscalibration. However, COME removes this component entirely and thereby trades overconfidence for underconfidence. This suggests that the two gradient components should not simply be retained or discarded in isolation, motivating an analysis of how they interact after backpropagation through the network.

4 Entropy-Minimization Geometry and Parameter-Space Conflict

We first study EM in logit space. The key idea is to isolate the norm-inflating (radial) component of the EM gradient, which reduces entropy without changing relative class ordering. We then analyze how the radial and tangential components change after backpropagation through the network. Although their geometric roles are clearly separated in logit space, their induced parameter gradients strongly interact, and their relative norms vary across network depth. These observations motivate both CAREM's gradient-surgery rule and our investigation of late-layer freezing for calibration.

4.1 Isolating the Norm-Inflating Entropy Gradient in Logit Space

Let $\boldsymbol{l} \in \mathbb{R}^K$ denote the logits for one input. Softmax probabilities are invariant to common shifts of the logits. We therefore use centered logits $\boldsymbol{z} = (\boldsymbol{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top)\boldsymbol{l}$. Consider a logit-space gradient step $\boldsymbol{z}^+ = \boldsymbol{z} - \eta\boldsymbol{g}$. Its squared norm is

$$\|\boldsymbol{z}^+\|_2^2 = \|\boldsymbol{z}\|_2^2 - 2\eta\boldsymbol{z}^\top\boldsymbol{g} + \eta^2\|\boldsymbol{g}\|_2^2. \tag{1}$$

We seek a decomposition $\boldsymbol{g} = \boldsymbol{g}_{\mathrm{rad}} + \boldsymbol{g}_{\tan}$ in which **the radial component produces only a magnitude change**. Thus, for sufficiently small $\eta > 0$, there must exist a scalar $\lambda(\eta) > 0$ such that

$$\boldsymbol{z} - \eta\boldsymbol{g}_{\mathrm{rad}} = \lambda(\eta)\boldsymbol{z} \quad \Longrightarrow \quad \boldsymbol{g}_{\mathrm{rad}} = \frac{1-\lambda(\eta)}{\eta}\boldsymbol{z} =: \alpha\boldsymbol{z}, \qquad \alpha \in \mathbb{R}. \tag{2}$$

Thus, the pure-magnitude constraint makes the radial component parallel to $\boldsymbol{z}$. We further constrain the remaining component to leave the centered logits' norm unchanged to first order

$$-2\eta\boldsymbol{z}^\top\boldsymbol{g}_{\tan} = 0 \quad \Longrightarrow \quad \boldsymbol{z}^\top\boldsymbol{g}_{\tan} = 0. \tag{3}$$

Substituting $\boldsymbol{g}_{\tan} = \boldsymbol{g} - \boldsymbol{g}_{\mathrm{rad}} = \boldsymbol{g} - \alpha\boldsymbol{z}$ into equation 3 gives

$$\boldsymbol{z}^\top(\boldsymbol{g} - \alpha\boldsymbol{z}) = 0 \quad \Longrightarrow \quad \alpha = \frac{\boldsymbol{z}^\top\boldsymbol{g}}{\|\boldsymbol{z}\|_2^2}. \tag{4}$$

For $\boldsymbol{z} \neq 0$, the unique decomposition is therefore

$$\boldsymbol{g}_{\mathrm{rad}} = \frac{\boldsymbol{z}^\top\boldsymbol{g}}{\|\boldsymbol{z}\|_2^2}\boldsymbol{z}, \qquad \boldsymbol{g}_{\tan} = \boldsymbol{g} - \boldsymbol{g}_{\mathrm{rad}}. \tag{5}$$

As $\boldsymbol{g}_{\mathrm{rad}}$ is parallel to $\boldsymbol{z}$ and $\boldsymbol{g}_{\tan}$ is orthogonal to $\boldsymbol{z}$, the two components are orthogonal. We further show in Appendix B that a gradient descent step along $\boldsymbol{g}_{\mathrm{rad}}$ in centered logit space always increases the centered-logit norm. $\boldsymbol{g}_{\mathrm{rad}}$ is therefore primarily responsible for the norm inflation during EM.

4.2 Empirical Conflict Between The Parameter Gradients

Let $\boldsymbol{z}_{\boldsymbol{\theta}}(\boldsymbol{x})$ be the centered logits of a network with parameters $\boldsymbol{\theta} \in \mathbb{R}^P$, and let $\boldsymbol{J} = \frac{\partial \boldsymbol{z}_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{K\times P}$ denote its Jacobian. Backpropagation pulls the two logit gradients into parameter space:

$$\boldsymbol{g}_{\boldsymbol{\theta},\mathrm{rad}} = \boldsymbol{J}^\top\boldsymbol{g}_{\mathrm{rad}}, \qquad \boldsymbol{g}_{\boldsymbol{\theta},\tan} = \boldsymbol{J}^\top\boldsymbol{g}_{\tan}. \tag{6}$$

Although the two logit gradients are orthogonal, their parameter-space inner product is $\boldsymbol{g}_{\boldsymbol{\theta},\mathrm{rad}}^\top\boldsymbol{g}_{\boldsymbol{\theta},\tan} = \boldsymbol{g}_{\mathrm{rad}}^\top\boldsymbol{J}\boldsymbol{J}^\top\boldsymbol{g}_{\tan}$, which is generally nonzero because $\boldsymbol{J}\boldsymbol{J}^\top$ need not preserve angles. The parameter gradients furthermore do not inherit the geometric behavior of their source logit gradients. A tangential parameter gradient can change the centered-logit norm at first order, since

$$\boldsymbol{z}_{\boldsymbol{\theta}-\eta\boldsymbol{g}_{\boldsymbol{\theta},\tan}} = \boldsymbol{z} - \eta\boldsymbol{J}\boldsymbol{J}^\top\boldsymbol{g}_{\tan} + \mathcal{O}(\eta^2). \tag{7}$$

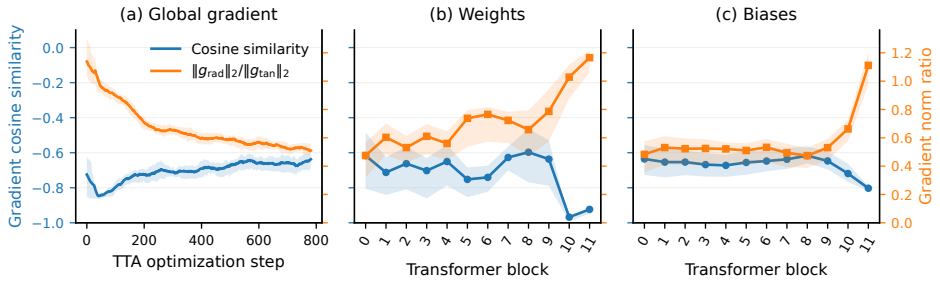


Figure 2: Cosine similarity and norm ratio of the radial and tangential parameter gradients during TTA. (a) Evolution of the global cosine similarity and gradient norm ratio. Curves are smoothed with a 50-step moving window. (b)-(c) Cosine similarity and norm ratio averaged over the adaptation run for the LayerNorm weights and biases, respectively, of each transformer block. The shaded areas represent the interquartile range (IQR). All experiments use a ViT-Base model on ImageNet-C with Gaussian noise at severity level 5 and a batch size of 64.

Expanding the resulting squared norm gives

$$\left\|\boldsymbol{z}_{\boldsymbol{\theta}-\eta\boldsymbol{g}_{\boldsymbol{\theta},\tan}}\right\|_2^2 = \left\|\boldsymbol{z}-\eta\boldsymbol{J}\boldsymbol{J}^\top\boldsymbol{g}_{\tan}\right\|_2^2 + \mathcal{O}(\eta^2) = \|\boldsymbol{z}\|_2^2 - 2\eta\boldsymbol{z}^\top\boldsymbol{J}\boldsymbol{J}^\top\boldsymbol{g}_{\tan} + \mathcal{O}(\eta^2). \tag{8}$$

Although $\boldsymbol{z}^\top\boldsymbol{g}_{\tan} = 0$ in logit-gradient space, this does not imply $\boldsymbol{z}^\top\boldsymbol{J}\boldsymbol{J}^\top\boldsymbol{g}_{\tan} = 0$. Thus, the term linear in η need not vanish, and a parameter update induced by the tangential logit gradient can change the centered-logit norm to first order, explaining why COME may become underconfident despite only retaining $\boldsymbol{g}_{\tan}$ during backpropagation (Zhang et al., 2025). Likewise, $\boldsymbol{J}\boldsymbol{J}^\top\boldsymbol{g}_{\mathrm{rad}}$ need not be parallel to $\boldsymbol{z}$, so a radial parameter update can change pairwise class ordering. As $\boldsymbol{J}\boldsymbol{J}^\top$ is positive semidefinite, however, the radial parameter update for a single sample is still guaranteed to increase the centered-logit norm to first order. We quantify the interaction of $\boldsymbol{g}_{\mathrm{rad}}$ and $\boldsymbol{g}_{\tan}$ after backpropagation using their cosine similarity and norm ratio

$$\rho = \frac{\boldsymbol{g}_{\boldsymbol{\theta},\mathrm{rad}}^\top\boldsymbol{g}_{\boldsymbol{\theta},\tan}}{\|\boldsymbol{g}_{\boldsymbol{\theta},\mathrm{rad}}\|_2\|\boldsymbol{g}_{\boldsymbol{\theta},\tan}\|_2} \qquad r = \frac{\|\boldsymbol{g}_{\boldsymbol{\theta},\mathrm{rad}}\|_2}{\|\boldsymbol{g}_{\boldsymbol{\theta},\tan}\|_2}. \tag{9}$$

As shown in Figure 2(a) for a ViT-Base model (Dosovitskiy et al., 2021; Wightman, 2019) on ImageNet-C with Gaussian noise at severity level 5 (Hendrycks & Dietterich, 2019; Deng et al., 2009), we empirically observe a strong negative global cosine similarity during EM with Tent (Wang et al., 2021), which quickly drops to -0.85 near the start of training before slightly increasing and ending at -0.64. Additionally, the radial gradient has a higher ℓ_2 norm at the beginning of training, but is quickly overtaken by the tangential gradient with a final norm ratio of 0.51. Figures 2(b) and (c) show that the cosine similarity remains strongly negative within each transformer block as well, with a sharp decrease for late blocks. The norm ratio similarly increases sharply for late blocks, indicating that the radial gradient is more prevalent in later layers than the tangential gradient. These observations are consistent with previous work that identifies the top layers of neural networks as primarily responsible for the "logit norm explosion" during EM (Mummadi et al., 2021; Choi et al., 2022), and may explain why SAR, which freezes the top layers during TTA (Niu et al., 2023), achieves the best ECE among TTA methods in Figure 1. This suggests that late-layer adaptation is particularly important for calibration. We observe analogous interactions between the radial and tangential parameter gradients in a ResNet-50-GN model (He et al., 2016; Wightman, 2019), as depicted in Appendix D.1. The radial and tangential gradients therefore hinder each other during optimization in parameter space, especially in later layers.

4.3 Empirical Roles of the Radial and Tangential Parameter Gradients

Negative gradient alignment alone does not reveal which component drives adaptation or how either component affects confidence after backpropagation. We therefore study their empirical roles by comparing Tent (Wang et al., 2021), which sums both parameter gradients, with tangential-only and radial-only variants in the same ViT-Base setting. We track the centered-logit norm and top-1 confidence throughout adaptation and report the final classification and calibration performance.

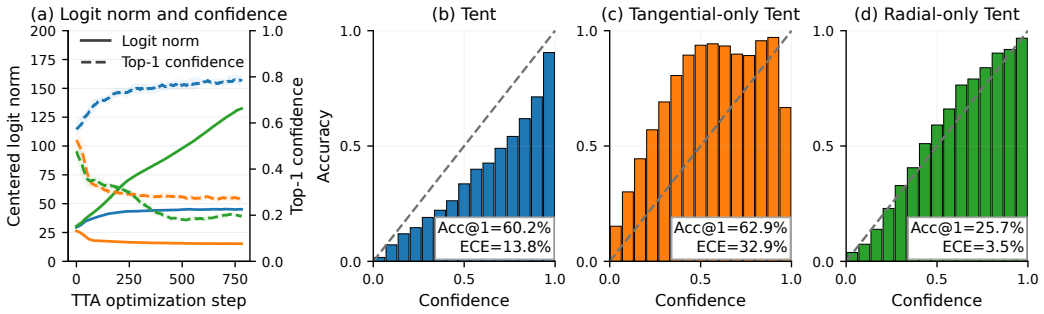


Figure 3: Roles of the radial and tangential parameter gradients during test-time adaptation. (a) Evolution of the centered-logit ℓ_2 norm (solid lines) and top-1 confidence (dashed lines) for standard Tent (blue), tangential-only Tent (orange), and radial-only Tent (green). Curves are smoothed with a 50-step moving window. The shaded area represents the interquartile range. (b)–(d) Reliability diagrams for the corresponding predictions; top-1 accuracy and expected calibration error (ECE) with 15 bins are reported in each panel. All experiments use a ViT-Base model on ImageNet-C with Gaussian noise at severity level 5 and a batch size of 64.

Figure 3(a) shows that standard Tent steadily increases the centered-logit norm and top-1 confidence. Tangential-only Tent instead decreases both quantities, yet achieves a final top-1 accuracy of 62.9%, compared with 60.2% for standard Tent and 56.8% without adaptation, as seen in Figures 1 and 3(b,c). This indicates that $\boldsymbol{g}_{\boldsymbol{\theta},\tan}$ provides an effective adaptation signal for improving classification under distribution shift. However, it produces strongly underconfident predictions with an ECE of 32.9% compared with 13.8% for standard Tent.

Radial-only Tent exhibits qualitatively different behavior. Its centered-logit norm grows rapidly and far exceeds that of standard Tent, but its top-1 confidence decreases nevertheless. Its classification accuracy also falls to 25.7%. This behavior illustrates the distinction established in Section 4.2: although g_{rad} corresponds to a pure rescaling direction in logit space, the induced parameter gradient $g_{\theta,\mathrm{rad}}$ need not preserve class ordering or monotonically increase confidence after backpropagation. The low ECE of 3.5% should therefore not be interpreted as successful adaptation; rather, confidence decreases together with a substantial loss in accuracy.

Taken together, these results suggest an asymmetric role for the two parameter gradients. The tangential parameter gradient provides the stronger standalone signal for improving classification, whereas the radial parameter gradient is not useful as an independent adaptation direction. At the same time, standard Tent is substantially better calibrated than tangential-only Tent, indicating that simply discarding the radial component is also undesirable. We therefore treat $g_{\theta,\tan}$ as the primary adaptation gradient and $g_{\theta,\mathrm{rad}}$ as an auxiliary gradient whose contribution is retained only when it does not oppose the primary direction. This motivates an asymmetric gradient-surgery rule for CAREM, introduced in the next section.

