Entropy Minimization without Model Collapse: Mitigating Prediction Bias in Medical Imaging

Tim Nielen[1]*, Sameer Ambekar[1,2,4,5]*, Daniel M. Lang[1,2#] , Julia A. Schnabel[1,2,3,4,5#]
[1]School of Computation, Information and Technology, Technical University of Munich, Germany
[2]Institute of Machine Learning in Biomedical Imaging, Helmholtz Munich, Germany
[3]School of Biomedical Engineering and Imaging Sciences, King's College London, UK
[4]Munich Center for Machine Learning (MCML)
[5]relAI – Konrad Zuse School of Excellence in Reliable AI

Abstract

Entropy minimization (EM) is the dominant objective for test-time adaptation, yet its failure mode, model collapse, remains poorly understood. In this work, we show that distribution shifts can cause feature clusters corresponding to distinct classes in the model's representation space to merge, while the decision boundary remains fixed. This induces a systematic skew in the predicted class distribution, referred to as prediction bias. Prediction bias refers to a shift in the predicted class distribution, with some classes overrepresented and others suppressed. We show that entropy minimization amplifies this prediction bias by tightening the existing clusters, reinforcing the incorrect groupings until all predictions collapse to a trivial solution. Next, to demonstrate the significance of prediction bias and mitigate it, we further propose *Distribution Shift Bias Reduction* (DSBR), a bias-correcting objective that specifically targets this failure mode by equalizing the contribution of each predicted class to the unsupervised entropy minimization loss. To study this failure mode, we design suitable adaptation settings using four medical-imaging datasets and additionally evaluate on ImageNet-C. We find that DSBR consistently stabilizes test-time adaptation, prevents model collapse, and matches or outperforms the state-of-the-art methods. Moreover, DSBR operates solely at test-time. Our code is available at `https://entropy-minimization-dsbr.github.io/`.

1 Introduction

Machine learning models face a fundamental challenge during real-world deployment: a model trained on source data must adapt to unseen data at inference time, without access to the training data or labels from the unseen deployment domain [10, 84, 91]. This challenge arises through distribution shifts caused by, e.g., changes in sensor properties [92], environmental conditions [33], and acquisition protocols [92] that alter the input data distribution while leaving the semantic task itself unchanged. Such shifts can lead to substantial drops in model performance when the learned representation is no longer well aligned with the target data [27, 35, 61]. Medical imaging is a particularly important example of this setting [80]. A pathology image acquired with a different scanner or stain remains semantically the same, yet becomes statistically misaligned with the source domain [36, 74]. In these scenarios, the model's confidence and prediction reliability can deteriorate significantly [5, 42, 74].

In natural vision, Entropy Minimization (EM) has emerged as the dominant objective for Test-Time Adaptation (TTA) methods designed to handle distribution shifts during deployment [51, 59, 64, 82]. A seminal work on EM, Tent [82] showed that minimizing the Shannon entropy [71] of the model's

