FAILURE-AWARE TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

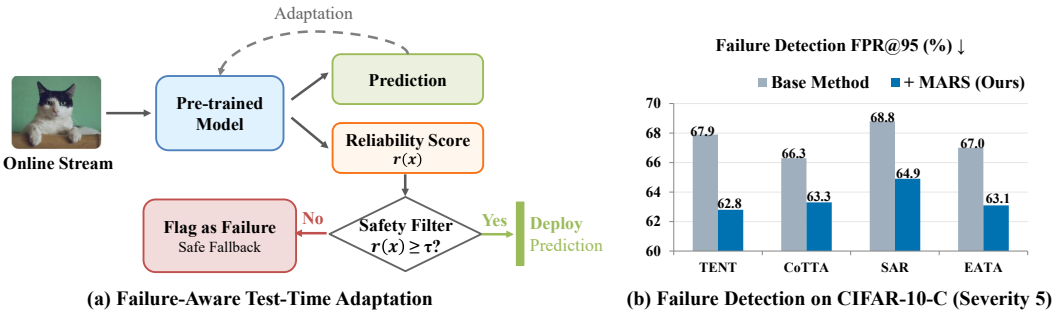
Test-time adaptation (TTA) mitigates distribution shifts by adapting pre-trained models to unlabeled target domains during deployment. While existing methods successfully boost target-domain accuracy via entropy minimization or self-training, they optimize for predictive performance while largely ignoring the reliability of the model's confidence estimates. Consequently, under severe domain shifts, these adaptation mechanisms break down the capacity for failure detection, causing models to become highly overconfident in their misclassifications. To overcome this limitation, we introduce Masked Augmentation for Reliable Self-adaptation (**MARS**), a plug-and-play framework for **failure-aware TTA** designed to simultaneously maximize adaptation accuracy and failure detection reliability. MARS leverages masked augmentations of input samples to serve two complementary objectives: (1) generating masking-ratio-conditioned soft labels to calibrate confidence rankings, and (2) measuring prediction consistency between original and masked samples to yield a robust uncertainty signal. Crucially, this dual-mechanism integrates seamlessly with existing TTA baselines without architectural modifications, supported by rigorous theoretical analysis. Across standard, continual, and mixed-domain setups on diverse benchmarks, MARS substantially improves failure detection metrics while matching or exceeding the target-domain accuracy of state-of-the-art methods. Our code will be made publicly available.

1 INTRODUCTION

Deep neural networks deployed in real-world environments frequently encounter test inputs drawn from distributions that differ from their training data. Test-time adaptation (TTA) addresses this shift by updating a subset of model parameters directly on unlabeled test streams, typically via entropy minimization or pseudo-label self-training (Wang et al., 2021; Niu et al., 2022; 2023; Wang et al., 2022). While prior work has predominantly focused on improving adaptation accuracy, a critical question remains underexplored: *when adaptation fails on individual samples, can the model reliably recognize its own errors?* This capability, known as failure detection (FD) (Hendrycks & Gimpel, 2016; Zhu et al., 2023), is essential in practical deployments where models must defer uncertain predictions, request human intervention, or trigger safe fallback mechanisms. In safety-critical applications like autonomous driving or medical imaging, an adapted model that makes highly confident mistakes is arguably more dangerous than an unadapted model with well-calibrated uncertainty. Motivated by this gap, we introduce *failure-aware TTA*: a dual objective that simultaneously maximizes target-domain accuracy and the reliable detection of prediction failures.

Pursuing this dual objective, however, reveals a critical vulnerability in current methodologies. Our empirical analysis demonstrates that while standard TTA methods effectively improve prediction accuracy, they struggle to reliably recognize their own prediction errors (Figure 1 (b)). Adaptation mechanisms based on entropy minimization, such as TENT (Wang et al., 2021), explicitly push all predictions toward highly confident, one-hot distributions, regardless of their actual correctness. Consequently, standard failure detection metrics like maximum softmax probability (MSP) (Hendrycks & Gimpel, 2016) and entropy (Shannon, 1948) become severely miscalibrated after adaptation. Because both correct and incorrect predictions collapse toward similarly low-entropy states, these metrics lose their power to distinguish failures from successes. This degradation is further exacerbated in continual and mixed-domain TTA settings, where confidence collapse accumulates across sequential domain shifts.

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.



(a) Failure-Aware Test-Time Adaptation (b) Failure Detection on CIFAR-10-C (Severity 5)

Figure 1: (a) Failure-aware test-time adaptation aims to jointly maximize target-domain accuracy while reliably detecting prediction failures. (b) The proposed MARS framework consistently enhances the failure detection capability of existing TTA methods.

To overcome this fundamental limitation, we propose Masked Augmentation for Reliable Self-adaptation (**MARS**), a plug-and-play framework designed to directly satisfy the failure-aware TTA objective. During adaptation, MARS applies multiple masked augmentations to the input, serving two complementary objectives. First, it generates soft pseudo-targets conditioned on masking ratios, adapting the model to preserve the ranking order of prediction confidence across these masked views. By structurally assigning lower confidence to augmentations with higher masking ratios, MARS explicitly enforces better calibration. Second, it measures prediction consistency between the original and masked views to produce a robust uncertainty score complementary to standard entropy. Because unreliable predictions naturally exhibit substantially higher variance across masked views than reliable ones, this consistency signal is highly effective for failure detection. Crucially, MARS integrates seamlessly into existing TTA baselines without architectural modifications and is supported by rigorous theoretical analysis. Extensive evaluations across four benchmarks (CIFAR-10-C, CIFAR-100-C, Tiny-ImageNet-C, and ImageNet-C), three adaptation protocols (standard, continual, and mixed-domain), and five baseline methods (TENT, CoTTA, EATA, SAR, and TPT) demonstrate that MARS consistently achieves superior target-domain accuracy while enhancing failure detection reliability. Our contributions are summarized as follows:

- We formalize *failure-aware TTA* as a dual objective: maximizing target-domain accuracy while maintaining reliable discrimination between correct and incorrect predictions. We demonstrate that while existing TTA methods often improve the former, they struggle with the latter.
- We introduce MARS, a versatile framework that leverages masked augmentations to jointly calibrate confidence rankings and extract a robust uncertainty signal. It integrates readily with existing TTA baselines without requiring architectural changes, supported by theoretical motivation analysis.
- Comprehensive evaluations across four benchmarks, three protocols, three corruption severities, and five baselines confirm that MARS consistently enhances failure detection metrics while matching or exceeding state-of-the-art prediction accuracy.

2 Related Work

Test-time Adaptation adapts a pretrained model to unlabeled target data on the fly, typically via entropy minimization or self-training. For instance, TENT (Wang et al., 2021) minimizes prediction entropy by updating the affine parameters of batch normalization layers. EATA (Niu et al., 2022) improves upon this by filtering for reliable, non-redundant samples and introducing a Fisher-regularized anti-forgetting term. To address stability under severe distribution shifts, SAR (Niu et al., 2023) incorporates sharpness-aware minimization. For continual TTA, CoTTA (Wang et al., 2022) employs a teacher-student architecture with weight-averaged pseudo-labels and stochastic restoration. Notably, these approaches evaluate success almost exclusively based on target-domain accuracy. While a recent line of work expands TTA to open-set settings (Lee et al., 2023; Li et al., 2023; Gao et al., 2024), our focus is fundamentally complementary: we explicitly target the failure modes that arise within known classes, a regime where existing methods often yield high accuracy but severely miscalibrated confidence.

Failure Detection. Foundational work on failure prediction originates from selective classification, which formalizes the trade-off between prediction accuracy and abstention under uncertainty (Chow, 1970). In the deep learning era, this concept has re-emerged as failure detection, where the goal is to identify test samples on which a model is likely to make an incorrect prediction (Qiu & Miikkulainen, 2022; Dong et al., 2025). While thresholding the maximum softmax probability (Hendrycks & Gimpel, 2016) remains a ubiquitous baseline, its susceptibility to overconfidence severely limits its reliability (Pinto et al., 2022; Qiu & Miikkulainen, 2022). Beyond MSP, a large body of work aims to estimate prediction risk more effectively. One stream of research focuses on training auxiliary modules, such as dedicated classification heads or independent networks, to explicitly predict correctness based on intermediate features or prediction logits (Corbière et al., 2019; Jiang et al., 2018; Granese et al., 2021). A distinct line of research adopts an integrated approach, jointly optimizing failure detection capabilities alongside the primary model's training objective (Luo et al., 2021; Moon et al., 2020; Liu et al., 2024; 2026). However, these approaches are inherently designed for the source training phase and are not directly applicable to test-time adaptation, where models must operate on unlabeled target streams without access to ground-truth supervision.