4.4 CAREM: Minimum-Change Auxiliary Projection

Let $\boldsymbol{g}_{\mathrm{main}} = \boldsymbol{g}_{\boldsymbol{\theta},\tan}$ and $\boldsymbol{g}_{\mathrm{aux}} = \boldsymbol{g}_{\boldsymbol{\theta},\mathrm{rad}}$. We seek a combined gradient $\boldsymbol{g}_{\mathrm{comb}} = \boldsymbol{g}_{\mathrm{main}} + \widetilde{\boldsymbol{g}}_{\mathrm{aux}}$ that preserves the progress associated with $\boldsymbol{g}_{\mathrm{main}}$ while retaining as much of $\boldsymbol{g}_{\mathrm{aux}}$ as possible. The inner product $\boldsymbol{g}_{\mathrm{main}}^{\top}\boldsymbol{g}_{\mathrm{comb}}$ measures the first-order progress of $\boldsymbol{g}_{\mathrm{main}}$. Using $\boldsymbol{g}_{\mathrm{main}}$ alone gives $\|\boldsymbol{g}_{\mathrm{main}}\|_2^2$. Therefore, requiring the auxiliary contribution not to reduce this progress gives

$$\boldsymbol{g}_{\mathrm{main}}^{\top}\boldsymbol{g}_{\mathrm{comb}} \geq \|\boldsymbol{g}_{\mathrm{main}}\|_2^2 \quad \Longleftrightarrow \quad \boldsymbol{g}_{\mathrm{main}}^{\top}\widetilde{\boldsymbol{g}}_{\mathrm{aux}} \geq 0. \tag{10}$$

Among all $\widetilde{\boldsymbol{g}}_{\mathrm{aux}}$ satisfying this constraint, CAREM selects the one closest to $\boldsymbol{g}_{\mathrm{aux}}$:

$$\widetilde{\boldsymbol{g}}_{\mathrm{aux}} = \arg\min_{\boldsymbol{g}'} \|\boldsymbol{g}' - \boldsymbol{g}_{\mathrm{aux}}\|_2^2 \quad \text{s.t.} \quad \boldsymbol{g}_{\mathrm{main}}^{\top}\boldsymbol{g}' \geq 0, \tag{11}$$

resulting in the following closed-form solution:

$$\widetilde{\boldsymbol{g}}_{\mathrm{aux}} = \begin{cases} \boldsymbol{g}_{\mathrm{aux}} - \dfrac{\boldsymbol{g}_{\mathrm{main}}^{\top}\boldsymbol{g}_{\mathrm{aux}}}{\|\boldsymbol{g}_{\mathrm{main}}\|_2^2}\boldsymbol{g}_{\mathrm{main}}, & \boldsymbol{g}_{\mathrm{main}}^{\top}\boldsymbol{g}_{\mathrm{aux}} < 0, \\ \boldsymbol{g}_{\mathrm{aux}}, & \boldsymbol{g}_{\mathrm{main}}^{\top}\boldsymbol{g}_{\mathrm{aux}} \geq 0, \end{cases} \qquad \boldsymbol{g}_{\mathrm{comb}} = \boldsymbol{g}_{\mathrm{main}} + \widetilde{\boldsymbol{g}}_{\mathrm{aux}}. \tag{12}$$

The resulting one-sided PCGrad-style rule (Yu et al., 2020) is asymmetric and has no hyperparameters. In practice, the logit-space decomposition is performed independently for each sample and CAREM is applied globally to the raw, flattened, and batch-aggregated parameter gradients. The resulting combined gradient is then passed to the optimizer. CAREM is applied to every gradient that flows through the logits; gradients that do not, such as EATA's Fisher regularization (Niu et al., 2022), are left unchanged. The calculation of $\boldsymbol{g}_{\boldsymbol{\theta},\tan}$ and $\boldsymbol{g}_{\boldsymbol{\theta},\mathrm{rad}}$ requires two backward passes. Therefore, CAREM doubles the number of backward passes for each TTA method.

5 Experiments

We design our experiments to answer the following questions: (Q1) Does CAREM improve accuracy and calibration under standard online TTA? (Q2) How does freezing the top/later layers (FTL) affect accuracy and calibration? (Q3) Does the combination of CAREM and FTL improve both metrics? (Q4) Do these gains persist under the imbalanced and mixed shifts introduced by SAR (Niu et al., 2023) and with varying batch sizes?

5.1 Experimental Setup

Datasets, evaluation scenarios, and metrics. We evaluate on ImageNet-C at severity level 5 in three settings (Hendrycks & Dietterich, 2019; Deng et al., 2009). In *standard* TTA, each corruption is processed as a separate online stream and the model is reset between corruptions. Following SAR (Niu et al., 2023), *imbalanced label shift* orders samples by class ($q_{\max}/q_{\min} = \infty$), while *mixed shift* combines all corruptions into one stream without resetting the model. We report top-1 classification accuracy and ECE with 15 equal-width bins (Naeini et al., 2015). Every entry is the mean $\pm$ standard deviation over four runs with different random seeds. Additional results on ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), and VisDA-2021 (Bashkirova et al., 2022) are in Appendix D.3.

Compared methods and implementation details. Our main backbone is an ImageNet-pretrained ViT-Base with LayerNorm (Dosovitskiy et al., 2021; Wightman, 2019)[1]. ResNet-50-GN results are in Appendix D.2. We use Tent (Wang et al., 2021), EATA (Niu et al., 2022), SAR (Niu et al., 2023), DeYO (Lee et al., 2024), and ZeroSiam (Chen et al., 2026) as EM baseline methods. We additionally include EATA-C (Tan et al., 2025) and CETA (Yang et al., 2024), two EM approaches aimed at improving calibration, as direct comparisons to CAREM with respect to ECE. Both approaches introduce additional terms to the EM objective to improve calibration. We also report COME (Zhang et al., 2025) and FTL as baselines. Following SAR (Niu et al., 2023), FTL freezes ViT blocks 9-11 for ViT-Base and ResNet layer 4. We adapt the affine parameters of the normalization layers with SGD (momentum 0.9), batch size 64, and learning rates 0.001 (ViT-Base) and 0.00025 (ResNet-50-GN). EATA-C follows its original configuration and uses learning rates of 0.1 and 0.005, respectively (Tan et al., 2025). Additional experimental details can be found in Appendix C.

5.2 Standard Corruption Shifts

Table 1 shows that CAREM improves the average accuracy and ECE of Tent, EATA, SAR, and ZeroSiam, with ZeroSiam+CAREM achieving the joint-highest average accuracy of 67.5%. CAREM improves DeYO's accuracy on 11 of the 15 corruptions. However, DeYO is unstable on Fog and Contrast, and CAREM worsens its collapse behavior, leading to a lower average accuracy despite the 11 improvements. FTL results in a lower ECE than the baseline and CAREM for all TTA methods. However, FTL decreases the average accuracy of EATA, DeYO, and ZeroSiam.

FTL and CAREM complement each other well, as their combination achieves a lower ECE than FTL alone for all TTA methods. It also improves CAREM's accuracy for Tent, SAR, and DeYO. EATA, SAR, and ZeroSiam attain the three lowest average ECEs (3.2%, 3.5%, 4.2%) in combination with CAREM+FTL, with accuracies very close to 67.5% (65.9%, 66.9%, 67.4%). EATA-C has the fourth-lowest ECE of 4.4%, albeit with an accuracy of 61.0%. In contrast, COME generally raises accuracy but substantially increases ECE.

[1]We use the correct normalization for ViT-Base, with mean and standard deviation equal to $(0.5, 0.5, 0.5)$. EATA, SAR, DeYO, and ZeroSiam used the standard ImageNet normalization, yielding different results.

Table 1: Standard online TTA results on ImageNet-C at severity level 5 with ViT-Base. The best result in each column is bold, and the best result within each method family is underlined.

	Noise			Blur				Weather				Digital				Avg.	
Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elast.	Pixel	JPEG	Acc.↑	ECE↓
No Adapt	56.8±0.0	56.8±0.0	57.5±0.0	46.9±0.0	35.6±0.0	53.1±0.0	44.8±0.0	62.2±0.0	62.5±0.0	65.7±0.0	77.7±0.0	32.6±0.0	46.0±0.0	67.0±0.0	67.6±0.0	55.5±0.0	10.5±0.0
Tent	60.3±0.1	61.6±0.1	61.8±0.1	59.0±0.1	56.5±0.0	63.5±0.1	59.2±0.1	56.9±2.0	64.4±0.5	19.8±33.9	79.2±0.0	67.3±0.1	61.2±0.2	72.8±0.1	70.6±0.1	60.9±2.3	16.9±2.9
Tent+COME	62.5±0.1	63.8±0.1	64.2±0.1	60.1±0.1	58.1±0.0	64.8±0.1	61.2±0.1	69.4±0.1	66.5±0.1	72.8±0.3	80.4±0.0	68.9±0.1	65.4±0.1	75.0±0.1	73.1±0.1	67.1±0.0	31.0±0.0
Tent+FTL	59.2±0.1	60.6±0.0	60.8±0.1	57.7±0.1	56.1±0.1	62.1±0.2	58.2±0.1	57.4±1.3	63.7±0.4	36.2±38.5	78.9±0.0	65.3±0.1	62.4±0.1	72.6±0.1	70.5±0.0	61.5±2.5	10.1±3.3
Tent+CAREM	63.9±0.1	65.2±0.1	65.2±0.1	61.8±0.1	61.4±0.1	66.7±0.1	63.9±0.1	71.5±0.1	68.7±0.3	20.2±36.1	81.1±0.1	69.9±0.2	70.5±0.0	76.4±0.0	74.7±0.1	65.4±2.4	13.9±2.9
Tent+CAREM+FTL	63.7±0.1	65.2±0.0	65.1±0.1	62.1±0.1	61.8±0.1	67.0±0.1	64.5±0.1	71.7±0.1	68.9±0.2	38.3±41.6	80.9±0.1	**70.2**±0.1	70.9±0.0	76.3±0.1	74.6±0.1	66.8±2.8	6.4±3.5
EATA	62.0±0.1	63.2±0.2	62.8±0.9	60.7±0.2	61.0±0.2	66.0±0.0	63.8±0.0	70.5±0.1	68.4±0.2	55.2±35.7	79.9±0.1	33.6±38.1	69.5±0.8	75.2±0.1	73.3±0.2	64.3±2.2	12.7±0.4
EATA-C	63.6±0.1	65.2±0.2	64.9±0.1	57.6±0.3	61.3±0.1	66.8±0.1	65.8±0.1	73.1±0.2	**71.6**±0.2	19.0±35.9	81.1±0.1	1.5±1.2	72.6±0.2	76.9±0.3	74.7±0.0	61.0±2.3	4.4±0.7
EATA+COME	63.3±0.1	64.7±0.1	64.8±0.0	61.9±0.1	62.5±0.1	67.0±0.1	64.8±0.0	71.8±0.1	69.1±0.1	74.2±0.4	80.8±0.0	45.5±30.8	71.4±0.1	76.1±0.0	74.1±0.0	**67.5**±2.1	18.6±0.8
EATA+FTL	58.9±0.4	60.6±0.1	60.5±0.3	57.7±0.4	58.1±0.1	63.3±0.2	61.2±0.1	68.0±0.1	65.0±0.1	67.6±6.2	78.7±0.0	16.5±32.0	67.0±0.2	73.0±0.2	71.3±0.1	61.8±1.8	9.5±0.2
EATA+CAREM	**64.4**±0.2	**66.2**±0.0	**65.8**±0.1	62.6±0.1	**64.4**±0.1	68.8±0.0	67.5±0.2	**73.7**±0.0	**71.6**±0.1	**76.0**±0.2	**81.3**±0.0	15.9±30.4	**74.2**±0.1	**77.4**±0.1	**75.4**±0.0	67.0±2.0	11.2±0.5
EATA+CAREM+FTL	64.3±0.1	66.1±0.1	**65.8**±0.0	**63.2**±0.2	64.2±0.0	**69.1**±0.1	**67.8**±0.0	73.5±0.1	71.2±0.0	75.7±0.1	81.0±0.1	0.6±0.2	73.7±0.1	77.2±0.1	75.0±0.0	65.9±0.0	**3.2**±0.1
CETA	60.3±0.0	61.6±0.2	61.8±0.1	59.0±0.1	56.6±0.1	63.6±0.1	59.5±0.1	57.7±1.8	64.5±0.3	2.1±0.1	79.3±0.1	66.2±0.1	62.1±0.4	73.0±0.1	70.8±0.0	59.9±0.1	18.9±0.1
SAR	59.2±0.0	60.7±0.1	60.9±0.0	57.7±0.1	56.3±0.1	62.2±0.1	58.2±0.1	58.6±0.3	64.0±0.1	68.3±2.5	78.9±0.0	64.9±0.3	62.5±0.1	72.6±0.1	70.5±0.0	63.7±0.2	7.2±0.0
SAR+COME	63.0±0.1	64.3±0.0	64.5±0.1	60.7±0.1	59.7±0.1	65.7±0.1	62.4±0.1	70.2±0.2	67.3±0.1	55.8±33.9	80.6±0.0	68.8±0.1	67.7±0.0	75.4±0.1	73.4±0.1	66.6±2.3	21.7±1.6
SAR+CAREM	63.9±0.1	65.3±0.0	65.3±0.1	62.2±0.1	62.0±0.1	67.2±0.1	64.6±0.1	71.8±0.1	69.0±0.2	74.6±1.0	81.0±0.0	34.2±39.0	71.0±0.1	76.4±0.0	74.7±0.0	66.9±2.6	3.5±0.2
DeYO	62.7±0.2	64.2±0.1	64.0±0.2	60.2±0.4	61.2±0.1	66.5±0.1	63.4±0.5	71.4±0.1	69.6±0.2	54.8±35.3	80.6±0.1	32.0±36.6	71.4±0.1	76.4±0.1	74.2±0.1	64.8±2.2	20.3±2.4
DeYO+COME	63.5±0.0	64.8±0.1	64.8±0.1	60.5±0.2	61.4±0.3	66.7±0.3	63.5±0.5	71.7±0.0	69.1±0.2	73.5±1.1	80.7±0.0	34.6±39.6	70.9±0.3	76.3±0.1	74.2±0.1	66.4±2.7	34.0±1.0
DeYO+FTL	59.4±0.3	61.2±0.3	60.7±0.4	57.9±0.4	58.9±0.3	64.6±0.2	62.1±0.1	68.9±0.1	66.5±0.4	54.4±35.1	78.9±0.1	32.4±36.9	69.3±0.5	74.3±0.0	72.0±0.1	62.8±4.3	15.5±5.3
DeYO+CAREM	62.7±0.1	64.6±0.1	64.3±0.3	59.4±0.4	62.0±0.1	66.6±0.2	64.5±0.7	72.2±0.3	70.3±0.3	37.1±41.8	80.8±0.1	0.3±0.1	73.2±0.1	76.7±0.2	74.8±0.1	62.0±2.8	22.6±4.1
DeYO+CAREM+FTL	63.8±0.1	65.6±0.1	65.1±0.1	62.6±0.0	63.8±0.2	68.8±0.1	67.6±0.1	73.2±0.1	71.0±0.1	37.8±43.1	80.8±0.0	17.2±33.9	73.9±0.1	77.1±0.1	74.9±0.1	64.2±2.4	10.4±4.5
ZeroSiam	60.3±0.1	61.4±0.1	61.7±0.0	58.8±0.1	56.4±0.1	63.2±0.1	59.3±0.2	67.5±0.0	65.0±0.5	70.0±0.5	79.0±0.0	67.4±0.3	62.7±0.1	72.4±0.1	70.2±0.0	65.0±0.0	11.2±0.0
ZeroSiam+COME	62.5±0.1	63.7±0.1	64.1±0.0	59.9±0.1	58.0±0.1	64.6±0.0	60.4±0.1	68.7±0.0	66.5±0.2	72.0±0.8	80.2±0.0	68.8±0.1	63.3±0.2	74.8±0.1	72.7±0.0	66.7±0.1	31.9±0.0
ZeroSiam+FTL	59.5±0.1	60.7±0.0	61.0±0.1	57.8±0.1	56.2±0.2	62.2±0.1	58.5±0.2	67.1±0.1	64.3±0.3	69.3±0.3	78.8±0.1	65.7±0.1	63.0±0.1	72.4±0.1	70.2±0.1	64.5±0.0	5.8±0.0
ZeroSiam+CAREM	62.0±0.1	63.4±0.1	63.6±0.2	60.8±0.1	59.4±0.1	65.3±0.1	62.1±0.1	70.0±0.0	67.3±0.1	73.5±0.3	80.4±0.1	69.6±0.1	67.8±0.1	75.2±0.1	72.3±0.0	**67.5**±0.0	10.0±0.1
ZeroSiam+CAREM+FTL	61.6±0.1	63.0±0.1	63.2±0.2	60.6±0.1	59.8±0.1	65.1±0.1	62.1±0.1	70.0±0.0	67.2±0.1	73.4±0.1	80.2±0.0	70.0±0.1	68.3±0.1	75.0±0.1	71.8±0.1	67.4±0.0	4.2±0.0