*Shared first-authorship #Shared last-authorship

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

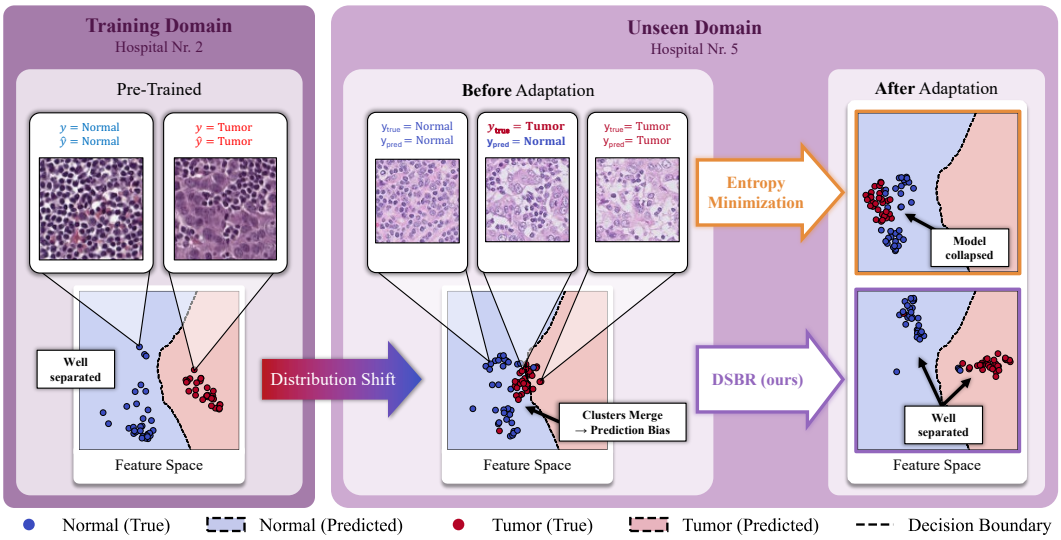

Figure 1: **Feature space clusters merge under distribution shift, inducing severe prediction bias, which is prone to failure for Entropy Minimization (EM).** We show the 2-dimensional t-SNE embeddings from WILDSCamelyon [42] dataset featurized by ResNet50-GN. The background colors indicate the area of the feature space the model classifies as each class which we obtain by fitting a k-nearest neighbors classifier on the 2D embeddings and corresponding model predictions. Entropy minimization [46, 82] is prone to model collapse. In contrast, **DSBR (ours)** mitigates model collapse and prevents cluster merging across multiple iterations.

predictions over the affine parameters of batch normalization layers reliably reduces prediction error under mild distribution shift. Since then, EM-based unsupervised loss has been widely used as an unsupervised objective for many TTA methods [58, 59, 95, 96]. Subsequent work has focused on making EM more robust in wild scenarios, including limited batch sizes and mixed distribution shifts [59], through filtering unreliable samples [46, 59], modified loss functions [96], sharpness-aware optimization [59], and preventing model collapse [59]. With model collapse depicting a critical failure mode in which the model outputs only one or a few classes, irrespective of the input [18, 19].

The progress of TTA methods for natural-image tasks has been largely driven by benchmarks that introduce synthetic distribution shifts in the form of severe corruptions or do not fully capture the real-world shifts [32, 63, 93]. In contrast, the shifts encountered in medical imaging are typically much subtler [39]. They often preserve the underlying pathology while being driven by clinically relevant factors such as changes in patient demographics, scanner hardware, staining protocols, or acquisition parameters. These variations modify the input statistics in ways that are difficult to detect visually [92], thereby posing a challenging problem for TTA methods. Moreover, as stated above, model collapse poses a critical risk because it produces single-dominant class outputs regardless of the input [18, 19]. In clinical settings, such failures may go unnoticed in practice, allowing compromised models to influence decisions in ways that can lead to serious, potentially life-threatening effects [7, 21, 47, 89].

Therefore, a rethinking of entropy minimization for medical imaging is needed. In this setting, deployment shifts are not only unknown in advance but also subtle and heterogeneous across acquisition protocols, scanners, and patient populations [39, 92]. Consequently, heuristics that perform well on synthetic corruption benchmarks do not transfer reliably to the medical domain [21, 39, 65]. While Press et al. [64] show that entropy minimization can succeed or fail under distribution shifts, the mechanisms driving model collapse remain unclear. In particular, it remains unclear from a theoretical perspective why the same EM objective improves adaptation under some shifts yet leads to catastrophic collapse under others.

As illustrated in Fig. 1, distribution shift distorts the feature space geometry. As class clusters drift and partially merge while the classifier boundary remains fixed, prediction bias amplifies, with some classes absorbing ambiguous samples and others becoming suppressed. To understand and improve entropy minimization, we make the following contributions:

- **Prediction bias as the root cause of model collapse, amplified by entropy minimization.** We identify prediction bias, a skew in the predicted class distribution induced by distribution shift, as the central failure mode of entropy minimization. We show that the shift distorts feature geometry, causing class clusters to drift and partially merge while the classifier boundary remains fixed. Entropy minimization amplifies this bias by reinforcing already dominant clusters, ultimately leading to model collapse.
- **Mitigating model collapse with Distribution Shift Bias Reduction (DSBR).** We propose DSBR, an importance reweighting method that equalizes the contributions of each predicted class to the entropy loss. By preventing primarily predicted classes from dominating the optimization signal, DSBR mitigates prediction bias and promotes a more stable separation of feature clusters during adaptation, as shown in Fig. 1.
- **We design challenging entropy-minimization scenarios for medical imaging.** We evaluate state-of-the-art baselines and DSBR on challenging entropy-minimization settings, including clinically grounded scenarios and clinical shifts in medical imaging. In addition, we demonstrate the effectiveness of DSBR in preventing model collapse.