3 METHODOLOGY

3.1 FAILURE-AWARE TEST-TIME ADAPTATION

We consider a scenario where a source-pretrained model $f_\theta : \mathcal{X} \to \mathbb{R}^C$, parameterized by θ, receives a continuous stream of unlabeled target domain data samples $\mathbf{x} \sim \mathcal{D}_t$. Conventional TTA methods focus on improving classification accuracy by adapting model parameters via:

$$\theta_t = \min_\theta \mathcal{L}_{\text{tta}}(\mathbf{x}; \theta), \tag{1}$$

where $\mathcal{L}_{\text{tta}}$ denotes the adaptation objective (e.g., entropy minimization or self-training). However, classification accuracy alone is insufficient for safety-critical systems. A reliable model must additionally estimate the likelihood of its own errors, enabling it to defer highly uncertain predictions to a human operator. *Failure-aware TTA* requires a per-sample reliability score $r(\mathbf{x}) \in \mathbb{R}$ such that a decision function $G(\mathbf{x})$ can detect failures based on a threshold τ:

$$G(\mathbf{x}) = \begin{cases} \text{correct} & \text{if } r(\mathbf{x}) \geq \tau, \\ \text{misclassified} & \text{otherwise.} \end{cases} \tag{2}$$

Ideally, for the adapted model f_{θ_t}, we expect the reliability score $r(\mathbf{x})$ to be larger for correctly classified samples than for misclassified ones.

3.2 THE MARS FRAMEWORK

Entropy minimization forces predictions to become highly overconfident during adaptation. While this reduces uncertainty when predictions are correct, it creates a confidence collapse when the model adapts toward an incorrect class: the prediction fails, yet the entropy remains deceptively low. Consequently, under imperfect adaptation, entropy loses its discriminative capacity to separate correct and misclassified samples, critically impairing failure detection. To address this, we propose Masked Augmentation for Reliable Self-adaptation (**MARS**). MARS leverages masked augmentations to directly counteract confidence collapse via two complementary objectives (Figure 2): (1) generating soft-label pseudo-targets conditioned on masking ratios to explicitly calibrate confidence rankings, and (2) measuring prediction consistency between original and masked samples to produce a robust uncertainty signal that compensates for the flaws of standard entropy. Crucially, MARS achieves this dual objective while integrating seamlessly with existing TTA baselines without architectural modifications.

3.2.1 MASKED AUGMENTATIONS GENERATION

To generate masked augmentations, we apply a random block-wise masking strategy to the input images. Given an input image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ and a target masking ratio $\rho \in (0, 1)$, we partition the spatial domain into non-overlapping patches of size $P \times P$. The total number of patches to mask is determined by $N = \lfloor (H \cdot W \cdot \rho)/P^2 \rfloor$. We sample N patches uniformly at random without

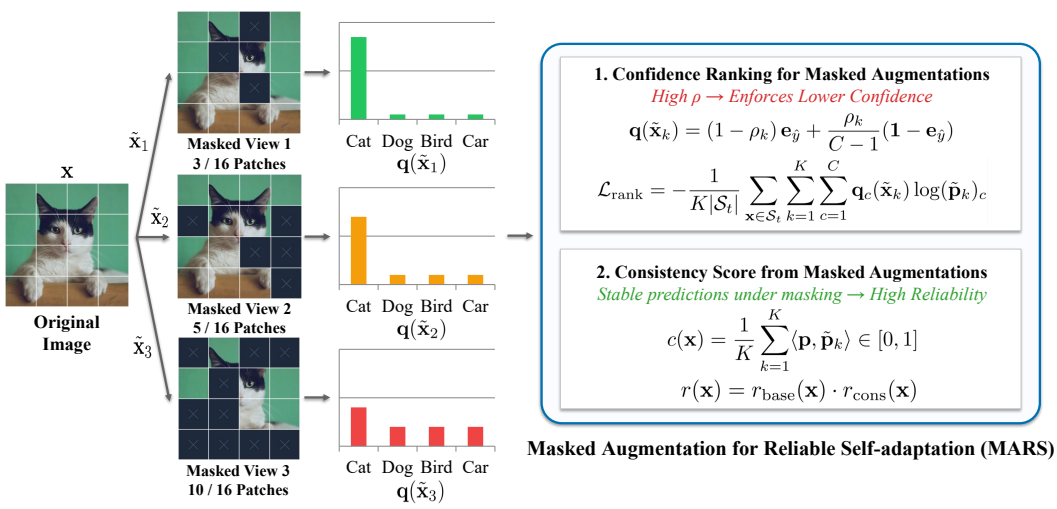


Figure 2: Overview of the proposed MARS framework.

replacement. A binary mask $\mathbf{M}$ is first constructed at the coarse patch resolution $\lfloor H/P \rfloor \times \lfloor W/P \rfloor$, where ones designate the patches to be dropped. This coarse mask is then spatially upsampled to the original image dimensions via element repetition. To accommodate spatial dimensions that are not perfectly divisible by P, we apply zero-padding to the bottom and right edges of the mask. The final augmented image is computed via element-wise multiplication $\tilde{\mathbf{x}} = \mathbf{x} \odot (\mathbf{1} - \mathbf{M})$, effectively zeroing out the randomly selected spatial regions.

For each input image $\mathbf{x}$, we sample K independent masking ratios $\rho_k \sim \mathcal{U}(r_{\min}, r_{\max})$ and generate corresponding masked augmentations $\tilde{\mathbf{x}}_k$ as described above. The current model f_{θ_t} then produces predictions for both the original input and masked augmentations, denoted as $\mathbf{p} = f_{\theta_t}(\mathbf{x})$ and $\tilde{\mathbf{p}}_k = f_{\theta_t}(\tilde{\mathbf{x}}_k)$, respectively.

3.2.2 Confidence Ranking for Masked Augmentations

We first generate soft pseudo-targets conditioned on the masking ratios, encouraging the model to preserve the ranking of prediction confidence. Specifically, heavily masked augmentations are assigned lower target confidence, which explicitly penalizes overconfidence on degraded inputs.

Let $\mathcal{S}_t$ denote the subset of samples for adaptation at step t. For each original sample $\mathbf{x} \in \mathcal{S}_t$, we generate K masked augmentations. For a given augmentation $\tilde{\mathbf{x}}_k$ with masking ratio ρ_k, we construct its soft pseudo-target $\mathbf{q}(\tilde{\mathbf{x}}_k)$ as:

$$\mathbf{q}(\tilde{\mathbf{x}}_k) = (1 - \rho_k)\, \mathbf{e}_{\hat{y}} + \frac{\rho_k}{C-1}(\mathbf{1} - \mathbf{e}_{\hat{y}}), \tag{3}$$

where $\hat{y} = \arg\max \mathbf{p}$ is the pseudo-label from $\mathbf{x}$, $\mathbf{e}_{\hat{y}}$ is the one-hot encoding of $\hat{y}$, and C is the total number of classes. Equation equation 3 smoothly interpolates from a one-hot distribution (when $\rho = 0$) to a uniform distribution (as ρ approaches $1 - 1/C$). This formalizes the intuition that as a larger portion of the input is masked, the model's certainty should naturally decrease. The **confidence ranking loss** is then defined as the cross-entropy between these soft pseudo-targets and the model's predictions $\tilde{\mathbf{p}}_k$ for the masked augmentations:

$$\mathcal{L}_{\text{rank}} = -\frac{1}{K|\mathcal{S}_t|} \sum_{\mathbf{x} \in \mathcal{S}_t} \sum_{k=1}^{K} \sum_{c=1}^{C} \mathbf{q}_c(\tilde{\mathbf{x}}_k) \log(\tilde{\mathbf{p}}_k)_c. \tag{4}$$

Assuming the base TTA method optimizes an intrinsic objective $\mathcal{L}_{\text{tta}}$ (e.g., entropy minimization), MARS integrates the confidence ranking loss using a single trade-off hyperparameter λ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{tta}} + \lambda\, \mathcal{L}_{\text{rank}}. \tag{5}$$

We apply $\mathcal{L}_{\text{rank}}$ only to reliable samples, where predictions are likely to be correct. For TTA methods that include built-in sample filtering (e.g., SAR and EATA), we directly use their selected samples. Otherwise, we identify reliable samples using a simple entropy-based filtering strategy.