Table 2: Online imbalanced label-shift results on ImageNet-C at severity level 5 with ViT-Base. The best result in each column is bold, and the best result within each method family is underlined.

	Noise			Blur				Weather				Digital				Avg.	
Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elast.	Pixel	JPEG	Acc.↑	ECE↓
No Adapt	56.8±0.0	56.8±0.0	57.5±0.0	46.9±0.0	35.6±0.0	53.1±0.0	44.8±0.0	62.2±0.0	62.5±0.0	65.7±0.0	77.7±0.0	32.6±0.0	46.0±0.0	67.0±0.0	67.6±0.0	55.5±0.0	10.5±0.0
Tent	60.3±0.0	61.5±0.0	61.7±0.0	58.8±0.0	56.4±0.0	63.3±0.0	58.5±0.0	64.5±0.1	64.6±0.0	2.7±0.0	79.2±0.0	66.5±0.1	62.0±0.1	72.4±0.0	70.7±0.0	60.2±0.0	17.7±0.0
Tent+COME	62.7±0.0	64.0±0.0	64.5±0.0	60.5±0.0	58.7±0.0	65.2±0.0	61.5±0.0	70.1±0.0	67.1±0.0	73.4±0.0	80.4±0.0	69.0±0.0	66.9±0.0	75.1±0.0	73.2±0.0	67.5±0.0	30.8±0.0
Tent+FTL	59.3±0.0	60.6±0.0	60.8±0.0	57.6±0.0	56.0±0.0	62.2±0.0	57.7±0.0	62.8±0.1	63.7±0.1	19.0±32.0	78.8±0.0	65.1±0.1	62.6±0.3	72.4±0.0	70.5±0.0	60.6±2.1	11.2±2.9
Tent+CAREM	**63.3**±0.0	64.6±0.0	**64.8**±0.0	61.2±0.0	60.9±0.0	66.3±0.0	63.1±0.0	71.2±0.0	68.6±0.0	37.9±41.2	**80.8**±0.0	69.4±0.1	69.6±0.0	76.0±0.0	74.2±0.0	66.1±2.7	12.4±3.4
Tent+CAREM+FTL	63.2±0.0	64.6±0.0	64.7±0.0	**61.6**±0.0	61.3±0.0	66.7±0.0	63.8±0.0	71.4±0.0	68.6±0.1	73.5±0.1	80.6±0.0	**69.6**±0.1	70.8±0.0	76.0±0.0	74.2±0.0	68.7±0.0	3.4±0.0
EATA	58.5±1.5	60.3±0.3	58.7±1.6	54.9±1.6	56.2±0.7	62.3±0.2	59.2±0.9	69.2±0.1	65.2±0.7	69.9±0.3	77.9±1.2	11.3±8.1	50.6±23.4	73.5±1.0	71.8±0.6	60.0±1.7	15.5±1.7
EATA-C	60.0±0.6	47.6±29.5	61.8±0.9	37.8±24.0	39.8±23.4	53.5±17.5	59.8±1.8	69.8±0.8	57.1±23.3	18.9±33.9	80.0±0.1	4.8±7.8	70.8±0.4	75.4±0.5	56.6±33.6	52.9±7.1	5.1±2.5
EATA+COME	62.6±0.1	64.1±0.1	64.0±0.1	44.7±27.9	60.1±0.3	65.2±0.3	63.2±0.2	71.0±0.1	65.9±4.1	73.6±0.1	80.6±0.0	57.5±4.0	70.8±0.2	75.7±0.0	73.6±0.1	66.2±2.2	19.2±1.2
EATA+FTL	57.3±0.3	57.7±1.4	58.0±0.4	52.6±1.4	43.4±22.4	60.9±0.9	58.3±0.2	66.8±0.1	61.9±0.3	68.7±0.7	78.2±0.1	11.1±18.8	66.7±0.3	72.0±0.3	70.2±0.5	58.9±2.2	11.7±0.5
EATA+CAREM	61.0±0.2	62.7±0.1	62.9±0.4	54.1±1.5	58.5±2.3	64.1±1.0	62.6±0.6	71.4±0.7	69.2±0.6	70.8±2.3	79.7±1.5	14.4±23.4	72.6±0.1	76.2±0.2	74.3±0.0	63.6±1.7	14.3±1.7
EATA+CAREM+FTL	62.2±0.0	64.0±0.1	63.8±0.1	58.5±0.8	61.0±0.2	66.4±0.4	65.3±0.4	72.0±0.1	69.5±0.2	62.7±20.8	80.1±0.1	19.6±27.7	72.6±0.1	75.9±0.1	74.1±0.1	64.5±2.7	5.8±2.1
CETA	60.6±0.1	61.6±0.0	61.7±0.0	58.8±0.0	56.7±0.1	63.5±0.0	59.1±0.1	65.8±0.3	64.5±0.2	3.0±0.1	79.2±0.0	65.8±0.1	63.0±0.0	72.5±0.1	70.8±0.0	60.4±0.0	18.1±0.0
SAR	59.3±0.0	60.6±0.0	60.8±0.0	57.6±0.0	56.1±0.0	62.1±0.0	57.6±0.0	62.4±0.1	64.1±0.1	69.7±0.1	78.8±0.0	65.1±0.1	62.7±0.2	72.3±0.0	70.5±0.0	64.0±0.0	6.8±0.0
SAR+COME	62.8±0.0	64.2±0.0	64.3±0.0	60.6±0.0	59.8±0.0	65.7±0.0	62.3±0.0	70.7±0.0	67.4±0.0	73.1±0.1	80.5±0.0	68.5±0.0	68.6±0.0	75.3±0.0	73.5±0.0	67.8±0.0	20.6±0.0
SAR+CAREM	**63.3**±0.0	**64.7**±0.0	**64.8**±0.0	**61.6**±0.0	61.6±0.0	66.8±0.0	63.8±0.0	71.6±0.0	68.6±0.0	**74.2**±0.1	80.7±0.0	69.4±0.1	70.9±0.0	76.1±0.0	74.2±0.0	**68.8**±0.0	**3.3**±0.0
DeYO	61.8±0.1	63.4±0.1	63.5±0.2	58.9±0.4	60.7±0.1	65.7±0.3	62.8±0.3	71.2±0.1	69.2±0.2	54.4±35.6	80.2±0.2	62.7±1.5	71.5±0.2	75.7±0.1	73.6±0.1	66.4±2.4	18.3±2.6
DeYO+COME	62.8±0.0	64.4±0.1	64.2±0.1	54.6±6.7	60.6±0.3	65.8±0.4	63.9±0.3	71.7±0.1	69.3±0.1	73.1±0.2	80.4±0.0	66.8±0.5	70.9±0.2	76.1±0.1	73.9±0.0	67.9±0.4	34.6±0.3
DeYO+FTL	58.6±0.4	60.1±0.2	59.7±0.2	56.0±0.4	57.6±0.2	63.5±0.2	61.1±0.1	68.5±0.1	65.6±0.6	52.9±34.6	78.6±0.1	62.9±0.4	68.8±0.2	73.7±0.2	71.4±0.1	63.9±2.3	13.8±2.8
DeYO+CAREM	62.2±0.2	63.9±0.2	63.5±0.2	58.7±0.4	61.3±0.4	66.3±0.3	64.6±0.4	71.7±0.1	69.5±0.2	72.5±0.3	80.4±0.1	15.6±30.5	72.4±0.1	**76.3**±0.1	**74.5**±0.1	64.9±2.0	20.7±2.3
DeYO+CAREM+FTL	62.2±0.0	64.1±0.0	63.8±0.1	60.4±0.2	**62.5**±0.1	**67.8**±0.1	**66.5**±0.1	**72.2**±0.0	**70.1**±0.1	73.2±0.1	80.2±0.1	16.7±32.7	**73.3**±0.0	76.2±0.1	74.2±0.1	65.6±2.2	9.7±2.3
ZeroSiam	61.0±0.0	62.2±0.0	62.4±0.0	59.1±0.0	56.9±0.0	63.4±0.0	59.4±0.0	68.1±0.0	66.1±0.1	70.6±0.2	79.2±0.0	67.4±0.1	63.9±0.0	72.9±0.0	70.5±0.0	65.5±0.0	10.6±0.0
ZeroSiam+COME	62.8±0.0	64.0±0.0	64.3±0.0	60.5±0.0	58.6±0.0	65.2±0.0	61.1±0.0	69.8±0.0	67.0±0.0	73.6±0.1	80.3±0.0	69.0±0.0	66.3±0.0	75.1±0.0	73.1±0.0	67.4±0.0	26.9±0.0
ZeroSiam+FTL	60.3±0.0	61.7±0.0	61.8±0.0	58.4±0.1	56.8±0.1	62.6±0.0	58.6±0.0	67.6±0.0	64.9±0.1	69.6±0.5	79.0±0.0	66.1±0.1	64.2±0.1	72.8±0.0	70.5±0.0	65.0±0.0	5.3±0.0
ZeroSiam+CAREM	61.8±0.0	63.0±0.0	63.3±0.0	60.3±0.0	58.9±0.0	64.6±0.0	61.4±0.0	69.2±0.0	66.5±0.1	73.6±0.1	79.6±0.0	69.1±0.0	66.4±0.1	73.9±0.0	71.2±0.0	66.8±0.0	10.1±0.0
ZeroSiam+CAREM+FTL	61.2±0.0	62.6±0.0	62.8±0.0	60.0±0.0	59.0±0.0	64.2±0.0	60.9±0.0	68.8±0.0	66.0±0.1	71.6±0.8	79.4±0.0	69.5±0.0	67.0±0.1	73.7±0.0	71.2±0.0	66.5±0.1	4.4±0.0

5.3 Online Imbalanced Label Shifts

The same pattern can be seen in Table 2 for the imbalanced label shifts. CAREM+FTL again improves the average accuracy and ECE of Tent, EATA, SAR, and ZeroSiam. SAR+CAREM achieves the best average accuracy and ECE with 68.8% and 3.3%, respectively. EATA-C also has a low ECE (5.1%) but is unstable ($52.9\% \pm 7.1\%$ accuracy). Interestingly, the simple Tent+CAREM+FTL almost matches SAR+CAREM in this scenario with 68.7% accuracy and 3.4% ECE.

5.4 Mixed Corruption Shifts

Table 3 shows a different pattern for the continuous mixed-shift stream. CAREM alone improves neither the average accuracy nor ECE of the TTA methods without FTL, worsening the collapse behavior of Tent and DeYO. FTL alone lowers the ECE of every TTA method, but it decreases the accuracy of EATA and ZeroSiam. CAREM+FTL improves the average accuracy and ECE of every TTA method, achieving the overall best accuracy of 65.1% with EATA. EATA-C achieves the second-lowest ECE among TTA methods (9.3%) with an accuracy of 64.0%. Surprisingly, ZeroSiam+COME attains the lowest ECE among TTA methods with 8.2% ECE and 64.5% accuracy.

A deeper analysis of these results can be found in Appendix E. For ZeroSiam, CAREM provides a clear accuracy gain early in adaptation, but this advantage disappears over the long mixed-shift adaptation. CAREM+FTL shows the same transient early gain, while removing the accuracy loss caused by FTL alone. Appendix E also shows that ZeroSiam+COME's low ECE masks a transition from strong underconfidence to overconfidence. It is thus an artifact of ECE binning, rather than a sign of a calibrated method.

Table 3: Mixed-shift results (Acc./ECE) on ImageNet-C at severity level 5 with ViT-Base. The best result in each column is bold, and the best result within each method family is underlined.