2 Background and Related Work

Notations. We focus on the task of image classification. Let $\mathcal{X}$ denote the input space, where random samples X are distributed according to a probability distribution $p_X(x)$ for $x \in \mathcal{X}$, and let $\mathcal{K} := \{1, \dots, K\}$ denote the set of classes. We consider a pre-trained model $f_\theta : \mathcal{X} \to \Delta^{K-1}$ with parameters θ, and write it as the composition $f_\theta(x) = h(\phi(x))$, where ϕ is a featurizer mapping inputs to a representation space and h is a classifier acting on these representations. We define the per-sample entropy loss as $l(x) := H(f_\theta(x))$ and the predicted class as $\hat{y}(x) := \arg\max f_\theta(x)$. Finally, $p_{\hat{Y}}(k)$ describes the predicted class distribution given $\hat{Y} = \hat{y}(X)$ and $k \in \mathcal{K}$.

The Entropy Minimization Enigma. In test-time adaptation, a model trained on source data, must adapt to unlabeled test data at inference time, without access to source data or any labels [82]. Unlike offline adaptation methods, including source-free domain adaptation [48, 91], which typically assume access to the entire target set, test-time adaptation operates solely online on a stream of incoming test data. Approaches include, but are not limited to, updating batch normalization statistics using an unsupervised loss [57, 67], leveraging surrogate tasks to learn test-time representations [79], utilizing pseudo-labels [44], and minimizing the entropy of model predictions [82]. Among these, entropy minimization (EM) has become the dominant choice for a wide range of methods [58, 59, 85, 96], and is also the focus of this work. Specifically, it minimizes the Shannon entropy [71] of the softmax output, $H(\mathbf{p}) = -\sum_{i=1}^{K} p_i \log p_i$ where $\mathbf{p} = f_\theta(\mathbf{x}) \in \Delta^{K-1}$.

This equation assumes that lower-entropy predictions correlate with higher accuracy [82]. Tent [82] uses EM to update the affine parameters of batch normalization layers and achieves strong results on image corruption benchmarks such as ImageNet-C and CIFAR-10/100-C. However, Niu et al. [59] showed that Tent is prone to failure in more realistic *wild* TTA scenarios. Overall, the inner mechanisms driving both the successes and failures of EM remain poorly understood, a phenomenon Press et al. [64] termed *Entropy Enigma*. They showed that EM acts as a cluster refinement process, tightening feature representations within clusters of predicted classes, and identified a failure case where these clusters drift away from their original class centers. We show that this failure is even seeded before adaptation begins: distribution shifts can cause clusters of distinct classes to merge, leading EM to reinforce incorrect groupings.

Stabilizing Entropy Minimization. The standard approach to mitigate EM failure is to identify and discard unreliable samples [46, 58, 59]. A common proxy for reliability is prediction entropy: samples above a fixed entropy threshold are considered too uncertain and excluded from the backward pass [46, 58, 59]. DeYO [46] argues that entropy alone is an unreliable signal and augments it with a shape-based score that measures how much a structure-destroying augmentation changes the model's prediction. Notably, such methods require dataset- and model-dependent hyperparameters to be fixed. However, high-entropy samples are not always harmful as stated in Press et al. [64]. Moreover, filtering is wasteful and becomes problematic in scenarios where data is scarce, which is often the case for medical imaging. Several methods further stabilize EM by modifying the loss or the optimization. EATA [58] adds a Fisher information-based regularizer to prevent the model from forgetting its source-domain knowledge during adaptation. SAR [59] applies sharpness-aware