3.2.3 CONSISTENCY SCORE FROM MASKED AUGMENTATIONS_ORIGINAL

We then measure prediction consistency between the original and masked samples to yield a robust uncertainty metric that is complementary to entropy. Unreliable predictions inherently exhibit substantially higher variance across masked views than reliable ones, making this signal particularly effective for failure detection under distribution shifts. We define soft consistency as the average dot product between the prediction distributions of the unmasked and masked views:

$$c(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K}\langle \mathbf{p}, \tilde{\mathbf{p}}_k \rangle \in [0,1]. \tag{6}$$

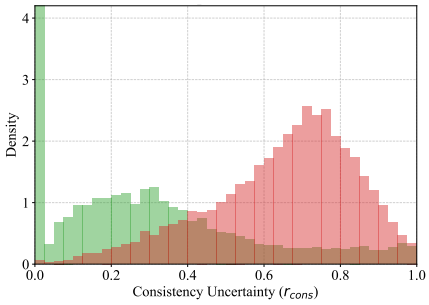


Figure 3: Clear separation between correct (green) and failure (red) samples based on the consistency score r_{cons}.

A high $c(\mathbf{x})$ indicates that the masked views strongly agree with the unmasked prediction, whereas a low $c(\mathbf{x})$ suggests that small input perturbations destabilize the model. We define the corresponding consistency score directly as $r_{\text{cons}}(\mathbf{x}) = 1 - c(\mathbf{x})$. Figure 3 verifies the effectiveness of r_{cons} in separating correct and incorrect predictions. We then combine r_{cons} with a base reliability score r_{base} derived from the unmasked logits, which can be instantiated as MSP ($\max \mathbf{p}$), negative Shannon entropy ($-H(\mathbf{p})$), top-2 margin ($\mathbf{p}_{(1)} - \mathbf{p}_{(2)}$), or any standard metric. We use negative Shannon entropy as our default. The final reliability score is the product of these two components:

$$r(\mathbf{x}) = r_{\text{base}}(\mathbf{x}) \cdot r_{\text{cons}}(\mathbf{x}) = -H(\mathbf{p}) \cdot (1 - c(\mathbf{x})). \tag{7}$$

We explicitly choose multiplicative over additive fusion because the two metrics are conceptually conjunctive: a prediction is reliable only if it is both confident and stable. Additive fusion would incorrectly allow a highly confident but unstable prediction to compensate for its instability. Multiplicative fusion prevents this, ensuring a strict requirement for both criteria.

4 THEORETICAL INSIGHTS

To provide a theoretical motivation for the MARS framework, we characterize its core properties through two theoretical results, and the detailed proofs are in Appendix A.

Setup. Let X denote the unmasked input, Y the ground-truth label, and $\tilde{X} = X \odot (1 - M)$ the masked view with masking ratio $\rho \in [0, 1 - 1/C]$. I is the mutual information and H is the entropy.

Proposition 1. *For any masking ratios $\rho_1 < \rho_2$, assuming label-relevant information is spatially distributed across at least two patches, then $I(\tilde{X}^{(\rho_2)}; Y) < I(\tilde{X}^{(\rho_1)}; Y)$ and $H(Y|\tilde{X}^{(\rho_2)}) > H(Y|\tilde{X}^{(\rho_1)})$.*

Takeaway. Masking strictly removes label-relevant information, causing the true posterior entropy to increase monotonically with ρ. This defines the ground-truth uncertainty pattern that a well-calibrated model is expected to exhibit.

Proposition 2. *At the global minimum of $\mathcal{L}_{\text{rank}}$, the model's predictive entropy on a masked view with ratio ρ is given by $H(\tilde{\mathbf{p}}) = -(1-\rho)\log(1-\rho) - \rho \log \frac{\rho}{C-1}$. This function is strictly increasing with respect to ρ on $[0, 1 - 1/C]$, perfectly matching the entropy ordering dictated by Proposition 1.*

Takeaway. MARS ensures that the model's predictive entropy is structurally aligned with the true posterior entropy across different masking ratios. This property provides a theoretical explanation for why MARS mitigates the entropy collapse commonly observed in standard adaptation methods.

Summary. Proposition 1 shows that masking inherently reduces mutual information, which in turn requires a corresponding increase in posterior entropy. MARS enforces this property through the confidence ranking loss $\mathcal{L}_{\text{rank}}$. As established in Proposition 2, the global optimum of this objective explicitly constrains the model's predictive entropy to follow the strict monotonic increase dictated by Proposition 1. Therefore, MARS ensures that the adapted model cannot falsely show certainty when critical semantic features are obscured.

Table 1: Failure-aware test-time adaptation performance under **standard** setup.

Method	CIFAR-10-C				CIFAR-100-C				Tiny-ImageNet-C			
	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑
Source	227.17	77.13	81.97	56.48	224.09	81.45	73.36	53.55	582.71	77.84	76.62	22.32
TENT	57.96	84.02	67.86	81.54	113.39	84.75	66.32	68.39	366.96	79.48	74.70	39.48
+MARS	**46.57**	**86.07**	**62.82**	**82.52**	**102.14**	**86.02**	**64.36**	**69.16**	**351.03**	**82.02**	**71.20**	39.46
CoTTA	49.24	85.94	66.28	81.54	128.15	84.66	67.47	65.34	371.49	80.90	72.97	37.71
+MARS	50.08	**86.35**	**63.32**	**81.55**	**125.61**	**85.78**	**64.52**	**65.44**	**368.29**	**81.51**	**72.77**	**37.74**
SAR	62.25	84.52	68.75	79.46	112.14	85.90	64.90	67.25	365.86	81.07	72.40	38.20
+MARS	**58.85**	**85.54**	**64.88**	**79.47**	**111.71**	**86.00**	**64.43**	**67.27**	**362.23**	**81.66**	**72.10**	**38.24**
EATA	53.05	84.79	67.00	82.06	103.28	85.98	64.55	68.89	356.20	81.29	71.92	39.01
+MARS	**46.55**	**86.03**	**63.13**	**82.55**	**102.08**	**86.18**	**63.83**	**69.00**	**351.33**	**82.03**	**71.15**	**39.04**

Table 2: Failure-aware test-time adaptation performance under **continual** setup.

Method	CIFAR-10-C				CIFAR-100-C				Tiny-ImageNet-C			
	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑
Source	227.17	77.13	81.97	56.48	224.09	81.45	73.37	53.55	582.71	77.85	76.58	22.32
TENT	121.52	73.08	76.27	76.78	867.25	30.94	95.76	20.37	918.84	35.81	94.79	12.42
+MARS	**96.55**	**80.92**	**68.95**	76.30	**609.88**	**50.22**	**88.03**	**40.67**	**768.64**	**55.71**	**86.57**	**21.18**
CoTTA	39.33	86.98	63.42	83.61	116.75	84.99	66.64	67.19	378.59	77.79	75.53	39.40
+MARS	40.16	**87.00**	**61.97**	83.54	**109.05**	**86.13**	**63.49**	**67.75**	381.63	**78.07**	**75.06**	38.95
SAR	62.25	84.52	68.77	79.46	125.99	83.81	66.36	67.09	368.87	79.33	74.07	39.25
+MARS	**58.85**	**85.54**	**64.87**	**79.47**	**97.46**	**85.37**	**65.08**	**70.79**	**366.37**	**80.00**	**73.06**	39.21
EATA	70.53	80.66	70.49	81.12	128.12	83.52	68.08	66.58	405.66	77.78	76.15	36.90
+MARS	**53.01**	**84.29**	**65.99**	**82.48**	**121.73**	**84.24**	**67.59**	**67.14**	**390.54**	**79.59**	**73.92**	**37.21**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and Settings. We primarily evaluate our method on three standard corruption benchmarks: CIFAR-10-C, CIFAR-100-C, and Tiny-ImageNet-C (Hendrycks & Dietterich, 2019). Each dataset contains 15 types of corruptions with 5 severity levels. Pre-trained models are trained on the clean training set, then tested and adapted on the corrupted test set. To comprehensively assess robustness, we adopt three distinct adaptation settings: *standard* (parameters reset between domains), *continual* (parameters carry over across all 15 domains sequentially), and *mixed-domain* (samples from all corruptions are randomly interleaved). While severity level 5 serves as our default, we additionally report results at severities 3 and 1 under the standard protocol. Furthermore, we extend our evaluation to ImageNet-C (Hendrycks & Dietterich, 2019) using CLIP (Radford et al., 2021).

Implementation Details In accordance with standard TTA benchmarks (Döbler et al., 2023; Gao et al., 2024), our network architectures include WideResNet-28 (Zagoruyko & Komodakis, 2016) for CIFAR-10-C, ResNeXt-29 (Xie et al., 2017) for CIFAR-100-C, and ResNet-50 (He et al., 2016) for Tiny-ImageNet-C. For CLIP-based evaluations, we deploy both ResNet-50 and ViT-B/16 (Dosovitskiy et al., 2020) models. We integrate MARS into four established TTA baselines: TENT (Wang et al., 2021), EATA (Niu et al., 2022), SAR (Niu et al., 2023), and CoTTA (Wang et al., 2022). Additionally, we apply MARS to TPT (Manli et al., 2022), a vision-language model-based TTA approach. We further compare MARS with recent state-of-the-art baselines in Table 18. The hyperparameters for all baseline methods are kept strictly at their defaults. For MARS, our default configuration utilizes $K = 8$ masked views, masking ratios $\rho \sim \mathcal{U}(0.2, 0.4)$, a patch size of $P = 4$, and a trade-off coefficient of $\lambda = 0.2$. We run each experiment three times using different random seeds and report the averaged results. All experiments are conducted on NVIDIA RTX 3090 GPUs.