Method	Original	COME	FTL	CAREM	CAREM+FTL
No Adapt	55.5 ± 0.0 / 8.3 ± 0.0	/	/	/	/
Tent	49.9 ± 8.9 / 37.6 ± 10.7	58.5 ± 0.4 / 34.9 ± 0.3	60.7 ± 0.1 / 16.6 ± 0.1	27.1 ± 3.0 / 65.9 ± 3.6	63.3 ± 0.1 / 14.0 ± 0.1
EATA	63.7 ± 0.1 / 24.1 ± 0.2	63.9 ± 0.1 / 23.9 ± 0.1	60.6 ± 0.4 / 20.4 ± 0.3	62.8 ± 0.2 / 26.0 ± 0.2	**65.1** ± 0.1 / 14.1 ± 0.2
EATA-C	64.0 ± 0.1 / 9.3 ± 0.0	/	/	/	/
CETA	63.0 ± 0.2 / 22.8 ± 0.2	/	/	/	/
SAR	60.7 ± 0.0 / 16.5 ± 0.0	64.4 ± 0.0 / 23.4 ± 0.0	/	63.8 ± 0.1 / 13.2 ± 0.1	/
DeYO	33.4 ± 3.6 / 57.9 ± 4.2	64.3 ± 0.0 / 31.0 ± 0.0	56.8 ± 0.4 / 27.8 ± 0.4	28.9 ± 1.8 / 63.4 ± 2.1	56.6 ± 0.5 / 28.3 ± 0.6
ZeroSiam	63.5 ± 0.1 / 19.4 ± 0.0	64.5 ± 0.0 / **8.2** ± 0.1	62.8 ± 0.0 / 10.9 ± 0.0	63.4 ± 0.1 / 20.0 ± 0.0	63.6 ± 0.0 / 10.3 ± 0.0

5.5 Effect of Test Batch Size

Figure 4 evaluates SAR+CAREM from batch size 1 to 256 for ViT-Base. We set the learning rate to $0.001 \cdot \text{batchsize}/64$ according to SAR (Niu et al., 2023). CAREM improves both accuracy and ECE at every batch size. The mean radial-tangential cosine similarity ρ remains strongly negative for every batch size and more than 99% of adaptation batches have $\rho < 0$. The radial and tangential gradients' conflict and CAREM's gains are therefore not tied to specific batch size regimes.

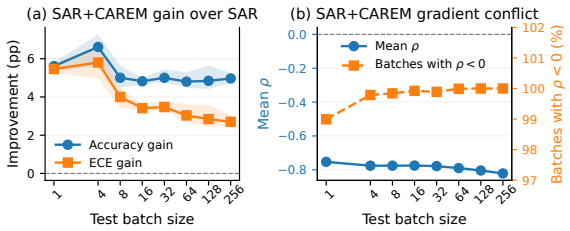


Figure 4: SAR+CAREM batch-size robustness (ViT-Base, ImageNet-C, level 5): (a) mean Acc./ECE gains over SAR (IQR shading); (b) mean radial-tangential cosine similarity ρ and fraction of batches with $\rho < 0$.

6 Conclusion

We studied entropy-based TTA by decomposing its gradient into radial and tangential components in centered-logit space. Although orthogonal in logit space, their parameter gradients strongly conflict, particularly in later layers. Tangential-only adaptation improves accuracy but becomes underconfident, whereas radial-only adaptation inflates the centered-logit norm and harms accuracy. CAREM protects the tangential gradient by minimally projecting the radial gradient when they conflict, without introducing additional hyperparameters.

Across ImageNet-C (Hendrycks & Dietterich, 2019), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), and VisDA-2021 (Bashkirova et al., 2022), CAREM generally improves accuracy, but its effect on calibration is less consistent. Consistent with the concentration of radial gradients in later layers, freezing later layers substantially reduces ECE, and its combination with CAREM often achieves strong accuracy and calibration across methods and backbones. The accuracy and calibration gains also persist across test batch sizes for SAR. Under continuous mixed shifts, however, CAREM alone can worsen collapse in unstable methods, and its early accuracy gains can fade over time. Future work could use layer-wise gradient surgery to account for the different gradient interactions across the network layers. Combining CAREM with methods from lifelong learning (Parisi et al., 2019) may also help retain its gains and prevent collapse during long mixed-shift streams.

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, and interpret results. We have not used generative AI tools to generate synthetic datasets or to clean or reformat datasets. Translation and qualitative or thematic data analysis were not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures, create and edit software code, draft and revise parts of the manuscript, prepare LaTeX tables, summarize and analyze existing literature, identify relevant work and potential research gaps, brainstorm, search for information, improve readability, and propose titles and keywords. We reviewed all AI-assisted work. The authors manually verified the mathematical arguments and proofs, checked literature suggestions, and independently conducted the experiments underlying the reported results. One author inspected and tested AI-assisted code, conducted a manual literature survey in addition to the AI-assisted one, and ensured the correctness of the AI-assisted generation of figures and tables. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our results in Section 5 and Appendix C. We specify the datasets (ImageNet-C, ImageNet-A, ImageNet-R, VisDA-2021) and online adaptation protocols, including the standard, class-imbalanced, and mixed-shift settings, the evaluated model architectures, pretrained checkpoints, checkpoint-specific preprocessing, all optimization settings, adapted parameters, and evaluation metrics. Reported results are means and standard deviations over four runs with different seeds. An anonymized implementation is included with the submission and contains the code and configurations required to reproduce CAREM and its integration with the evaluated TTA methods. Appendix A additionally provides pseudocode for CAREM. The complete source code will be made publicly available after the review process.

REFERENCES

Dina Bashkirova, Dan Hendrycks, Donghyun Kim, Haojin Liao, Samarth Mishra, Chandramouli Rajagopalan, Kate Saenko, Kuniaki Saito, Burhan Ul Tayyab, Piotr Teterwak, et al. VisDA-2021 competition: Universal domain adaptation to improve performance on out-of-distribution data. In *NeurIPS 2021 Competitions and Demonstrations Track*, pp. 66–79. PMLR, 2022.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. *International Conference on Learning Representations*, 2019.

Guohao Chen, Shuaicheng Niu, Deyu Chen, Jiahao Yang, Zitian Zhang, Mingkui Tan, Pengcheng Wu, and Zhiqi Shen. ZeroSiam: An efficient asymmetry for test-time entropy optimization without collapse. In *International Conference on Learning Representations*, volume 2026, pp. 144893–144925, 2026.

Sungha Choi, Seunghan Yang, Seokeon Choi, and Sungrack Yun. Improving test-time adaptation via shift-agnostic weight regularization and nearest source prototypes. In *European Conference on Computer Vision*, pp. 440–458. Springer, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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. *International Conference on Learning Representations*, 2021.

Yunshu Du, Wojciech M Czarnecki, Siddhant M Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *International Conference on Learning Representations*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*, 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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349. IEEE, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Yu-Guan Hsieh, James Thornton, Eugene Ndiaye, Michal Klein, Marco Cuturi, and Pierre Ablin. Careful with that scalpel: Improving gradient surgery with an EMA. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19085–19100. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/hsieh24a.html.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.

Andreas Kirsch and Yarin Gal. A Note on "Assessing Generalization of SGD via Disagreement". *Transactions on Machine Learning Research*, 2022.

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*, volume 2024, pp. 46977–47002, 2024.

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

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021.

Chaithanya Kumar Mummadi, Robin Hutmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-time adaptation to distribution shift by confidence maximization and input transformation. *arXiv preprint arXiv:2106.14999*, 2021.

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, 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*, 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. *International Conference on Learning Representations*, 2023.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

Ori Press, Ravid Shwartz-Ziv, Yann LeCun, and Matthias Bethge. The entropy enigma: Success and failure of entropy minimization. In *International Conference on Machine Learning*. PMLR, 2024.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Mingkui Tan, Guohao Chen, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Peilin Zhao, and Shuaicheng Niu. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6274–6289, 2025.

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

Ross Wightman. PyTorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Hao Yang, Min Wang, Jinshen Jiang, and Yun Zhou. Towards test time adaptation via calibrated entropy minimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3736–3746, 2024.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. OpenOOD: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: Test-time adaption by conservatively minimizing entropy. *International Conference on Learning Representations*, 2025.

APPENDIX

CONTENTS

A PyTorch-Style CAREM Pseudocode **14**

B Proofs **14**

C Additional Experimental Details **15**

C.1 Datasets and Evaluation Protocol 15

C.2 Models and Implementation Details 15

D Additional Results **16**

D.1 Empirical Conflict Between The Parameter Gradients of ResNet-50-GN 16

D.2 ImageNet-C Results for ResNet-50-GN 16

D.3 Results on ImageNet-R, ImageNet-A, and VisDA-2021 18

E Accuracy and Calibration Dynamics Under Mixed Corruption Shifts **20**

F Comparison of Different Gradient Surgery Rules **21**

A PyTorch-Style CAREM Pseudocode

Algorithm 1 shows the pseudocode of CAREM. The logit gradient is decomposed for each sample, while the parameter gradients are flattened and projected globally. CAREM simply replaces each `backward()` call of the TTA method with `carem_backwards`. Gradients of logit-independent regularizers are added separately before the optimizer step.

Algorithm 1: Pseudocode of CAREM in a PyTorch-like style.

```
def carem_backwards(loss, logits, params):
    g = torch.autograd.grad(loss, logits, retain_graph=True)[0]
    z = (logits - logits.mean(dim=-1, keepdim=True)).detach()
    u = z / z.norm(dim=-1, keepdim=True)
    g_rad = (g * u).sum(dim=-1, keepdim=True) * u
    g_tan = g - g_rad

    grad_rad = flatten(torch.autograd.grad(
        logits, params, grad_outputs=g_rad, retain_graph=True))
    grad_tan = flatten(torch.autograd.grad(
        logits, params, grad_outputs=g_tan))

    combined = gradient_surgery_rule(grad_tan, grad_rad)
    pieces = torch.split(combined, [p.numel() for p in params])
    for p, piece in zip(params, pieces):
        p.grad = piece.view_as(p)
```

B Proofs

The following result characterizes gradient descent steps in centered logit space and does not describe a neural network parameter update.

Proposition 1 (The radial EM component increases the centered-logit norm)**.** *Let* $\boldsymbol{z} \neq \mathbf{0}$*,* $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{z})$*,* $\mathcal{H}(\boldsymbol{p}) = -\sum_{k=1}^{K} p_k \log p_k$*, and* $\boldsymbol{g} = \nabla_{\boldsymbol{z}} \mathcal{H}(\mathrm{softmax}(\boldsymbol{z}))$*. For every* $\eta \geq 0$*,*

$$\|\boldsymbol{z} - \eta \boldsymbol{g}_{\mathrm{rad}}\|_2^2 \geq \|\boldsymbol{z}\|_2^2. \tag{13}$$

Proof. The derivative of the i-th softmax probability with respect to z_k is

$$\frac{\partial p_i}{\partial z_k} = p_i(\delta_{ik} - p_k), \tag{14}$$

where δ_{ik} denotes the Kronecker delta. Differentiating the entropy directly therefore gives

$$g_k = \frac{\partial \mathcal{H}(\boldsymbol{p})}{\partial z_k} = -\sum_i (\log p_i + 1)\frac{\partial p_i}{\partial z_k} = p_k \left(\sum_i p_i \log p_i - \log p_k \right) = p_k \left(\sum_i p_i z_i - z_k \right). \tag{15}$$

The final equality follows from $\log p_i = z_i - \log \sum_j \exp(z_j)$ and $\sum_i p_i = 1$. Consequently,

$$\boldsymbol{z}^\top \boldsymbol{g} = \sum_k z_k g_k = \left(\sum_i p_i z_i \right)^2 - \sum_i p_i z_i^2 = -\frac{1}{2} \sum_{i,j} p_i p_j (z_i - z_j)^2. \tag{16}$$

Because all softmax probabilities are positive and the nonzero centered vector $\boldsymbol{z}$ has at least two distinct entries, $\boldsymbol{z}^\top \boldsymbol{g} < 0$.

From the definition of $\boldsymbol{g}_{\mathrm{rad}} = \frac{\boldsymbol{z}^\top \boldsymbol{g}}{\|\boldsymbol{z}\|_2^2} \boldsymbol{z}$, it therefore follows that

$$\|\boldsymbol{z} - \eta \boldsymbol{g}_{\mathrm{rad}}\|_2^2 = \left(1 - \eta \frac{\boldsymbol{z}^\top \boldsymbol{g}}{\|\boldsymbol{z}\|_2^2} \right)^2 \|\boldsymbol{z}\|_2^2 \geq \|\boldsymbol{z}\|_2^2, \tag{17}$$

because $\eta \geq 0$ and $\boldsymbol{z}^\top \boldsymbol{g} < 0$. □

Thus, the radial component of the EM gradient always increases the centered logits' norm under a direct logit-space descent step.

C Additional Experimental Details

C.1 Datasets and Evaluation Protocol

ImageNet-C. ImageNet-C (Hendrycks & Dietterich, 2019) applies 15 common corruptions from the noise, blur, weather, and digital categories to the ImageNet validation images. Each corruption has five severity levels; we evaluate all 15 corruptions at severity 5. The dataset and corruption code are available from the authors' repository[2]. In the standard setting, we reset the model to its pretrained state before each corruption and process its images online. Following SAR (Niu et al., 2023), we also evaluate a class-ordered stream with an extreme class imbalance ($q_{\max}/q_{\min} = \infty$) and a mixed-corruption stream in which the model continues adapting across all corruption types without resets.

ImageNet-A and ImageNet-R. ImageNet-A[3] contains naturally occurring images that are difficult for ImageNet-trained classifiers (Hendrycks et al., 2021b). ImageNet-R[4] contains artistic and other renditions of ImageNet classes (Hendrycks et al., 2021a). Both datasets cover 200 ImageNet classes and test shifts that differ from the algorithmic corruptions in ImageNet-C. The original dataset repositories provide the images and class mappings. We report results for both pretrained backbones in Appendix D.3.

VisDA-2021. The VisDA-2021 domain-adaptation benchmark uses ImageNet-1K as its source domain and evaluates on images from shifted target distributions, including ObjectNet and ImageNet variants (Bashkirova et al., 2022). We use it as an additional natural-shift benchmark; the official development kit describes the target data[5]. Results for both backbones are reported in Table 9.

C.2 Models and Implementation Details

Pretrained models and common settings. We use the ImageNet-pretrained ViT-Base with LayerNorm `vit_base_patch16_224` and ResNet-50 with GroupNorm `resnet50_gn` from `timm`[6]. We use the checkpoint-specific input normalization: the ViT-Base checkpoint requires channel mean and standard deviation $(0.5, 0.5, 0.5)$, while ResNet-50-GN uses the standard ImageNet normalization. For the standard configuration, we update the affine parameters of the normalization layers using SGD with momentum 0.9, a batch size of 64, and learning rates 0.001 for ViT-Base and 0.00025 for ResNet-50-GN. For the batch-size sweep in Section 5.5, we scale SAR's learning rate proportionally to the batch size. Frozen-top-layers (FTL) follows SAR (Niu et al., 2023): we freeze ViT blocks 9–11 or ResNet layer 4 during adaptation.

Tent[7] (Wang et al., 2021). We follow all hyperparameters that are set in Tent. We use the common SGD settings: momentum 0.9, batch size 64, and learning rates 0.001 for ViT-Base and 0.00025 for ResNet-50-GN.

EATA[8] (Niu et al., 2022). We follow all hyperparameters that are set in EATA. We use the specified reliability threshold $E_0 = 0.4 \ln 1000$, redundancy threshold $\epsilon = 0.05$, Fisher coefficient $\beta = 2000$, and 2000 samples for the Fisher estimate. We find the aggressive reliability threshold $E_0 = 0.4 \ln 1000$ to still be adequate for ViT-Base, despite the different input normalization, due to the usage of E_0 in sample reweighting. We use the common SGD settings above.

EATA-C (Tan et al., 2025). We follow all hyperparameters that are set in EATA-C. We implemented EATA-C from its paper because no public implementation was available for these experiments. We use the specified reliability thresholds $E_0 = 0.4 \ln 1000$ for ViT-Base and $E_0 = 0.5 \ln 1000$ for

[2] ImageNet-C: `https://github.com/hendrycks/robustness`
[3] ImageNet-A: `https://github.com/hendrycks/natural-adv-examples`
[4] ImageNet-R: `https://github.com/hendrycks/imagenet-r`
[5] `https://github.com/VisionLearningGroup/visda21-dev`
[6] `https://github.com/huggingface/pytorch-image-models`
[7] `https://github.com/DequanWang/tent`
[8] `https://github.com/mr-eggplant/EATA`

ResNet-50-GN, redundancy thresholds $\epsilon = 0.05$ for ViT-Base and $\epsilon = 0.07$ for ResNet-50-GN, stochastic-depth rate 0.2, Fisher coefficient $\beta = 50$, and 2000 samples for the Fisher estimate. We use SGD with momentum 0.9, batch size 64, and the paper's learning rates 0.1 for ViT-Base and 0.005 for ResNet-50-GN.

CETA (Yang et al., 2024). We implemented the calibrated-entropy objective from the paper because no public implementation was available for these experiments. We use the common SGD settings above and set $\gamma = 1$.

SAR[9] (Niu et al., 2023). We use the specified SAM radius $\rho = 0.05$, model recovery threshold $e_0 = 0.2$, and freeze ViT blocks 9-11 or ResNet layer 4 during adaptation. We find the reliability threshold $E_0 = 0.4 \ln 1000$ to be adequate for ResNet-50-GN, but too aggressive for ViT-Base due to the change in its input normalization. We therefore set $E_0 = 0.7 \ln 1000$ for ViT-Base. We use the common SGD settings above; only the SAR batch-size sweep changes the learning rate, to $0.001 \cdot \mathrm{batchsize}/64$ for ViT-Base and $0.00025 \min\left(\frac{\mathrm{batchsize}}{32}, 1\right)$ for ResNet-50-GN, following SAR (Niu et al., 2023). We double the learning rate for batch size 1 according to the official implementation (Niu et al., 2023).

DeYO[10] (Lee et al., 2024). We follow all hyperparameters that are set in DeYO. We use the specified entropy threshold for sample selection $\tau_{\mathrm{Ent}} = 0.5 \ln 1000$, reliability threshold $E_0 = 0.4 \ln 1000$, and PLPD threshold $\tau_{\mathrm{PLPD}} = 0.2$. We use the common SGD settings above.

ZeroSiam[11] (Chen et al., 2026). We set the coefficient $\alpha = 1$ and use a linear predictor. We use the common SGD settings above.

COME[12] (Zhang et al., 2025). We follow all hyperparameters that are set in COME. We use the specified magnitude multiplier $\tau = 1$ and norm $p = 2$. The remaining trainable parameters and settings follow the corresponding parent method.

CAREM and FTL. CAREM replaces the entropy backward pass with the projection in Section 4.4; it adds no hyperparameters. FTL freezes ViT blocks 9–11 or ResNet layer 4 during adaptation, following SAR (Niu et al., 2023). The remaining trainable parameters and settings follow the corresponding parent method.

D Additional Results

D.1 Empirical Conflict Between The Parameter Gradients of ResNet-50-GN

To examine whether the gradient interactions identified for ViT-Base extend beyond transformer architectures, we repeat the gradient analysis in Section 4.2 with an ImageNet-pretrained ResNet-50-GN (He et al., 2016; Wightman, 2019). Figure 5(a) shows that the global cosine similarity remains negative throughout adaptation, confirming that the radial and tangential parameter gradients also conflict in ResNet-50-GN. The conflict is strongest early on and declines during adaptation, while the global radial-to-tangential norm ratio steadily decreases, indicating that the tangential gradient increasingly dominates the aggregate update. The block-wise results in Figures 5(b) and 5(c) show a similar pattern to ViT-Base as well. The gradients remain negatively aligned for both GroupNorm weights and biases across all residual blocks, whereas the norm ratio rises sharply in the final residual stage and exceeds one in the last block. Thus, global aggregation masks a concentration of the radial gradient in the later layers. This reproduces the qualitative behavior observed for ViT-Base and further motivates FTL as a complementary mechanism for limiting late-layer confidence inflation.

D.2 ImageNet-C Results for ResNet-50-GN

We further evaluate CAREM in the standard online TTA, imbalanced shifts, and mixed shifts settings from Section 5 for ResNet-50-GN. We follow the adaptation protocols, metrics, and optimization settings described in Section 5 and Appendix C.

[9] https://github.com/mr-eggplant/SAR
[10] https://github.com/Jhyun17/DeYO
[11] https://github.com/Cascol-Chen/ZeroSiam
[12] https://github.com/BlueWhaleLab/COME

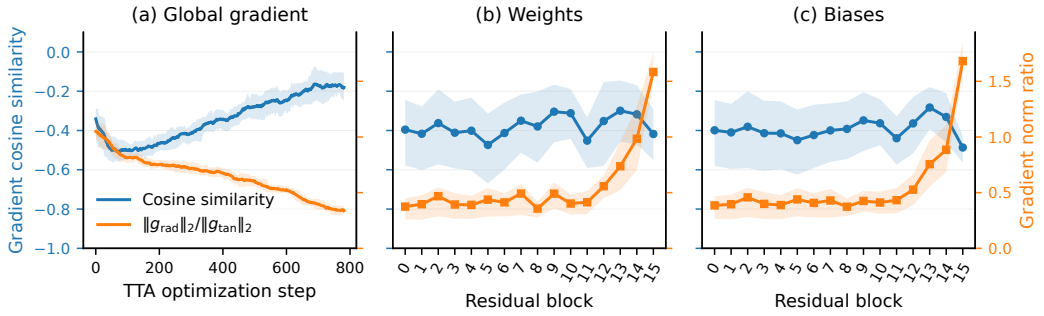


Figure 5: Cosine similarity and norm ratio of the radial and tangential parameter gradients during TTA. (a) Evolution of the global cosine similarity and gradient norm ratio. Curves are smoothed with a 50-step moving window. (b)–(c) Cosine similarity and norm ratio averaged over the adaptation run for the GroupNorm weights and biases, respectively, of each residual block. The shaded areas represent the interquartile range. All experiments use a ResNet-50-GN model on ImageNet-C with Gaussian noise at severity level 5 and a batch size of 64.

Table 4 shows that CAREM raises the average accuracy of all five TTA baselines under standard corruption shifts, most notably SAR (35.0% to 42.4%) and EATA (45.7% to 49.7%). On its own, CAREM does not consistently lower ECE, whereas FTL lowers ECE for each method. The combination CAREM+FTL improves the average accuracy and ECE over both CAREM and FTL for all TTA methods except Tent. The highest accuracies come from DeYO+CAREM+FTL (50.9%, 6.0% ECE) and EATA+CAREM+FTL (50.6%, 4.4% ECE); the latter nearly matches EATA-C's 4.3% ECE at higher accuracy. COME barely changes the ResNet-50-GN no adapt baseline. This is due to the fact that COME's fixed uncertainty evidence makes its TTA objective sensitive to the model's absolute logit scale and offset (Zhang et al., 2025). For ResNet-50-GN's logits, this reduces adaptation gradients by approximately three orders of magnitude due to the uncertainty term overpowering the logits in the loss.

Table 4: Standard online TTA on ImageNet-C at severity level 5 with ResNet-50-GN. We report top-1 accuracy and ECE (15 bins), in percent, as mean ± standard deviation over four seeds. Each corruption is adapted independently. The best result in each column is bold, and the best result within each method family is underlined.

	Noise			Blur				Weather				Digital				Avg.	
Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elast.	Pixel	JPEG	Acc.↑	ECE↓
No Adapt	22.1 ±0.0	23.0 ±0.0	22.0 ±0.0	19.8 ±0.0	11.4 ±0.0	21.5 ±0.0	25.0 ±0.0	40.3 ±0.0	47.0 ±0.0	34.0 ±0.0	68.8 ±0.0	36.2 ±0.0	18.5 ±0.0	29.2 ±0.0	52.6 ±0.0	31.4 ±0.0	7.6 ±0.0
Tent	23.3 ±1.5	28.2 ±0.6	24.6 ±1.5	15.0 ±0.1	9.3 ±0.9	21.8 ±0.3	23.0 ±0.4	26.0 ±0.6	33.2 ±0.2	3.5 ±0.0	69.9 ±0.1	42.2 ±0.1	11.4 ±0.4	48.4 ±0.1	54.6 ±0.1	29.0 ±0.1	40.1 ±0.2
Tent+COME	22.1 ±0.0	23.0 ±0.0	22.0 ±0.0	19.8 ±0.0	11.4 ±0.0	21.5 ±0.0	25.0 ±0.0	40.3 ±0.0	47.0 ±0.0	34.0 ±0.0	68.8 ±0.0	36.2 ±0.0	18.5 ±0.0	29.2 ±0.0	52.6 ±0.0	31.4 ±0.0	7.6 ±0.0
Tent+FTL	28.9 ±0.3	31.9 ±0.3	30.0 ±0.4	16.6 ±0.1	12.4 ±1.0	25.7 ±0.2	29.2 ±0.2	31.1 ±0.9	39.5 ±0.2	5.2 ±0.1	70.6 ±0.0	43.9 ±0.0	14.9 ±0.7	49.8 ±0.1	55.7 ±0.1	32.4 ±0.1	16.1 ±0.1
Tent+CAREM	31.9 ±0.2	33.9 ±0.9	33.2 ±0.6	16.7 ±0.2	4.1 ±0.4	27.8 ±0.3	23.9 ±0.7	24.6 ±0.7	33.2 ±0.3	3.5 ±0.3	71.0 ±0.1	44.7 ±0.1	6.7 ±0.4	51.2 ±0.1	55.7 ±0.1	30.8 ±0.2	40.6 ±0.3
Tent+CAREM+FTL	35.4 ±0.1	38.1 ±0.2	36.9 ±0.3	19.8 ±0.1	5.7 ±0.7	33.0 ±0.1	33.1 ±0.4	25.9 ±0.7	41.6 ±0.9	8.7 ±1.9	72.3 ±0.1	46.8 ±0.0	5.9 ±0.3	52.9 ±0.1	56.6 ±0.1	34.2 ±0.1	19.1 ±0.7
EATA	39.0 ±0.2	40.9 ±0.2	40.0 ±0.2	28.7 ±0.7	27.5 ±0.3	36.8 ±0.2	38.9 ±0.1	51.0 ±0.0	49.0 ±0.4	55.7 ±0.2	71.9 ±0.1	50.0 ±0.2	41.5 ±0.3	56.0 ±0.1	58.3 ±0.1	45.7 ±0.1	24.5 ±0.1
EATA-C	37.2 ±0.3	40.0 ±0.2	38.5 ±0.3	25.8 ±0.3	26.0 ±0.5	36.2 ±0.6	38.5 ±0.3	51.6 ±0.3	50.9 ±0.4	54.8 ±0.3	72.7 ±0.1	50.1 ±0.2	39.3 ±0.3	55.6 ±0.3	57.9 ±0.2	45.0 ±0.1	**4.3** ±0.0
EATA+COME	22.0 ±0.0	22.8 ±0.0	21.9 ±0.0	19.7 ±0.0	11.3 ±0.0	21.3 ±0.0	25.0 ±0.0	40.1 ±0.0	47.1 ±0.0	33.9 ±0.0	68.6 ±0.0	35.9 ±0.0	18.3 ±0.0	28.6 ±0.0	52.3 ±0.0	31.3 ±0.0	7.9 ±0.0
EATA+FTL	37.9 ±0.2	40.0 ±0.2	39.2 ±0.2	25.4 ±0.8	25.8 ±0.3	36.7 ±0.1	38.3 ±0.5	50.1 ±0.1	47.0 ±0.1	54.0 ±0.1	71.8 ±0.0	48.0 ±0.2	40.5 ±0.2	53.5 ±0.2	56.8 ±0.0	44.3 ±0.0	6.1 ±0.0
EATA+CAREM	42.7 ±0.2	44.6 ±0.2	43.7 ±0.2	34.2 ±0.3	**33.1** ±0.0	42.0 ±0.1	44.4 ±0.2	53.9 ±0.2	52.6 ±0.1	59.7 ±0.2	73.0 ±0.1	53.1 ±0.2	48.7 ±0.3	60.0 ±0.2	60.2 ±0.1	49.7 ±0.0	21.4 ±0.1
EATA+CAREM+FTL	44.2 ±0.1	46.2 ±0.1	45.6 ±0.1	**35.6** ±0.2	32.7 ±0.1	43.5 ±0.1	**45.9** ±0.2	55.2 ±0.1	53.5 ±0.2	60.5 ±0.2	73.8 ±0.1	53.5 ±0.1	49.8 ±0.1	59.6 ±0.1	59.6 ±0.1	50.6 ±0.0	4.4 ±0.0
CETA	26.4 ±0.5	29.7 ±0.6	26.8 ±0.6	14.9 ±0.2	5.2 ±0.4	21.9 ±0.3	22.7 ±0.1	25.0 ±0.2	31.9 ±0.8	4.4 ±0.1	69.8 ±0.0	42.1 ±0.1	11.0 ±0.4	48.1 ±0.0	54.4 ±0.1	28.9 ±0.1	40.5 ±0.1
SAR	32.3 ±0.2	34.3 ±0.3	33.5 ±0.3	18.6 ±0.2	18.6 ±0.6	30.1 ±0.4	30.9 ±0.3	42.1 ±0.4	43.4 ±0.1	6.3 ±1.8	70.3 ±0.0	43.8 ±0.2	15.8 ±1.4	48.9 ±0.1	55.4 ±0.0	35.0 ±0.2	12.9 ±0.3
SAR+COME	22.1 ±0.0	23.0 ±0.0	22.0 ±0.0	19.8 ±0.0	11.4 ±0.0	21.5 ±0.0	25.0 ±0.0	40.3 ±0.0	47.0 ±0.0	34.0 ±0.0	68.8 ±0.0	36.2 ±0.0	18.5 ±0.0	29.2 ±0.0	52.6 ±0.0	31.4 ±0.0	7.6 ±0.0
SAR+CAREM	39.1 ±0.1	41.3 ±0.1	40.8 ±0.1	28.3 ±0.6	24.0 ±0.5	37.0 ±0.3	37.5 ±0.2	49.0 ±0.2	49.3 ±0.2	52.9 ±1.0	72.4 ±0.1	48.3 ±0.1	5.4 ±0.5	54.1 ±0.1	56.9 ±0.1	42.4 ±0.1	9.3 ±0.1
DeYO	23.5 ±1.1	26.9 ±4.3	27.0 ±5.4	12.5 ±1.9	18.2 ±0.6	26.7 ±0.9	17.2 ±0.4	42.4 ±0.9	42.0 ±1.3	2.1 ±1.3	69.2 ±0.3	44.2 ±0.4	26.0 ±0.3	54.6 ±0.5	53.9 ±0.5	32.4 ±0.3	53.3 ±0.4
DeYO+COME	22.0 ±0.0	22.9 ±0.0	21.9 ±0.0	19.8 ±0.0	11.3 ±0.0	21.4 ±0.0	25.0 ±0.0	40.2 ±0.0	47.1 ±0.0	34.0 ±0.0	68.6 ±0.0	35.9 ±0.0	18.4 ±0.0	28.8 ±0.0	52.4 ±0.0	31.3 ±0.0	7.8 ±0.0
DeYO+FTL	40.7 ±0.2	42.9 ±0.2	42.0 ±0.2	21.8 ±0.4	24.1 ±0.3	38.2 ±0.2	37.2 ±1.4	50.9 ±0.2	48.9 ±0.3	29.2 ±31.2	72.9 ±0.1	50.1 ±0.1	41.8 ±0.5	56.1 ±0.1	57.9 ±0.1	43.6 ±2.0	10.6 ±3.1
DeYO+CAREM	27.7 ±5.7	35.3 ±1.9	28.0 ±5.7	17.2 ±2.1	19.4 ±2.7	31.2 ±0.3	18.7 ±1.9	45.4 ±1.6	45.6 ±1.1	40.1 ±26.0	70.4 ±0.1	47.1 ±0.5	27.3 ±3.6	56.6 ±0.3	56.2 ±0.0	37.7 ±2.6	48.2 ±3.0
DeYO+CAREM+FTL	**45.3** ±0.0	**47.2** ±0.1	**46.5** ±0.1	34.6 ±0.7	30.5 ±0.2	**44.4** ±0.1	41.9 ±1.8	**56.4** ±0.2	**54.8** ±0.2	**61.0** ±0.4	**74.0** ±0.0	**54.5** ±0.1	**50.4** ±0.4	**61.4** ±0.1	**60.5** ±0.1	**50.9** ±0.2	6.0 ±0.3
ZeroSiam	32.8 ±0.2	34.8 ±0.2	34.1 ±0.1	19.1 ±0.0	20.5 ±0.1	30.3 ±0.1	31.1 ±0.1	44.5 ±0.1	43.1 ±0.1	46.7 ±0.1	70.1 ±0.1	44.5 ±0.2	29.8 ±0.1	49.5 ±0.1	55.0 ±0.1	39.1 ±0.0	24.3 ±0.0
ZeroSiam+COME	22.1 ±0.0	23.0 ±0.0	22.0 ±0.0	19.8 ±0.0	11.4 ±0.0	21.5 ±0.0	25.0 ±0.0	40.3 ±0.0	47.0 ±0.0	34.0 ±0.0	68.8 ±0.0	36.2 ±0.0	18.5 ±0.0	29.2 ±0.0	52.6 ±0.0	31.4 ±0.0	7.6 ±0.0
ZeroSiam+FTL	32.5 ±0.2	34.6 ±0.2	33.8 ±0.1	19.2 ±0.1	20.2 ±0.1	30.7 ±0.1	32.6 ±0.2	45.9 ±0.1	44.2 ±0.0	47.0 ±0.1	70.6 ±0.0	44.7 ±0.1	31.0 ±0.1	49.9 ±0.1	55.6 ±0.1	39.5 ±0.1	6.6 ±0.0
ZeroSiam+CAREM	34.5 ±0.3	36.6 ±0.1	35.9 ±0.1	19.6 ±0.1	21.5 ±0.1	31.6 ±0.0	32.1 ±0.1	45.5 ±0.2	43.6 ±0.1	48.8 ±0.2	70.5 ±0.1	45.5 ±0.1	31.0 ±0.1	51.0 ±0.1	55.4 ±0.0	40.2 ±0.0	25.1 ±0.0
ZeroSiam+CAREM+FTL	35.4 ±0.3	37.7 ±0.2	36.9 ±0.1	20.6 ±0.1	22.0 ±0.1	33.2 ±0.1	34.9 ±0.3	47.9 ±0.1	46.0 ±0.1	50.6 ±0.1	71.4 ±0.1	46.4 ±0.1	34.2 ±0.1	52.0 ±0.1	56.1 ±0.0	41.7 ±0.0	6.3 ±0.0