Evaluation Metrics. The performance is measured via top-1 accuracy. To rigorously evaluate failure detection reliability, we report the Area Under the Receiver Operating Characteristic curve (AUROC), False Positive Rate at 95% True Positive Rate (FPR@95), and Area Under the Risk-

Table 3: Failure-aware test-time adaptation performance under **mixed-domain** setup.

Method	CIFAR-10-C				CIFAR-100-C				Tiny-ImageNet-C			
	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑
Source	227.10	77.13	81.97	56.48	224.09	81.45	73.37	53.55	582.71	77.86	76.59	22.32
TENT	334.23	59.55	85.65	60.93	951.83	24.97	97.85	9.42	976.40	23.13	97.21	5.78
+MARS	**133.05**	**81.56**	**72.16**	**67.40**	**772.74**	**42.70**	**93.40**	**23.61**	**927.44**	**37.49**	**93.25**	**10.32**
CoTTA	122.34	81.56	75.01	68.81	183.03	83.26	71.81	57.63	518.59	76.13	79.11	27.80
+MARS	**98.88**	**83.69**	**70.36**	**71.54**	**158.04**	**84.89**	**66.74**	**60.22**	**516.26**	**76.69**	**78.81**	27.76
SAR	151.68	79.38	77.91	66.13	422.65	64.40	79.40	50.84	576.10	72.00	78.93	27.65
+MARS	**143.18**	**80.98**	**73.34**	**66.15**	**389.80**	**67.47**	**78.51**	**52.49**	**574.39**	**72.78**	**78.23**	**27.73**
EATA	188.28	71.92	80.17	68.20	165.62	82.45	70.71	61.33	542.43	74.08	80.31	27.67
+MARS	189.21	**74.44**	**76.72**	67.15	**151.98**	**83.33**	**69.73**	**62.63**	**515.13**	**76.24**	**76.98**	**28.71**

Coverage curve (AURC). Specifically, AUROC captures the ability of the uncertainty score to distinguish successes from failures; FPR@95 reflects the incidence of falsely accepted errors at a high coverage threshold; and AURC provides a holistic measure of the risk-coverage trade-off. As is standard practice (Zhu et al., 2023), we scale all reported AURC values by 10^3.

5.2 Main Results

Standard Adaptation. Table 1 demonstrates the effectiveness of MARS under the standard TTA setup. As a plug-and-play module, MARS consistently enhances the failure detection capabilities of established TTA baselines. Notably, integrating MARS yields substantial reductions in AURC and FPR@95, alongside steady improvements in AUROC across most configurations. Crucially, MARS achieves this enhanced reliability without compromising target-domain accuracy. In fact, when integrated with SAR and EATA, MARS frequently boosts classification accuracy, indicating a strong synergy between failure-aware regularization and adaptation accuracy. Overall, these results validate that MARS successfully mitigates the severe overconfidence typical of standard TTA, achieving a robust balance between accurate predictions and reliable failure detection.

Continual Adaptation. Table 2 evaluates MARS under the challenging continual TTA setup, where error accumulation frequently triggers catastrophic forgetting. Notably, while standard TENT suffers severe degradation on CIFAR-100-C and Tiny-ImageNet-C, integrating MARS dramatically mitigates this collapse, doubling target accuracy and vastly improving failure detection metrics. Across robust baselines like SAR and EATA, MARS consistently lowers AURC and FPR@95 while preserving or enhancing accuracy. Although we observe minor AURC and accuracy trade-offs with the CoTTA baseline, AUROC and FPR@95 metrics still improve. Overall, MARS serves as a crucial stabilizer in continual TTA, preventing model collapse while ensuring reliable uncertainty estimation.

Mixed-domain Adaptation. Table 3 evaluates MARS in the highly challenging mixed-domain setup, where randomly interleaved corruptions often destabilize adaptation. MARS demonstrates remarkable resilience in this volatile setup. Notably, standard TENT suffers catastrophic collapse across all datasets, but integrating MARS strongly mitigates this, dramatically recovering both target accuracy and failure detection metrics (e.g., more than doubling accuracy on CIFAR-100-C). For baselines like CoTTA and SAR, MARS consistently reduces FPR@95 and AURC while maintaining or boosting accuracy. Overall, MARS effectively prevents model collapse and ensures reliable uncertainty estimation even under rapidly shifting, unpredictable domain distributions.

Mild Domain Shifts. Table 4 evaluates MARS under milder distribution shifts (severity 3) in the standard setup. Even in these lower-severity regimes where baseline models already exhibit strong accuracy and reasonable calibration, MARS consistently improves failure detection. Across all tested baselines and datasets, integrating MARS systematically reduces AURC and FPR@95 while boosting AUROC. These results demonstrate that MARS is a robust, universally applicable module that effectively enhances uncertainty estimation without compromising primary classification tasks, regardless of corruption intensity.

Extension to Vision-Language Models (VLMs). Table 5 extends our evaluation to VLMs, integrating MARS with the Test-Time Prompt Tuning (TPT) baseline across CLIP's ResNet-50 and ViT-B/16 architectures. Across all evaluated datasets, MARS consistently enhances failure detection. Adding

Table 4: Failure-aware test-time adaptation performance under standard setup with **severity 3**.

Method	CIFAR-10-C				CIFAR-100-C				Tiny-ImageNet-C			
	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑
Source	84.10	83.73	72.04	74.92	118.45	85.07	67.47	66.75	413.06	79.46	74.61	34.88
TENT	29.89	87.25	59.10	87.42	84.63	86.38	63.06	72.73	267.35	81.93	71.44	47.99
+MARS	**25.14**	**89.24**	**55.66**	87.06	**78.91**	**87.27**	**62.02**	**73.10**	**257.91**	**83.43**	**68.82**	**48.04**
CoTTA	25.90	88.84	58.30	87.08	94.26	86.04	64.87	70.63	276.56	82.56	70.37	46.44
+MARS	**25.67**	**89.27**	**56.08**	86.94	**92.11**	**87.02**	**61.52**	70.34	**274.14**	**83.01**	**69.71**	**46.45**
SAR	30.40	87.96	60.13	86.26	84.52	87.00	62.66	71.94	272.43	82.63	70.59	46.85
+MARS	**28.48**	**88.72**	**56.57**	**86.26**	**84.08**	**87.14**	**62.30**	**71.94**	**270.06**	**83.10**	**69.85**	46.84
EATA	28.07	87.72	58.04	87.62	79.03	87.03	62.30	73.15	263.72	82.65	70.05	47.81
+MARS	**23.80**	**89.14**	**54.34**	**87.84**	**78.31**	**87.19**	**62.16**	**73.20**	**260.53**	**83.20**	**69.68**	**47.86**

Table 5: Failure-aware test-time adaptation performance under standard setup with **VLMs**

Method	CIFAR-10-C				CIFAR-100-C				ImageNet-C			
	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑
ResNet-50 Backbone												
TPT	451.26	70.33	88.05	36.64	723.79	72.91	84.59	14.52	759.47	79.28	72.60	11.13
+ MARS	**423.92**	**77.07**	**80.38**	36.62	**690.16**	**76.66**	**80.85**	**15.61**	**757.79**	**81.19**	**68.03**	10.76
ViT-B/16 Backbone												
TPT	175.59	82.19	75.50	59.53	427.79	82.70	68.14	32.11	580.15	80.39	71.17	21.80
+ MARS	**174.90**	**83.92**	**71.08**	**59.56**	**425.26**	**83.11**	**67.25**	**32.21**	**578.12**	**81.93**	**64.91**	21.56

Table 6: Ablation on different numbers of masked views K. We conducted experiments on 150,000 corrupted images from CIFAR-10-C, comprising 10,000 images for each of the 15 corruption types.

Method	AURC ↓	AUROC ↑	FPR@95 ↓	ACC ↑	Wall-Clock Time (s)	Memory (MB)
EATA	53.05	84.79	67.00	82.06	75.4	2268.0
+ MARS ($K = 2$)	49.93	85.19	63.93	82.47	264.5	10362.9
+ MARS ($K = 3$)	48.75	85.53	63.46	82.49	359.3	12239.5
+ MARS ($K = 4$)	47.41	85.71	63.80	82.60	453.7	14116.1
+ MARS ($K = 8$)	46.55	86.03	63.13	82.55	791.6	21624.7
+ MARS ($K = 16$)	46.33	86.09	62.83	82.53	1509.7	36636.2

MARS yields notable reductions in FPR@95 and AURC, alongside steady AUROC improvements. While this enhanced reliability incurs a marginal trade-off in target-domain accuracy, the safety gains are pronounced. Ultimately, these results demonstrate MARS's broad versatility, proving it effectively enforces reliable uncertainty estimation even when adapting complex VLMs.

5.3 Ablation studies

Number of masked views and inference overhead. Table 6 ablates the number of masked views K in MARS on CIFAR-10-C. Integrating MARS strictly improves both target-domain accuracy and failure detection over the EATA baseline across all evaluated K. While increasing K yields monotonic improvements in failure detection metrics, accuracy remains stable and performance gains marginally plateau beyond $K = 8$. Consequently, we adopt $K = 8$ as our default, striking an optimal balance between computational efficiency and reliable failure detection. Besides, $K = 2$ retains 79% of the FPR@95 reduction and 48% of the AURC reduction while already improving accuracy.

Influence of each component and scoring functions. The ranking loss improves calibration and accuracy on reliable samples, while the consistency score provides a robust uncertainty signal across the distribution (Table 7). While standard MSP and Entropy with our MARS improve upon the EATA baseline, relying solely on masked prediction consistency (r_{cons}) degrades performance (Table 8). Our proposed composite score ($r_{\text{base}} * r_{\text{cons}}$) achieves the best overall results across all metrics. This reveals a key synergy: combining base score with consistency leads to more reliable calibration for failure detection.

Table 7: Ablation on each component in MARS.

Method	AURC ↓	AUROC ↑	FPR@95 ↓	ACC ↑
EATA	53.05	84.79	67.00	82.06
+ $\mathcal{L}_{\text{rank}}$	49.21	85.25	65.93	**82.55**
+ $r(\mathbf{x})$	50.53	85.58	63.14	82.06
+ MARS (full)	**46.55**	**86.03**	**63.13**	**82.55**

Table 8: Ablation on different confidence scores. We use Entropy for r_{base}.

Method	AURC ↓	AUROC ↑	FPR@95 ↓
EATA	53.05	84.79	67.00
+ MARS (MSP)	49.21	85.25	65.93
+ MARS (Entropy)	49.21	85.32	64.79
+ MARS (r_{cons})	62.80	79.08	82.09
+ MARS ($r_{\text{base}} + r_{\text{cons}}$)	60.81	80.93	64.91
+ MARS ($r_{\text{base}} * r_{\text{cons}}$)	46.55	86.03	63.13

Table 9: Ablation on loss weight λ.

λ	AURC ↓	AUROC ↑	FPR@95 ↓	ACC ↑
EATA	53.05	84.79	67.00	82.06
0.1	47.38	85.86	63.16	82.55
0.2	46.55	86.03	63.13	82.55
0.5	46.80	86.22	63.10	82.27
1.0	48.04	86.22	63.69	81.82

Table 10: Ablation on mask ratios.

Mask ratio	AURC ↓	AUROC ↑	FPR@95 ↓	ACC ↑
EATA	103.28	85.98	64.55	68.89
0.2-0.6	102.56	86.19	64.31	68.94
0.1-0.9	102.51	86.14	64.31	68.95
0.4-0.6	102.23	86.19	64.03	68.96
0.2-0.4	102.28	86.16	64.24	69.00
0.01-0.6	102.54	86.18	64.03	68.92

Sensitivity to loss weights. Table 9 ablates the loss weight λ. MARS consistently improves failure detection over the EATA baseline across all evaluated weights. We observe an optimal balance at $\lambda = 0.2$, which minimizes AURC while maximizing accuracy. Larger weights marginally degrade these metrics, demonstrating that moderate regularization best prevents overconfidence without disrupting the primary adaptation process.

Influence of mask ratio. Table 10 ablates the impact of varying random mask ratios in MARS on CIFAR-100-C. Regardless of the chosen range, MARS consistently outperforms the EATA baseline across all failure detection metrics and target-domain accuracy. The minimal performance variance across different intervals highlights that MARS is highly robust and insensitive to specific mask ratio hyperparameters.

Confidence ranking of masked views. As shown in Proposition 1, a well-calibrated model's confidence should degrade monotonically with increasing information removal. To verify this, we generate two masked views $\tilde{\mathbf{x}}^{(1)}$ and $\tilde{\mathbf{x}}^{(2)}$ for each sample $\mathbf{x}$ with mask ratios $\rho_1 < \rho_2$ and measure the fraction of samples violating $\text{conf}(\mathbf{x}) > \text{conf}(\tilde{\mathbf{x}}^{(1)}) > \text{conf}(\tilde{\mathbf{x}}^{(2)})$. As shown in Table 11, EATA+MARS consistently achieves the lowest violation rate across all four ratio combinations on CIFAR-10-C, reducing EATA's violation by 6-8 points. This confirms that MARS produces confidence estimates better aligned with input informativeness, validating the calibration mechanism predicted by our theory.

Table 11: Masked-view monotonicity violation rate.

ρ_1	ρ_2	**Violation rate ↓**		
		Source	EATA	EATA+MARS
0.10	0.70	73.30%	59.38%	51.24%
0.10	0.60	72.06%	59.71%	51.78%
0.20	0.70	73.18%	61.85%	54.00%
0.05	0.70	75.19%	57.93%	50.34%

6 Conclusion

In this work, we point out an important but overlooked weakness of test-time adaptation: existing methods improve accuracy on the target domain, but they are poor at recognizing when their own predictions are wrong. To address this, we introduce failure-aware TTA, which aims to improve accuracy and, at the same time, to better separate correct predictions from incorrect ones. We then propose MARS, a simple plug-and-play method that uses masked inputs to achieve both goals. Experiments on four benchmarks, three adaptation protocols, three corruption levels, and five baseline methods show that MARS consistently improves failure detection while matching or exceeding state-of-the-art accuracy. We hope this work encourages the community to judge TTA methods not only by how accurate they are, but also by how trustworthy their predictions are.

AI USE STATEMENT

We used generative AI tools only to improve the grammar, clarity, and readability of the manuscript. The authors checked all AI-assisted edits to ensure they are accurate and keep the intended meaning, and take full responsibility for the final content.

REFERENCES

Normand J Beaudry and Renato Renner. An intuitive proof of the data processing inequality. *arXiv preprint arXiv:1107.0740*, 2011.

Pierre Brémaud. *An introduction to probabilistic modeling*. Springer Science & Business Media, 2012.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *NeurIPS*, 2019.

Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *CVPR*, 2023.

Hao Dong, Moru Liu, Jian Liang, Eleni Chatzi, and Olga Fink. To trust or not to trust your vision-language model's prediction. *arXiv preprint arXiv:2505.23745*, 2025.

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. *arXiv preprint arXiv:2010.11929*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *CVPR*, 2024.

Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. Doctor: A simple method for detecting misclassification errors. In *NeurIPS*, 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 *CVPR*, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *NeurIPS*, 2018.

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, 2024a.

Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *ICCV*, 2023.

Taeckyung Lee, Sorn Chottananurak, Taesik Gong, and Sung-Ju Lee. Aetta: Label-free accuracy estimation for test-time adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28643–28652. IEEE, 2024b.

Yushu Li, Xun Xu, Yongyi Su, and Kui Jia. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *ICCV*, 2023.

Moru Liu, Hao Dong, Olga Fink, and Mario Trapp. Adaptive confidence regularization for multimodal failure detection. *arXiv preprint arXiv:2603.02200*, 2026.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Yijun Liu, Jiequan Cui, Zhuotao Tian, Senqiao Yang, Qingdong He, Xiaoling Wang, and Jingyong Su. Typicalness-aware learning for failure detection. In *NeurIPS*, 2024.

Yan Luo, Yongkang Wong, Mohan S Kankanhalli, and Qi Zhao. Learning to predict trustworthiness with steep slope loss. In *NeurIPS*, 2021.

Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Goldstein Tom, Anandkumar Anima, and Xiao Chaowei. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *ICML*, 2020.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022.

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

Shuaicheng Niu, Guohao Chen, Peilin Zhao, Tianyi Wang, Pengcheng Wu, and Zhiqi Shen. Self-bootstrapping for versatile test-time adaptation. *arXiv preprint arXiv:2504.08010*, 2025.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In *NeurIPS*, 2022.

Xin Qiu and Risto Miikkulainen. Detecting misclassification errors in neural networks with a gaussian process model. In *AAAI*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Devavrat Tomar, Guillaume Vray, Behzad Bozorgtabar, and Jean-Philippe Thiran. Tesla: Test-time self-learning with automatic adversarial augmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20341–20350. IEEE, 2023.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Openmix: Exploring outlier samples for misclassification detection. In *CVPR*, 2023.