Table 5 shows the same accuracy gains under imbalanced label shifts: CAREM improves every baseline, and DeYO+CAREM+FTL achieves the highest average accuracy (51.7%). EATA+CAREM+FTL reaches 49.8% accuracy and 4.6% ECE, close to EATA-C's 4.1% ECE but substantially more accurate (43.1%). FTL again reduces the ECE of Tent, EATA, DeYO, and Ze-

Table 5: Online imbalanced label-shift results on ImageNet-C at severity level 5 with ResNet-50-GN and $q_{\max}/q_{\min} = \infty$. We report top-1 accuracy and ECE (15 bins), in percent, as mean $\pm$ standard deviation over four seeds. The best result in each column is bold, and the best result within each method family is underlined.

	Noise			Blur				Weather				Digital				Avg.	
Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elast.	Pixel	JPEG	Acc.↑	ECE↓
No Adapt	22.1 ± 0.0	23.0 ± 0.0	22.0 ± 0.0	19.8 ± 0.0	11.4 ± 0.0	21.5 ± 0.0	25.0 ± 0.0	40.3 ± 0.0	47.0 ± 0.0	34.0 ± 0.0	68.8 ± 0.0	36.2 ± 0.0	18.5 ± 0.0	29.2 ± 0.0	52.6 ± 0.0	31.4 ± 0.0	7.6 ± 0.0
Tent	15.5 ± 0.0	26.6 ± 0.0	17.1 ± 0.0	17.1 ± 0.0	17.2 ± 0.0	26.1 ± 0.0	20.6 ± 0.0	31.6 ± 0.0	36.1 ± 0.0	7.1 ± 0.0	70.1 ± 0.0	42.8 ± 0.0	13.3 ± 0.0	48.8 ± 0.0	54.8 ± 0.0	29.7 ± 0.0	40.6 ± 0.0
Tent+COME	22.1 ± 0.0	23.0 ± 0.0	22.0 ± 0.0	19.8 ± 0.0	11.4 ± 0.0	21.5 ± 0.0	25.0 ± 0.0	40.3 ± 0.0	47.0 ± 0.0	34.0 ± 0.0	68.8 ± 0.0	36.2 ± 0.0	18.5 ± 0.0	29.2 ± 0.0	52.6 ± 0.0	31.4 ± 0.0	7.6 ± 0.0
Tent+FTL	27.6 ± 0.0	31.8 ± 0.0	29.6 ± 0.0	18.0 ± 0.0	18.8 ± 0.0	27.9 ± 0.0	29.1 ± 0.0	35.2 ± 0.0	41.1 ± 0.0	39.5 ± 0.0	70.7 ± 0.0	43.9 ± 0.0	16.4 ± 0.0	49.9 ± 0.0	55.5 ± 0.0	35.7 ± 0.0	11.5 ± 0.0
Tent+CAREM	21.5 ± 0.0	29.1 ± 0.0	27.0 ± 0.0	19.6 ± 0.0	18.8 ± 0.0	31.9 ± 0.0	20.8 ± 0.0	34.1 ± 0.0	39.7 ± 0.0	19.9 ± 0.0	71.3 ± 0.0	45.5 ± 0.0	9.3 ± 0.0	52.1 ± 0.0	56.2 ± 0.0	33.1 ± 0.0	38.5 ± 0.0
Tent+CAREM+FTL	35.1 ± 0.0	38.4 ± 0.0	37.1 ± 0.0	21.5 ± 0.0	21.1 ± 0.0	35.3 ± 0.0	34.4 ± 0.0	42.3 ± 0.0	47.7 ± 0.0	48.0 ± 0.0	72.5 ± 0.0	47.4 ± 0.0	9.2 ± 0.0	53.4 ± 0.0	56.6 ± 0.0	40.0 ± 0.0	11.2 ± 0.0
EATA	34.7 ± 0.5	37.2 ± 0.3	36.4 ± 0.5	21.3 ± 0.4	24.4 ± 0.4	31.3 ± 0.5	33.7 ± 0.7	46.7 ± 0.3	41.7 ± 1.1	51.9 ± 0.2	71.1 ± 0.1	47.4 ± 0.2	37.5 ± 0.5	53.8 ± 0.3	56.8 ± 0.1	41.7 ± 0.2	28.9 ± 0.2
EATA-C	35.3 ± 0.4	37.1 ± 1.0	35.6 ± 0.6	24.5 ± 0.2	21.4 ± 2.2	34.7 ± 0.4	35.5 ± 1.2	50.3 ± 0.1	50.1 ± 0.2	52.5 ± 0.4	72.4 ± 0.2	49.1 ± 0.3	35.8 ± 1.4	55.0 ± 0.3	57.8 ± 0.3	43.1 ± 0.1	**4.1** ± 0.1
EATA+COME	22.0 ± 0.0	23.0 ± 0.0	22.0 ± 0.0	19.8 ± 0.0	11.4 ± 0.0	21.4 ± 0.0	25.0 ± 0.0	40.2 ± 0.0	47.1 ± 0.0	34.0 ± 0.0	68.6 ± 0.0	36.0 ± 0.0	18.4 ± 0.0	28.9 ± 0.0	52.4 ± 0.0	31.4 ± 0.0	7.8 ± 0.0
EATA+FTL	35.9 ± 0.3	38.4 ± 0.2	37.8 ± 0.1	21.9 ± 0.2	24.7 ± 0.1	34.8 ± 0.1	37.2 ± 0.1	47.6 ± 0.3	45.2 ± 0.4	52.1 ± 0.2	71.6 ± 0.0	46.3 ± 0.2	40.3 ± 0.2	53.1 ± 0.1	56.5 ± 0.0	42.9 ± 0.0	6.7 ± 0.0
EATA+CAREM	38.7 ± 0.1	41.0 ± 0.3	40.2 ± 0.2	30.7 ± 0.2	29.8 ± 0.7	37.6 ± 0.1	40.4 ± 0.4	49.8 ± 0.7	47.9 ± 1.2	57.0 ± 0.2	72.5 ± 0.0	50.7 ± 0.3	44.4 ± 0.9	58.8 ± 0.2	59.2 ± 0.3	46.6 ± 0.1	25.2 ± 0.1
EATA+CAREM+FTL	43.2 ± 0.1	45.1 ± 0.1	44.4 ± 0.1	34.3 ± 0.2	31.7 ± 0.3	42.0 ± 0.3	44.4 ± 0.1	54.4 ± 0.1	52.4 ± 0.2	60.1 ± 0.1	73.7 ± 0.0	52.7 ± 0.2	50.3 ± 0.1	59.2 ± 0.1	59.6 ± 0.0	49.8 ± 0.0	4.6 ± 0.0
CETA	20.3 ± 0.0	27.8 ± 0.0	27.7 ± 0.0	16.9 ± 0.0	16.6 ± 0.0	25.8 ± 0.0	20.3 ± 0.0	31.2 ± 0.0	35.0 ± 0.0	7.6 ± 0.0	70.0 ± 0.0	42.6 ± 0.0	13.4 ± 0.0	48.4 ± 0.0	54.7 ± 0.0	30.5 ± 0.0	39.7 ± 0.0
SAR	31.9 ± 0.0	34.8 ± 0.0	33.4 ± 0.0	20.2 ± 0.0	19.4 ± 0.0	30.9 ± 0.0	29.4 ± 0.0	41.3 ± 0.0	43.1 ± 0.0	45.9 ± 0.0	70.3 ± 0.0	43.9 ± 0.0	15.7 ± 0.0	49.0 ± 0.0	55.2 ± 0.0	37.6 ± 0.0	9.3 ± 0.0
SAR+COME	22.1 ± 0.0	23.0 ± 0.0	22.0 ± 0.0	19.8 ± 0.0	11.4 ± 0.0	21.5 ± 0.0	25.0 ± 0.0	40.3 ± 0.0	47.0 ± 0.0	34.0 ± 0.0	68.8 ± 0.0	36.2 ± 0.0	18.5 ± 0.0	29.2 ± 0.0	52.6 ± 0.0	31.4 ± 0.0	7.6 ± 0.0
SAR+CAREM	39.5 ± 0.0	41.3 ± 0.0	41.1 ± 0.0	29.6 ± 0.0	24.4 ± 0.0	38.2 ± 0.0	37.5 ± 0.0	50.5 ± 0.0	49.7 ± 0.0	55.9 ± 0.0	72.5 ± 0.0	49.2 ± 0.0	6.2 ± 0.0	54.7 ± 0.0	57.0 ± 0.0	43.1 ± 0.0	9.4 ± 0.0
DeYO	12.9 ± 2.1	30.9 ± 3.2	16.0 ± 1.5	18.7 ± 0.3	17.7 ± 0.6	31.4 ± 1.1	16.5 ± 1.3	42.1 ± 1.5	43.5 ± 0.9	50.3 ± 1.1	69.5 ± 0.2	45.3 ± 0.6	27.7 ± 2.6	54.5 ± 0.5	55.0 ± 0.2	35.5 ± 0.5	50.1 ± 0.6
DeYO+COME	22.0 ± 0.0	22.9 ± 0.0	22.0 ± 0.0	19.8 ± 0.0	11.3 ± 0.0	21.4 ± 0.0	25.0 ± 0.0	40.2 ± 0.0	47.1 ± 0.0	34.0 ± 0.0	68.6 ± 0.0	35.9 ± 0.0	18.4 ± 0.0	28.9 ± 0.0	52.4 ± 0.0	31.3 ± 0.0	7.8 ± 0.0
DeYO+FTL	39.7 ± 0.1	42.8 ± 0.0	41.3 ± 0.1	24.5 ± 0.1	24.7 ± 0.1	39.1 ± 0.0	6.8 ± 0.9	51.2 ± 0.1	49.6 ± 0.0	57.0 ± 0.1	72.8 ± 0.0	50.3 ± 0.0	44.6 ± 0.3	56.2 ± 0.1	57.8 ± 0.0	43.9 ± 0.1	11.3 ± 0.2
DeYO+CAREM	25.4 ± 6.2	32.3 ± 1.7	26.3 ± 3.1	22.2 ± 1.3	21.0 ± 1.4	34.0 ± 1.1	19.1 ± 1.0	46.1 ± 0.8	47.1 ± 1.0	53.9 ± 0.8	70.5 ± 0.1	48.5 ± 0.2	37.6 ± 1.6	57.6 ± 0.3	56.8 ± 0.2	39.9 ± 0.4	45.8 ± 0.4
DeYO+CAREM+FTL	**44.8** ± 0.1	**46.4** ± 0.1	**46.1** ± 0.0	**35.0** ± 0.1	**32.4** ± 0.2	**45.0** ± 0.1	**45.7** ± 0.2	**56.6** ± 0.1	**55.0** ± 0.1	**62.5** ± 0.1	**74.0** ± 0.0	**55.1** ± 0.0	**54.1** ± 0.1	**62.1** ± 0.0	**61.0** ± 0.0	**51.7** ± 0.0	5.5 ± 0.0
ZeroSiam	32.8 ± 0.0	35.0 ± 0.0	34.0 ± 0.0	19.5 ± 0.0	21.1 ± 0.0	30.6 ± 0.0	31.7 ± 0.0	45.4 ± 0.0	44.2 ± 0.0	47.8 ± 0.0	70.3 ± 0.0	44.7 ± 0.0	29.8 ± 0.0	49.9 ± 0.0	55.4 ± 0.0	39.5 ± 0.0	23.5 ± 0.0
ZeroSiam+COME	22.1 ± 0.0	23.0 ± 0.0	22.0 ± 0.0	19.8 ± 0.0	11.4 ± 0.0	21.5 ± 0.0	25.0 ± 0.0	40.3 ± 0.0	47.0 ± 0.0	34.0 ± 0.0	68.8 ± 0.0	36.3 ± 0.0	18.5 ± 0.0	29.2 ± 0.0	52.6 ± 0.0	31.4 ± 0.0	7.6 ± 0.0
ZeroSiam+FTL	32.5 ± 0.0	34.7 ± 0.0	33.7 ± 0.0	19.5 ± 0.0	20.6 ± 0.0	31.0 ± 0.0	32.5 ± 0.0	46.4 ± 0.0	44.3 ± 0.0	48.2 ± 0.0	70.7 ± 0.0	44.4 ± 0.0	30.5 ± 0.0	50.0 ± 0.0	55.6 ± 0.0	39.6 ± 0.0	6.5 ± 0.0
ZeroSiam+CAREM	34.5 ± 0.0	37.0 ± 0.0	36.1 ± 0.0	20.5 ± 0.0	22.4 ± 0.0	32.5 ± 0.0	33.4 ± 0.0	46.4 ± 0.0	44.9 ± 0.0	49.9 ± 0.0	70.9 ± 0.0	45.9 ± 0.0	31.6 ± 0.0	51.5 ± 0.0	55.8 ± 0.0	40.9 ± 0.0	24.2 ± 0.0
ZeroSiam+CAREM+FTL	35.4 ± 0.0	37.8 ± 0.0	37.0 ± 0.0	20.9 ± 0.0	22.5 ± 0.0	33.6 ± 0.0	34.9 ± 0.0	48.1 ± 0.0	46.0 ± 0.0	51.2 ± 0.0	71.4 ± 0.0	46.2 ± 0.0	33.8 ± 0.0	51.9 ± 0.0	56.1 ± 0.0	41.8 ± 0.0	6.2 ± 0.0