A Detailed Proofs for Theoretical Analysis

Setup. Let X denote the unmasked input, Y the ground-truth label, and $\tilde{X} = X \odot (1 - M)$ the masked view with masking ratio $\rho \in [0, 1 - 1/C]$. I is the mutual information and H is the entropy.

Proposition 1. *For any masking ratios $\rho_1 < \rho_2$, assuming label-relevant information is spatially distributed across at least two patches, then $I(\tilde{X}^{(\rho_2)}; Y) < I(\tilde{X}^{(\rho_1)}; Y)$ and $H(Y|\tilde{X}^{(\rho_2)}) > H(Y|\tilde{X}^{(\rho_1)})$.*

Proof: The random spatial mask M is sampled independently of the data generation process, applying a mask with ratio $\rho_2 > \rho_1$ can be mathematically formalized as applying an additional, independent masking operation M' to the already masked view $\tilde{X}^{(\rho_1)}$. This establishes the extended Markov chain:

$$Y \to X \to \tilde{X}^{(\rho_1)} \to \tilde{X}^{(\rho_2)}. \tag{8}$$

By the Data Processing Inequality (DPI) (Beaudry & Renner, 2011), post-processing cannot increase information. Thus, the mutual information satisfies $I(\tilde{X}^{(\rho_2)}; Y) \leq I(\tilde{X}^{(\rho_1)}; Y)$. Furthermore, the Proposition specifies that label-relevant information is spatially distributed across at least two patches. Because the sampling of M is uniform without replacement, any strictly larger masking ratio $\rho_2 > \rho_1$ has a non-zero probability of erasing at least one additional patch containing label-relevant features. Consequently, the transition from $\tilde{X}^{(\rho_1)}$ to $\tilde{X}^{(\rho_2)}$ involves a strictly positive loss of mutual information:

$$I(\tilde{X}^{(\rho_2)}; Y) < I(\tilde{X}^{(\rho_1)}; Y). \tag{9}$$

Using the Shannon entropy definition of mutual information, $I(A; B) = H(B) - H(B|A)$, we can substitute this into the inequality:

$$H(Y) - H(Y|\tilde{X}^{(\rho_2)}) < H(Y) - H(Y|\tilde{X}^{(\rho_1)}). \tag{10}$$

Since the marginal entropy of the labels $H(Y)$ is a constant independent of ρ, subtracting it from both sides and multiplying by -1 reverses the inequality, yielding the final strict monotonicity property:

$$H(Y|\tilde{X}^{(\rho_2)}) > H(Y|\tilde{X}^{(\rho_1)}). \tag{11}$$

This proves that the true posterior uncertainty must strictly increase as the masking ratio increases.

Proposition 2. *At the global minimum of $\mathcal{L}_{\text{rank}}$, the model's predictive entropy on a masked view with ratio ρ is given by $H(\tilde{\mathbf{p}}) = -(1-\rho)\log(1-\rho) - \rho \log \frac{\rho}{C-1}$. This function is strictly increasing with respect to ρ on $[0, 1 - 1/C]$, perfectly matching the entropy ordering dictated by Proposition 1.*

Proof: The MARS framework defines the soft pseudo-target $\mathbf{q}$ for a masked view $\tilde{\mathbf{x}}_k$ conditioned on the masking ratio ρ_k as:

$$\mathbf{q}(\tilde{\mathbf{x}}_k) = (1 - \rho_k)\mathbf{e}_{\hat{y}} + \frac{\rho_k}{C-1}(\mathbf{1} - \mathbf{e}_{\hat{y}}), \tag{12}$$

where $\mathbf{e}_{\hat{y}}$ is the one-hot encoded pseudo-label and C is the number of classes. The ranking loss $\mathcal{L}_{\text{rank}}$ is the cross-entropy between the target $\mathbf{q}$ and the model prediction $\tilde{\mathbf{p}}$. By Gibbs' inequality (Brémaud, 2012), the global minimum of the cross-entropy loss with respect to the predictive distribution occurs exactly when the predicted distribution matches the target distribution: $\tilde{\mathbf{p}} = \mathbf{q}(\tilde{\mathbf{x}}_k)$. At this global minimum, the predictive entropy $H(\tilde{\mathbf{p}})$ is exactly the entropy of the target distribution $H(\mathbf{q})$:

$$H(\tilde{\mathbf{p}}) = -\sum_{c=1}^{C} \mathbf{q}_c \log \mathbf{q}_c. \tag{13}$$

Since $\mathbf{q}$ consists of one class with probability $(1 - \rho)$ and $C - 1$ classes with probability $\frac{\rho}{C-1}$, the summation evaluates to:

$$H(\tilde{\mathbf{p}}) = -(1-\rho)\log(1-\rho) - \sum_{i=1}^{C-1} \left(\frac{\rho}{C-1}\right) \log \left(\frac{\rho}{C-1}\right) = -(1-\rho)\log(1-\rho) - \rho \log \frac{\rho}{C-1}. \tag{14}$$

This proves the first part of the Proposition. To prove that $H(\tilde{\mathbf{p}})$ is strictly increasing with respect to ρ on the interval $[0, 1 - 1/C]$, we take the first derivative with respect to ρ:

$$\frac{d}{d\rho}H(\tilde{\mathbf{p}}) = \log(1-\rho) + 1 - \left(\log\left(\frac{\rho}{C-1}\right) + 1\right) = \log\left(\frac{(1-\rho)(C-1)}{\rho}\right). \tag{15}$$

For the model's confidence to be better than random chance, we consider the interval $\rho \in [0, 1-1/C)$. In this domain, the probability mass on the predicted class is strictly greater than the uniformly distributed mass on the remaining classes: $(1-\rho) > \frac{\rho}{C-1}$. Rearranging this inequality yields $\frac{(1-\rho)(C-1)}{\rho} > 1$. Because the argument of the logarithm is strictly greater than 1, the derivative is strictly positive:

$$\frac{d}{d\rho}H(\tilde{\mathbf{p}}) > 0. \tag{16}$$

Thus, $H(\tilde{\mathbf{p}})$ is strictly monotonically increasing with ρ, perfectly matching the entropy ordering required by Proposition 1.

B Limitations, Broader Impacts, and Future Work

B.1 Limitations

Our evaluation focuses on image classification under synthetic corruptions and natural distribution shifts; extending the analysis to dense prediction tasks (e.g., segmentation, detection), non-vision modalities, or real-world deployment shifts would strengthen the generality of our findings.

B.2 Broader Impacts

MARS targets reliable failure detection during adaptation, that is directly relevant to the safe deployment of adaptive models in high-stakes settings such as autonomous driving, medical imaging, and industrial monitoring. By enabling adapted models to better recognize their own errors, our work supports human-in-the-loop workflows, selective prediction, and graceful fallback to safer alternatives, which can reduce the harm caused by silent failures under distribution shift. We see this as a positive societal contribution that complements ongoing work on robustness and trustworthy machine learning.

B.3 Future Work

Several directions naturally extend this work. First, beyond the random masking strategy explored here, learned or content-aware masking policies, potentially conditioned on intermediate model features or attention maps, may yield stronger calibration signals at lower computational cost. Second, integrating MARS with selective prediction and abstention frameworks could translate improved failure detection into concrete deployment-time decision rules with formal coverage-risk guarantees. Third, extending failure-aware TTA to structured prediction tasks, multimodal settings, and foundation models presents both methodological and evaluation challenges, particularly in defining failure at the level of complex outputs. Fourth, while our analysis emphasizes synthetic corruptions and standard TTA benchmarks, evaluating MARS under naturally occurring temporal shifts and in fully online streaming scenarios would clarify its behavior in realistic deployments. Finally, we believe failure-aware evaluation deserves to become a standard axis of comparison in the TTA literature, and we hope our framework and benchmarks contribute toward establishing that practice.

C ALGORITHM OVERVIEW

The complete execution flow of MARS is illustrated in Algorithm 1.

Algorithm 1 MARS: Masked Augmentation for Reliable Self-adaptation

Require: Source-pretrained model f_θ; target stream $\mathcal{D}_\mathcal{T}$; base TTA loss $\mathcal{L}_{\text{tta}}$; number of masked views K; masking ratio range $(r_{\min}, r_{\max})$; patch size P; trade-off λ; learning rate η

for each batch $\mathcal{B} = \{x^{(i)}\}_{i=1}^{B}$ from $\mathcal{D}_\mathcal{T}$ **do**
  // Forward pass on original inputs
  $p \leftarrow \mathrm{softmax}(f_\theta(\mathcal{B}))$; $\quad \hat{y} \leftarrow \arg\max p$
  Compute base reliability $r_{\text{base}}(x) = -H(p)$ for all $x \in \mathcal{B}$
  // Identify reliable subset $\mathcal{S}_t$
  if base TTA provides native filtering (e.g., SAR, EATA) **then**
    $\mathcal{S}_t \leftarrow \text{NativeFilter}(\mathcal{B})$
  else
    $\mathcal{S}_t \leftarrow \{x \in \mathcal{B} \mid H(p(x)) \le \bar{H}_\mathcal{B}\}$, where $\bar{H}_\mathcal{B} = \frac{1}{B}\sum_{x\in\mathcal{B}} H(p(x))$
  end if
  // Generate masked augmentations
  Sample $\rho_k \sim \mathcal{U}(r_{\min}, r_{\max})$ for $k = 1, \dots, K$
  Construct masked views $\tilde{x}_k = x \odot (\mathbf{1} - M_k)$ via block-wise masking with patch size P and ratio ρ_k, for all $x \in \mathcal{B}$
  $\tilde{p}_k \leftarrow \mathrm{softmax}(f_\theta(\tilde{x}_k))$ for $k = 1, \dots, K$
  // Confidence ranking loss on reliable subset
  if $|\mathcal{S}_t| > 0$ **then**
    Construct soft pseudo-targets: $q(\tilde{x}_k) = (1-\rho_k)\, e_{\hat{y}} + \frac{\rho_k}{C-1}(\mathbf{1} - e_{\hat{y}})$
    $\mathcal{L}_{\text{rank}} \leftarrow -\dfrac{1}{K|\mathcal{S}_t|} \displaystyle\sum_{x\in\mathcal{S}_t}\sum_{k=1}^{K}\sum_{c=1}^{C} q_c(\tilde{x}_k)\log(\tilde{p}_k)_c$
  else
    $\mathcal{L}_{\text{rank}} \leftarrow 0$
  end if
  // Joint update
  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{tta}} + \lambda\, \mathcal{L}_{\text{rank}}$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}}$
  // Reliability score for failure detection
  $c(x) \leftarrow \dfrac{1}{K}\displaystyle\sum_{k=1}^{K}\langle p,\, \tilde{p}_k\rangle$; $\quad r_{\text{cons}}(x) \leftarrow 1 - c(x)$
  $r(x) \leftarrow r_{\text{base}}(x) \cdot r_{\text{cons}}(x)$
end for
return $\hat{y}, r(x)$

Table 12: Calibration results of ECE, MCE, and Brier Score on CIFAR-10-C (%).

Method	Source	TENT	CoTTA	SAR	EATA
Base ECE ↓	33.99	13.40	12.64	13.79	13.04
+MARS	–	10.86	10.74	13.78	11.84
Base MCE ↓	52.82	34.82	86.73	33.58	34.85
+MARS	–	26.96	31.60	33.47	31.92
Base Brier Score ↓	74.78	31.72	31.58	33.72	30.31
+MARS	–	29.29	29.16	33.70	28.80

Table 13: Failure-aware test-time adaptation performance under standard setup with **severity 1**.

Method	CIFAR-10-C				CIFAR-100-C				Tiny-ImageNet-C			
	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑	AURC↓	AUROC↑	FPR@95↓	ACC↑
Source	27.09	89.10	56.49	86.90	76.10	87.09	63.40	73.81	255.44	82.48	71.30	48.95
TENT	18.38	89.17	51.34	90.60	68.49	87.39	61.29	75.41	224.72	82.84	70.03	52.48
+MARS	**15.69**	**90.62**	**50.46**	90.23	**66.01**	**87.87**	**60.68**	**75.62**	**218.78**	**84.25**	**67.38**	52.22
CoTTA	15.02	90.76	50.59	90.59	75.56	87.42	61.56	73.64	232.13	83.46	69.24	51.00
+MARS	**14.61**	**90.99**	**48.92**	90.51	**74.66**	**87.82**	**60.19**	73.55	**230.26**	**83.93**	**68.38**	50.97
SAR	17.12	90.02	51.80	90.22	69.66	87.67	61.11	74.81	230.06	83.50	68.58	51.26
+MARS	**15.82**	**90.64**	**49.81**	**90.22**	**69.23**	**87.85**	**60.56**	74.80	**228.00**	**83.95**	**68.44**	51.24
EATA	17.57	89.50	50.66	90.70	66.30	87.71	60.93	75.62	222.12	83.64	68.56	52.14
+MARS	**14.36**	**90.70**	**48.70**	**90.91**	**65.49**	**87.94**	**60.19**	**75.72**	**219.31**	**84.14**	**67.92**	**52.18**

D More Results

Calibration Analysis. Table 12 reports Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Brier Score on CIFAR-10-C. MARS consistently improves all three calibration metrics across every baseline. The gains are most pronounced for CoTTA, where MCE drops sharply from 86.73 to 31.60, indicating substantially mitigated worst-case miscalibration. SAR shows the smallest improvements, reflecting its existing reliability-aware filtering, while Tent and EATA exhibit consistent gains across all metrics. These results confirm that MARS yields more reliable confidence estimates, directly supporting its effectiveness for failure detection.

Mild Domain Shifts. Table 13 evaluates MARS under milder distribution shifts (severity 1) in the standard setup. Even in these lower-severity regimes where baseline models already exhibit strong accuracy and reasonable calibration, MARS consistently improves failure detection. Across all tested baselines and datasets, integrating MARS systematically reduces AURC and FPR@95 while boosting AUROC. These results demonstrate that MARS is a robust, universally applicable module that effectively enhances uncertainty estimation without compromising primary classification tasks, regardless of corruption intensity.

Selective-prediction Analysis. Table 14 reports two selective-classification analyses with *fixed-coverage and fixed-risk protocols*. These analyses test how well each method knows when to trust its own predictions. The idea is simple: the model handles the cases it is confident about automatically and passes the uncertain ones to a human. We compare against EATA in two ways, either by fixing how many cases are automated, or by fixing how many mistakes we are willing to tolerate. When the system automates 80–95% of cases, which is typical in real deployments, MARS makes fewer mistakes on the cases it handles automatically. When we allow almost no mistakes (at most 1% error), EATA can barely automate anything (0.1% of cases), while MARS still automates 1.8%, about 18$\times$ more at the same error level. Whenever the error budget is 5% or higher, MARS automates more cases than EATA at every setting we tested.

Effect of the reliable-sample filter on the ranking loss. To test this, we removed the confidence filter, so the ranking loss is trained on every sample, including those whose pseudo-labels are wrong. As Table 15 shows, the model still does not collapse. The reason is that MARS does not try to match each pseudo-label exactly. Instead, it only asks that predictions become less confident as more of the

Table 14: Selective prediction on CIFAR-10-C. Left: selective error (%) at fixed coverage. Lower is better. Right: max. coverage (%) under a selective-error ceiling. Higher is better.

Coverage	EATA	EATA+MARS
80%	9.42	**9.13**
90%	13.36	**12.92**
95%	15.58	**15.12**

Risk ceiling	EATA	EATA+MARS
1%	0.1	**1.8**
5%	63.1	**63.8**
10%	81.6	**82.6**
15%	93.7	**94.7**

Table 15: Ablating the confidence filter in $\mathcal{L}_{\text{rank}}$ (CIFAR-10-C, severity 5).

Base	Variant	AUROC↑	FPR@95↓	AURC↓
TENT	TENT	84.02	67.86	57.96
	+MARS (no filter)	84.35	66.43	56.78
	+MARS (with filter)	**86.07**	**62.82**	**46.57**
EATA	EATA	84.79	67.00	53.05
	+MARS (no filter)	85.30	65.15	51.92
	+MARS (with filter)	**86.03**	**63.13**	**46.55**

input is masked, in a consistent order. Because this rule depends on the overall pattern rather than on any single label, a few wrong labels cannot easily throw it off. That said, wrong pseudo-labels still add noise to training and weaken this ordering. Filtering out unreliable samples therefore matters a lot for getting the best results: adding the confidence filter clearly improves performance.

Comparison with other failure detection scores. As shown in Table 16, simple post-hoc baselines applied during test-time adaptation yield suboptimal or inconsistent failure detection performance, whereas MARS substantially outperforms all of them across every metric.

Table 16: Comparison with other failure detection scores on CIFAR-10-C.

Method	AURC ↓	AUROC ↑	FPR@95 ↓
EATA	53.05	84.79	67.00
+ Energy (Liu et al., 2020)	66.18	81.06	69.48
+ temperature scaling (Guo et al., 2017)	53.76	84.74	65.58
+ test-time augmentation variance (K=8)	49.69	85.06	70.46
+ MC-Dropout (Gal & Ghahramani, 2016)	54.82	84.22	68.77
+ MARS	46.55	86.03	63.13

Cross-domain generalization with vision-language models. Table 17 evaluates TPT with MARS on Flowers102 (Nilsback & Zisserman, 2008), DTD (Cimpoi et al., 2014), and OxfordPets (Parkhi et al., 2012). Integrating MARS consistently enhances failure detection capabilities during cross-domain adaptation while preserving or improving the model's underlying accuracy.