roSiam, while CAREM alone provides less consistent calibration gains. COME once again barely changes the no adapt baseline.

Table 6 shows that CAREM slightly improves every TTA method's accuracy and ECE except DeYO for mixed shifts. CAREM does not prevent Tent or DeYO from collapsing in this scenario. CAREM+FTL recovers both methods and gives EATA the highest accuracy (44.2%), with 16.5% ECE. EATA-C has lower ECE (9.4%) at 41.9% accuracy, while CETA collapses (10.8% accuracy, 83.3% ECE). The COME variants again remain close to the no-adaptation result (31.4% accuracy, 1.4% ECE), so their low ECE here accompanies almost no adaptation.

Table 6: Mixed-shift classification accuracy / expected calibration error (ECE) on ImageNet-C at severity level 5 with ResNet-50-GN. Values are the mean $\pm$ standard deviation across runs.

Method	Original	COME	FTL	CAREM	CAREM+FTL
No Adapt	31.4 ± 0.0 / **1.4** ± 0.0	/	/	/	/
Tent	13.5 ± 0.3 / 79.3 ± 0.3	31.4 ± 0.0 / **1.4** ± 0.0	32.7 ± 0.4 / 23.2 ± 0.8	13.9 ± 0.3 / 79.2 ± 0.5	39.6 ± 0.0 / 15.6 ± 0.0
EATA	38.3 ± 0.1 / 47.5 ± 0.2	31.1 ± 0.0 / 1.7 ± 0.1	41.1 ± 0.0 / 17.4 ± 0.1	39.0 ± 0.1 / 47.0 ± 0.1	**44.2** ± 0.1 / 16.5 ± 0.1
EATA-C	41.9 ± 0.1 / 9.4 ± 0.1	/	/	/	/
CETA	10.8 ± 1.0 / 83.3 ± 1.3	/	/	/	/
SAR	38.5 ± 0.0 / 14.6 ± 0.1	31.4 ± 0.0 / **1.4** ± 0.0	/	41.9 ± 0.0 / 13.4 ± 0.1	/
DeYO	18.1 ± 4.6 / 75.2 ± 5.2	30.9 ± 0.0 / 2.0 ± 0.1	38.7 ± 1.7 / 25.5 ± 2.7	15.7 ± 2.7 / 77.9 ± 3.0	40.7 ± 1.9 / 24.9 ± 2.9
ZeroSiam	31.8 ± 0.1 / 53.7 ± 0.1	31.4 ± 0.0 / **1.4** ± 0.0	38.9 ± 0.0 / 11.4 ± 0.0	32.5 ± 0.2 / 53.2 ± 0.2	39.4 ± 0.0 / 11.6 ± 0.0

D.3 Results on ImageNet-R, ImageNet-A, and VisDA-2021

We additionally evaluate CAREM on ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), and VisDA-2021 (Bashkirova et al., 2022). We keep the same metrics and optimization settings described in Section 5 and Appendix C.

Table 7 shows that CAREM slightly improves the accuracy of all five ViT-Base baselines on ImageNet-A. DeYO+CAREM achieves the highest accuracy (56.7%), while adding FTL lowers its ECE from 14.3% to 10.7% with a small accuracy loss. On ResNet-50-GN, gains are much smaller: DeYO+CAREM+FTL performs best at 17.0% accuracy and 28.7% ECE, compared with 14.2% and 28.8% without adaptation. COME often lowers ECE at the cost of accuracy, and shows the same no adapt behavior for ResNet-50-GN that was previously observed on ImageNet-C.

Table 7: Classification accuracy and expected calibration error (ECE) on ImageNet-A. Values are mean ± standard deviation across runs and are reported as Acc. / ECE. The best result within each model section is bold; the best variant within each method is underlined.

Model	Method	Original	COME	FTL	CAREM	CAREM+FTL
ViT-Base	No Adapt	50.6 ± 0.0 / 8.8 ± 0.0	/	/	/	/
	Tent	52.8 ± 0.2 / 11.7 ± 0.3	52.0 ± 0.2 / 7.3 ± 0.3	52.9 ± 0.1 / 10.3 ± 0.2	53.6 ± 0.2 / 11.0 ± 0.3	53.4 ± 0.2 / 9.7 ± 0.2
	EATA	53.8 ± 0.4 / 11.3 ± 0.3	54.0 ± 0.2 / 7.8 ± 0.3	53.7 ± 0.1 / 10.0 ± 0.3	54.8 ± 0.3 / 10.5 ± 0.1	54.3 ± 0.1 / 9.7 ± 0.2
	EATA-C	51.6 ± 1.0 / 9.5 ± 0.5	/	/	/	/
	CETA	52.9 ± 0.0 / 11.9 ± 0.1	/	/	/	/
	SAR	53.0 ± 0.1 / 10.2 ± 0.2	51.7 ± 0.3 / 8.2 ± 0.1	/	53.5 ± 0.3 / 9.8 ± 0.2	/
	DeYO	55.4 ± 0.4 / 15.1 ± 0.3	54.9 ± 0.2 / 15.2 ± 0.4	55.3 ± 0.2 / 11.6 ± 0.1	**56.7** ± 0.4 / 14.3 ± 0.3	56.2 ± 0.2 / 10.7 ± 0.2
	ZeroSiam	52.7 ± 0.1 / 11.6 ± 0.2	51.7 ± 0.2 / **7.2** ± 0.2	52.7 ± 0.1 / 10.1 ± 0.1	53.1 ± 0.2 / 11.1 ± 0.2	53.0 ± 0.1 / 10.1 ± 0.2
ResNet-50-GN	No Adapt	14.2 ± 0.0 / 28.8 ± 0.0	/	/	/	/
	Tent	14.4 ± 0.1 / 33.3 ± 0.0	14.2 ± 0.0 / 28.8 ± 0.0	14.7 ± 0.1 / 29.0 ± 0.1	14.5 ± 0.1 / 33.2 ± 0.1	14.7 ± 0.1 / 29.0 ± 0.1
	EATA	14.9 ± 0.1 / 33.2 ± 0.1	14.2 ± 0.0 / 28.8 ± 0.0	15.1 ± 0.2 / 29.1 ± 0.1	15.1 ± 0.1 / 33.0 ± 0.2	15.3 ± 0.2 / 29.0 ± 0.1
	EATA-C	14.6 ± 0.3 / 29.2 ± 0.2	/	/	/	/
	CETA	14.4 ± 0.0 / 33.6 ± 0.0	/	/	/	/
	SAR	14.7 ± 0.1 / 29.0 ± 0.1	14.2 ± 0.0 / 28.8 ± 0.0	/	14.8 ± 0.1 / 29.0 ± 0.1	/
	DeYO	15.6 ± 0.2 / 45.6 ± 0.3	14.2 ± 0.0 / 28.8 ± 0.0	16.8 ± 0.3 / 28.8 ± 0.2	16.2 ± 0.2 / 44.9 ± 0.3	**17.0** ± 0.4 / **28.7** ± 0.3
	ZeroSiam	14.4 ± 0.0 / 33.1 ± 0.0	14.2 ± 0.0 / 28.8 ± 0.0	14.6 ± 0.1 / 29.1 ± 0.1	14.5 ± 0.1 / 33.0 ± 0.1	14.7 ± 0.1 / 29.1 ± 0.1

Table 8 shows that CAREM raises accuracy for every ViT-Base baseline on ImageNet-R. DeYO+CAREM+FTL has the highest accuracy (69.9%, 5.3% ECE), while EATA+CAREM+FTL reaches 66.9% accuracy with lower ECE (2.9%); EATA-C achieves 69.1% and 4.1%. On ResNet-50-GN, DeYO+CAREM+FTL also performs best in accuracy (47.4%, 7.0% ECE), just ahead of EATA-C (46.5%, 7.2%). FTL is responsible for the ECE improvements.

Table 8: Classification accuracy and expected calibration error (ECE) on ImageNet-R. Values are mean ± standard deviation across runs and are reported as Acc. / ECE. The best result within each model section is bold; the best variant within each method is underlined.

Model	Method	Original	COME	FTL	CAREM	CAREM+FTL
ViT-Base	No Adapt	58.9 ± 0.0 / **2.5** ± 0.0	/	/	/	/
	Tent	63.5 ± 0.0 / 7.3 ± 0.1	61.2 ± 0.1 / 14.2 ± 0.1	63.9 ± 0.1 / 3.6 ± 0.2	65.0 ± 0.1 / 6.7 ± 0.1	65.0 ± 0.1 / 3.3 ± 0.1
	EATA	64.8 ± 0.1 / 5.6 ± 0.1	65.4 ± 0.1 / 6.3 ± 0.1	64.7 ± 0.2 / 3.4 ± 0.1	67.4 ± 0.1 / 5.0 ± 0.1	66.9 ± 0.1 / 2.9 ± 0.1
	EATA-C	69.1 ± 0.1 / 4.1 ± 0.1	/	/	/	/
	CETA	63.7 ± 0.1 / 7.8 ± 0.1	/	/	/	/
	SAR	64.3 ± 0.1 / 3.4 ± 0.1	61.8 ± 0.0 / 7.3 ± 0.0	/	65.4 ± 0.1 / 3.2 ± 0.1	/
	DeYO	68.5 ± 0.1 / 11.5 ± 0.1	66.3 ± 0.2 / 23.1 ± 0.4	68.2 ± 0.1 / 6.2 ± 0.2	69.6 ± 0.1 / 11.4 ± 0.1	**69.9** ± 0.1 / 5.3 ± 0.1
	ZeroSiam	63.6 ± 0.0 / 5.9 ± 0.0	61.1 ± 0.1 / 14.7 ± 0.1	63.8 ± 0.1 / 3.0 ± 0.1	64.1 ± 0.0 / 5.8 ± 0.1	64.2 ± 0.0 / 3.0 ± 0.1
ResNet-50-GN	No Adapt	40.3 ± 0.0 / **5.5** ± 0.0	/	/	/	/
	Tent	40.7 ± 0.0 / 21.9 ± 0.1	40.3 ± 0.0 / **5.5** ± 0.0	42.4 ± 0.0 / 6.1 ± 0.1	40.8 ± 0.0 / 21.8 ± 0.1	42.4 ± 0.0 / 6.1 ± 0.1
	EATA	43.9 ± 0.2 / 12.7 ± 0.2	40.2 ± 0.0 / **5.5** ± 0.0	43.7 ± 0.3 / 6.3 ± 0.1	44.4 ± 0.2 / 12.5 ± 0.2	44.3 ± 0.2 / 6.3 ± 0.2
	EATA-C	46.5 ± 0.2 / 7.2 ± 0.2	/	/	/	/
	CETA	40.5 ± 0.0 / 22.9 ± 0.1	/	/	/	/
	SAR	42.0 ± 0.1 / 5.9 ± 0.0	40.3 ± 0.0 / **5.5** ± 0.0	/	42.1 ± 0.1 / 5.9 ± 0.1	/
	DeYO	40.3 ± 0.7 / 41.2 ± 1.0	40.3 ± 0.0 / **5.5** ± 0.0	47.0 ± 0.1 / 7.1 ± 0.2	41.2 ± 0.5 / 40.1 ± 0.7	**47.4** ± 0.0 / 7.0 ± 0.1
	ZeroSiam	42.1 ± 0.1 / 15.9 ± 0.1	40.3 ± 0.0 / **5.5** ± 0.0	42.6 ± 0.1 / 5.8 ± 0.1	42.4 ± 0.1 / 16.4 ± 0.1	43.1 ± 0.1 / 5.8 ± 0.1

Table 9: Classification accuracy and expected calibration error (ECE) on VisDA-2021. Values are mean ± standard deviation across runs and are reported as Acc. / ECE. The best result within each model section is bold; the best variant within each method is underlined.