Comparison with additional SOTA methods. Table 18 shows that MARS can be easily added to recent state-of-the-art TTA methods. MARS is simply an extra loss term: it asks the model to become less confident as more of the input is masked. Because it does not depend on how any particular baseline works, it fits well alongside the sample selection and re-weighting strategies used in newer methods. To show this, we added MARS to SPA (Niu et al., 2025), DeYO (Lee et al., 2024a), AETTA (Lee et al., 2024b), and TeSLA (Tomar et al., 2023) on CIFAR10-C, CIFAR100-C, and Tiny-ImageNet-C. On every dataset, MARS helps the model better detect its own mistakes, while keeping or improving its accuracy.

Comparison of masking strategies. Proposition 1 assumes that label-relevant information is spread evenly across patches. Under this assumption, with masks sampled independently for each ratio, masking more patches removes more useful information on average, so the model's true uncertainty grows with the masking ratio in expectation. MARS relies on this average trend. When useful

Table 17: Cross-domain generalization on Flowers102, DTD, and OxfordPets with CLIP RN-50 and ViT-B/16.

Dataset	Method	Acc↑	AUROC↑	FPR@95↓	AURC↓
CLIP RN-50					
Flowers102	TPT	57.73	83.45	62.00	191.55
	+ MARS	**57.78** +0.05	**84.56** +1.11	**61.63** -0.37	**179.78** -11.77
DTD	TPT	40.19	76.76	82.58	372.41
	+ MARS	**40.35** +0.16	**78.90** +2.14	**77.87** -4.71	**355.59** -16.82
OxfordPets	TPT	**80.68**	84.16	64.46	57.64
	+ MARS	80.49 -0.19	**86.07** +1.91	**61.87** -2.59	**51.61** -6.03
CLIP ViT-B/16					
Flowers102	TPT	64.31	83.07	66.78	149.82
	+ MARS	**64.34** +0.03	**85.97** +2.90	**57.36** -9.42	**142.51** -7.31
DTD	TPT	41.43	80.10	73.93	342.23
	+ MARS	**41.84** +0.41	**80.43** +0.33	**69.93** -4.00	**340.30** -1.93
OxfordPets	TPT	83.37	85.46	62.16	44.97
	+ MARS	**85.09** +1.72	**88.23** +2.77	**57.38** -4.78	**31.65** -13.32

Table 18: MARS consistently improves failure detection while preserving or improving adaptation accuracy across four recent TTA baselines.

Dataset	Method	Acc↑	AUROC↑	FPR@95↓	AURC↓
CIFAR10-C	SPA	80.79	85.11	67.24	55.02
	+ MARS	**81.01** +0.22	**85.85** +0.74	**65.36** -1.88	**51.65** -3.37
	DeYO	81.60	84.69	66.97	54.91
	+ MARS	**82.13** +0.53	**86.08** +1.39	**62.73** -4.24	**47.79** -7.12
	AETTA	79.91	84.67	68.61	60.05
	+ MARS	**80.31** +0.40	**85.86** +1.19	**64.75** -3.86	**54.66** -5.39
	TeSLA	80.43	85.31	67.97	55.33
	+ MARS	**81.02** +0.59	**86.50** +1.19	**63.51** -4.46	**49.88** -5.45
CIFAR100-C	SPA	65.61	85.48	65.39	122.93
	+ MARS	**66.02** +0.41	**85.76** +0.28	**64.49** -0.90	**119.75** -3.18
	DeYO	69.52	86.06	63.91	100.14
	+ MARS	**69.53** +0.01	**86.18** +0.12	**63.80** -0.11	**99.70** -0.44
	AETTA	**66.66**	85.88	64.67	115.44
	+ MARS	66.45 -0.21	**86.19** +0.31	**63.77** -0.90	**115.41** -0.03
	TeSLA	66.30	85.90	64.85	117.31
	+ MARS	**66.35** +0.05	**86.29** +0.39	**63.48** -1.37	**115.63** -1.68
Tiny-ImageNet-C	SPA	37.96	81.40	72.27	366.44
	+ MARS	**38.01** +0.05	**81.71** +0.31	**71.98** -0.29	**363.89** -2.55
	DeYO	39.83	80.50	73.43	353.78
	+ MARS	**40.08** +0.25	**81.64** +1.14	**71.61** -1.82	**344.79** -8.99
	AETTA	37.42	80.87	72.57	375.70
	+ MARS	**37.73** +0.31	**82.03** +1.16	**71.64** -0.93	**364.82** -10.88
	TeSLA	37.78	80.83	72.74	371.62
	+ MARS	**37.96** +0.18	**82.13** +1.30	**70.80** -1.94	**361.99** -9.63

information is concentrated in a few patches, this ordering holds for every single image only if masks are nested, i.e., each larger mask contains the smaller one. We use independent masks because they are cheaper and simpler, and in our experiments (Table 19) they perform as well as or better than nested masks.

Table 19: Comparison of masking strategies on CIFAR-10-C. Independent masks sample mask ratios independently per augmentation. Nested masks sort the sampled ratios and apply masks cumulatively, so patches masked at a smaller ratio are strictly contained in those masked at larger ratios.

Method	ACC ↑	AUROC ↑	FPR@95 ↓	AURC ↓	Wall-Clock (s) ↓	Peak Mem. (MB) ↓
MARS (nested masks)	82.49	85.50	63.50	48.68	1,343.7	21,636
MARS (independent masks, default)	**82.55**	**86.03**	**63.13**	**46.55**	**453.7**	**14,116**

E More Discussions

E.1 The reliability score definition

The score is designed so that higher $r(x)$ means a more reliable sample. It combines two signals: entropy, which measures predictive uncertainty, and $r_{cons} = 1 - c(x)$, which measures how much the prediction changes under masking, where $c(x)$ is the agreement between the original and masked predictions. A failure-prone sample is both uncertain and unstable, so both terms are large and $r(x) = -\text{entropy} \times r_{cons}$ is strongly negative. A reliable sample is confident and stable, so r_{cons} is near zero and $r(x)$ is close to zero, its highest possible value.

Because $r(x) = -H(p)(1 - c(x))$ is a product, the score approaches 0 (maximum reliability) whenever either the entropy $H(p)$ or the inconsistency $1 - c(x)$ approaches 0. We checked how often this happens in practice. We call a sample **confident-but-unstable** if $H(p) < 0.1$ and $c(x) < 0.1$, and **uncertain-but-stable** if $H(p) > 0.9$ and $c(x) > 0.7$. On CIFAR10-C with EATA+MARS, uncertain-but-stable samples never appear (0.00%), and confident-but-unstable samples make up only 1.13%. So although these cases are possible in theory, they are rare enough that the score works well in practice.

E.2 The AUROC/FPR@95 improvement is not merely a byproduct of the large accuracy gain

Higher accuracy naturally improves failure detection, since a more accurate model separates correct and wrong predictions more clearly. However, MARS's gains in AUROC, FPR@95, and AURC do not come from accuracy alone. To show this, we look at cases where the base method and MARS reach almost the same accuracy. For SAR on CIFAR-10-C in the continual setting (Table 2), accuracy is 79.46% vs. 79.47%, yet MARS lowers AURC from 62.25 to 58.85 and FPR@95 from 68.77% to 64.87%, and raises AUROC from 84.52% to 85.54%. In the mixed-domain setting (Table 3), accuracy is 66.13% vs. 66.15%, yet MARS lowers AURC from 151.68 to 143.18 and FPR@95 from 77.91% to 73.34%, and raises AUROC from 79.38% to 80.98%. So even with no accuracy gain, MARS makes the model's confidence more reliable.

E.3 The gap between the theoretical claim of posterior-entropy alignment and the large residual violation rate

The remaining violations (Table 11) come from two factors. First, MARS enforces the ranking through a soft loss with a small weight ($\lambda = 0.2$), so that it does not override the main goal of improving accuracy on the target domain. Forcing zero violations would likely hurt adaptation. Second, TTA works on heavily corrupted data where features are already badly distorted, so perfect per-sample calibration is very hard for any deep network. Even so, MARS reduces the violation rate by 6–8 percentage points over the baseline.

E.4 The monotonicity range in Proposition 2

The two boundary cases are as follows. At $\rho = 0$, no information is lost, so the expected entropy is at its minimum. At the other end, entropy is highest when the model has lost all useful information and predicts every class with equal probability $1/C$; this corresponds to $\rho = 1 - 1/C$. Between these two points, the expected entropy strictly increases with ρ.