Model	Method	Original	COME	FTL	CAREM	CAREM+FTL
ViT-Base	No Adapt	55.0 ± 0.0 / **3.9** ± 0.0	/	/	/	/
	Tent	56.7 ± 0.0 / 15.0 ± 0.1	56.0 ± 0.1 / 23.4 ± 0.1	56.8 ± 0.1 / 9.3 ± 0.1	58.7 ± 0.1 / 13.2 ± 0.1	58.7 ± 0.0 / 6.0 ± 0.1
	EATA	57.9 ± 0.2 / 16.3 ± 0.2	57.8 ± 0.2 / 9.7 ± 0.1	55.9 ± 1.2 / 12.5 ± 0.6	59.3 ± 0.3 / 15.2 ± 0.2	59.1 ± 0.3 / 7.0 ± 0.3
	EATA-C	59.2 ± 0.3 / 7.2 ± 0.2	/	/	/	/
	CETA	56.8 ± 0.0 / 15.8 ± 0.1	/	/	/	/
	SAR	56.8 ± 0.1 / 9.3 ± 0.1	56.6 ± 0.2 / 13.2 ± 0.1	/	58.9 ± 0.1 / 5.8 ± 0.1	/
	DeYO	59.1 ± 0.4 / 21.1 ± 0.4	58.0 ± 0.1 / 31.3 ± 0.8	58.2 ± 0.2 / 14.1 ± 0.2	59.1 ± 0.2 / 22.9 ± 0.3	**60.2** ± 0.2 / 9.7 ± 0.2
	ZeroSiam	56.4 ± 0.1 / 14.0 ± 0.1	54.2 ± 0.5 / 31.4 ± 0.2	56.6 ± 0.1 / 8.7 ± 0.1	57.4 ± 0.2 / 13.2 ± 0.1	57.6 ± 0.2 / 7.5 ± 0.1
ResNet-50-GN	No Adapt	39.7 ± 0.0 / **6.9** ± 0.0	/	/	/	/
	Tent	38.8 ± 0.1 / 30.0 ± 0.1	39.7 ± 0.0 / **6.9** ± 0.0	40.2 ± 0.1 / 8.7 ± 0.1	39.3 ± 0.0 / 29.6 ± 0.0	40.5 ± 0.1 / 8.9 ± 0.1
	EATA	40.5 ± 0.1 / 27.9 ± 0.5	39.7 ± 0.0 / **6.9** ± 0.0	40.6 ± 0.2 / 9.7 ± 0.1	41.6 ± 0.1 / 26.8 ± 0.2	41.7 ± 0.1 / 9.7 ± 0.1
	EATA-C	41.6 ± 0.1 / 8.3 ± 0.0	/	/	/	/
	CETA	38.6 ± 0.0 / 31.7 ± 0.1	/	/	/	/
	SAR	40.2 ± 0.1 / 8.6 ± 0.0	39.7 ± 0.0 / **6.9** ± 0.0	/	40.8 ± 0.0 / 8.7 ± 0.1	/
	DeYO	36.4 ± 0.3 / 47.9 ± 0.4	39.7 ± 0.0 / **6.9** ± 0.0	41.6 ± 0.0 / 10.3 ± 0.1	37.4 ± 0.6 / 46.7 ± 0.7	**42.5** ± 0.2 / 10.3 ± 0.2
	ZeroSiam	39.3 ± 0.0 / 27.2 ± 0.0	39.7 ± 0.0 / **6.9** ± 0.0	40.3 ± 0.1 / 8.5 ± 0.1	39.6 ± 0.0 / 27.6 ± 0.1	40.5 ± 0.1 / 8.7 ± 0.1

Table 9 extends the comparison to VisDA-2021. With ViT-Base, DeYO+CAREM+FTL reaches the highest accuracy (60.2%, 9.7% ECE), while SAR+CAREM has the lowest ECE among TTA methods (5.8%) at 58.9% accuracy. For ResNet-50-GN, DeYO+CAREM+FTL again leads in accuracy (42.5%, 10.3% ECE), and EATA+CAREM+FTL achieves 41.7% with 9.7% ECE. CAREM alone generally raises accuracy but does not reduce ECE; FTL is responsible for the ECE reduction.

E Accuracy and Calibration Dynamics Under Mixed Corruption Shifts

The aggregate mixed-shift results in Table 3 hide changes during the continuous stream. We examine the ZeroSiam variants' accuracy gains and confidence trajectories in Figures 6 and 7.

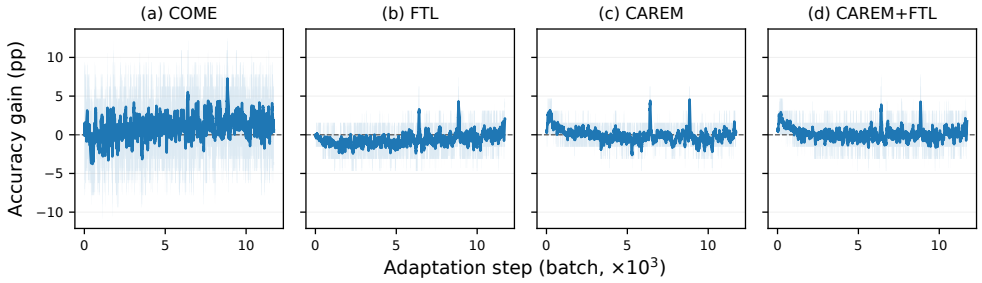


Figure 6: Top-1 accuracy difference from ZeroSiam for its COME, FTL, CAREM, and CAREM+FTL variants under mixed shifts. Curves use a 50-step moving window and IQR shading.

Figure 6 shows that CAREM initially improves ZeroSiam's accuracy by about 2–3 percentage points, but the gain fades as adaptation continues. This explains their similar aggregate accuracies (63.4% and 63.5%). CAREM+FTL also loses its early gain, yet avoids much of FTL's accuracy cost. COME's gains are very variable, but mostly positive, yielding 64.5% aggregate accuracy.

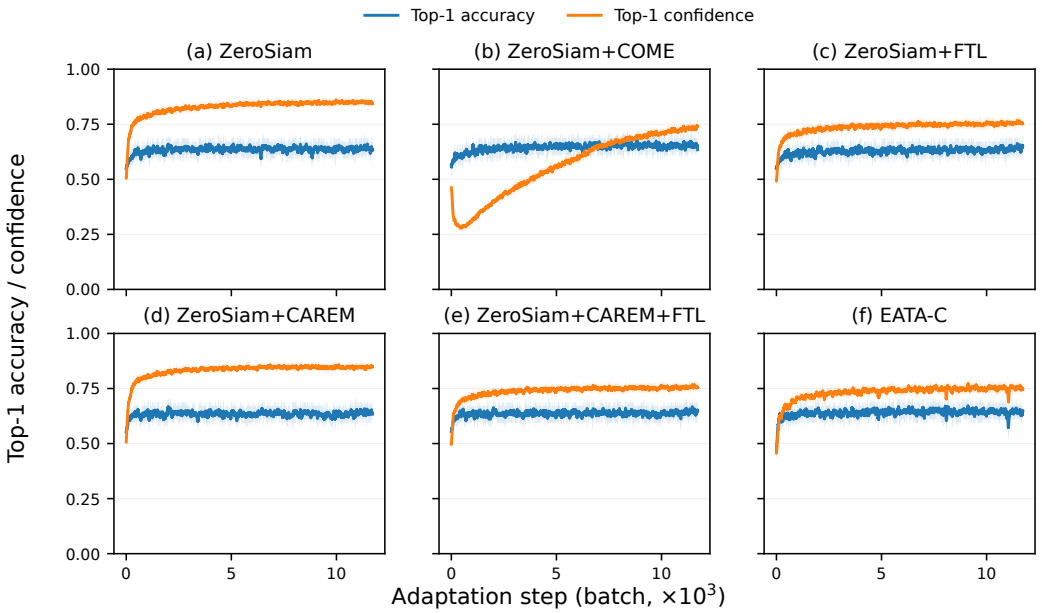


Figure 7: Top-1 accuracy and confidence for EATA-C and ZeroSiam and its variants under mixed shifts. Curves use a 50-step moving window and IQR shading.

Table 10: Results across the 15 ImageNet-C corruptions at severity level 5 using SAR+CAREM with ViT-Base and different gradient surgery rules. Results are reported as mean ± standard deviation over four runs with different random seeds. Accuracy and ECE are reported as percentages. For each metric, the best mean is shown in bold and the second-best mean is underlined.

Gradient combination	**Acc.** ↑	**ECE** ↓
Naive sum (SAR)	63.7 ± 0.2	<u>7.2 ± 0.0</u>
Tangential only	66.3 ± 2.6	23.3 ± 1.7
Cosine gating	66.3 ± 2.6	21.4 ± 2.5
Always orthogonalize aux.	<u>66.8 ± 2.5</u>	**3.5** ± 0.2
Summed-update projection	63.8 ± 0.2	<u>7.2 ± 0.2</u>
Min-change aux. projection	**66.9** ± 2.6	**3.5** ± 0.2

Figure 7 shows that ZeroSiam and ZeroSiam+CAREM become strongly overconfident. FTL reduces this gap, although CAREM+FTL is still overconfident, similarly to EATA-C. ZeroSiam+COME instead starts severely underconfident and becomes overconfident later. Its aggregate 8.2% ECE masks this reversal: pooling all the predictions into confidence bins allows oppositely signed errors within a bin to partly cancel. Its low aggregate ECE therefore does not imply consistent calibration over time and is rather an artifact of the metric.

F Comparison of Different Gradient Surgery Rules

We empirically compare four hyperparameter-free gradient surgery rules for combining the two parameter gradients to quantify the influence of each rule on CAREM's performance. Let $\boldsymbol{g}_{\text{main}} = \boldsymbol{g}_{\boldsymbol{\theta},\text{tan}}$, $\boldsymbol{g}_{\text{aux}} = \boldsymbol{g}_{\boldsymbol{\theta},\text{rad}}$, and $c = \boldsymbol{g}_{\text{main}}^{\top}\boldsymbol{g}_{\text{aux}}$. The resulting direction $\boldsymbol{g}_{\text{comb}}$ is used in the parameter update $\boldsymbol{\theta}^{+} = \boldsymbol{\theta} - \eta\boldsymbol{g}_{\text{comb}}$. We report the naive sum and tangential-only update as references. If $\|\boldsymbol{g}_{\text{main}}\|_2 = 0$, no main direction can be protected and the auxiliary gradient is left unchanged.

Cosine gating retains $\boldsymbol{g}_{\text{aux}}$ only when it agrees with $\boldsymbol{g}_{\text{main}}$ (Du et al., 2018):

$$\boldsymbol{g}_{\text{comb}} = \begin{cases} \boldsymbol{g}_{\text{main}} + \boldsymbol{g}_{\text{aux}}, & c \geq 0, \\ \boldsymbol{g}_{\text{main}}, & c < 0. \end{cases} \tag{18}$$

This protects the main direction but discards the complete auxiliary signal during conflict, including components orthogonal to the main gradient.

Always orthogonalizing the auxiliary gradient follows the projection used by Bloop (Hsieh et al., 2024) and removes the auxiliary component parallel to $\boldsymbol{g}_{\text{main}}$ at every step:

$$\boldsymbol{g}_{\text{comb}} = \boldsymbol{g}_{\text{main}} + \boldsymbol{g}_{\text{aux}} - \frac{c}{\|\boldsymbol{g}_{\text{main}}\|_2^2}\boldsymbol{g}_{\text{main}}. \tag{19}$$

It preserves the first-order progress of $\boldsymbol{g}_{\text{main}}$ exactly, but also removes constructive alignment.

Projecting the summed update adapts the minimum-ℓ_2 half-space projection of A-GEM (Chaudhry et al., 2019). It starts from $\boldsymbol{g}_0 = \boldsymbol{g}_{\text{main}} + \boldsymbol{g}_{\text{aux}}$ and finds its closest direction that is nonconflicting with $\boldsymbol{g}_{\text{main}}$:

$$\boldsymbol{g}_{\text{comb}} = \begin{cases} \boldsymbol{g}_0 - \dfrac{\boldsymbol{g}_{\text{main}}^{\top}\boldsymbol{g}_0}{\|\boldsymbol{g}_{\text{main}}\|_2^2}\boldsymbol{g}_{\text{main}}, & \boldsymbol{g}_{\text{main}}^{\top}\boldsymbol{g}_0 < 0, \\ \boldsymbol{g}_0, & \text{otherwise.} \end{cases} \tag{20}$$

This enforces only $\boldsymbol{g}_{\text{main}}^{\top}\boldsymbol{g}_{\text{comb}} \geq 0$ and can therefore reduce the first-order main progress to zero.

Minimum-change auxiliary projection is explained in Section 4.4.

Table 10 shows that always orthogonalizing the auxiliary gradient and CAREM both reach 3.5% ECE, with accuracies of 66.8% and 66.9%, respectively. These similar results are expected as Figure 4 shows that over 99% of the batches have a negative cosine similarity. Tangential-only updates and cosine gating improve accuracy over the naive sum but have much higher ECE, while projecting the summed update gives almost no improvement. Therefore, always orthogonalizing and the minimum-change auxiliary projection are both adequate gradient surgery rules for CAREM.