(a) **WILDSCamelyon**: Hospital Nr. 2 → 4

✓ *Stable*		Predicted	
Class	**True**	**Before EM**	**After EM**
Normal	49.8%	62.8%	52.4%
Tumor	50.2%	37.2%	47.6%

(b) **ImageNet-C**: Gaussian Noise level 3

✓ *Stable*		Predicted	
Class	**True**	**Before EM**	**After EM**
Christmas stocking	0.1%	0.2%	0.3% (max)
999 other (median)	0.1%	0.1%	0.1%

(c) **WILDSCamelyon**: Hospital Nr. 2 → 5

✘ Model Collapse		Predicted	
Class	**True**	**Before EM**	**After EM**
Normal	49.6%	85.0%	**100.0%**
Tumor	50.4%	15.0%	**0.0%**

(d) **ImageNet-C**: Fog level 5

✘ Model Collapse		Predicted	
Class	**True**	**Before EM**	**After EM**
Geyser	0.2%	25.5%	**99.3%**
999 other (median)	0.1%	0.1%	0.0%

Table 1: **Prediction bias leads to Model collapse.** Predicted class distributions before and after EM with ResNet50-GN pre-trained on WILDSCamelyon Hospital 2 (a,c) and ImageNet (b,d), on 20% holdouts of respective target domains. The color shadings indicate the prediction bias; the greater the color difference between *True* and *Before EM*, the more likely it is that model collapse occurs.

minimization [22] to encourage convergence to flat minima, making updates more robust to noisy gradients. ROID [54] maintains a running weight ensemble between the adapted and source model to prevent parameter drift. However, ROID, and mean-teacher methods in general, do not address the root cause of model collapse, and instead solve it by incurring increased compute costs due to auxiliary teacher and augmentations. Recently, COME [96] replaces the entropy loss with one derived from a Dirichlet prior and down-weights the low-evidence predictions. In addition, energy-based works, including Park et al. [62], augment the objective with additional loss terms for optimization. We deliberately avoid all of these strategies for our method, as our goal is to isolate and directly target prediction bias as the primary driver of EM failure.

Importance Reweighting. Importance reweighting is also used to address distribution shifts [26, 34, 50, 73, 77]. Existing work addresses input shift by reweighting the pre-training loss with $p_X(x)/q_X(x)$ [26, 34, 73, 77], but this requires access to source data and it's distribution q_X and is thus not applicable in fully test-time adaptation. In our setting, input shift induces a bias in the predicted class distribution $p_{\hat{Y}}$. To target this, ROID [54] performs diversity weighting through cosine dissimilarity over the model's softmax outputs. These are known to be miscalibrated in modern network architectures [28]. We however, reweight the entropy loss proportional to $1/p_{\hat{Y}}$ directly.

3 Prediction Bias Leads to Model Collapse and Is Amplified by Entropy Minimization

3.1 Prediction Bias Is the Root Cause of Model Collapse

Model collapse under entropy minimization is rooted in a skew of the predicted class prior on the target domain. We refer to this skew as *prediction bias*, consistent with prior work [17, 46, 54, 83]. Our central claim, however, is different: under distribution shift, prediction bias is already present before adaptation, and during adaptation, entropy minimization subsequently amplifies it until collapse.

To explain the origin of this bias, consider a standard model of the form $f_\theta(x) = h(\phi(x))$, where ϕ is a featurizer and h is a classifier. After source training, the learned representation typically organizes samples of different classes into compact and approximately separable regions in feature space. The classifier h then induces a decision boundary

$$\mathcal{H} := \{z \in \phi(\mathcal{X}) \mid h(z)_i = h(z)_j,\ i \neq j\},$$

which partitions the feature space into regions associated with different predicted classes. Under distribution shift, however, the feature distribution changes while the classifier remains fixed. Some class clusters remain stable, whereas others drift toward neighboring classes and partially cross the decision boundary. This produces a skew in the model's predictions: certain classes absorb ambiguous samples from others, becoming overrepresented, while others are suppressed. Crucially, this prediction bias is present before any adaptation takes place.

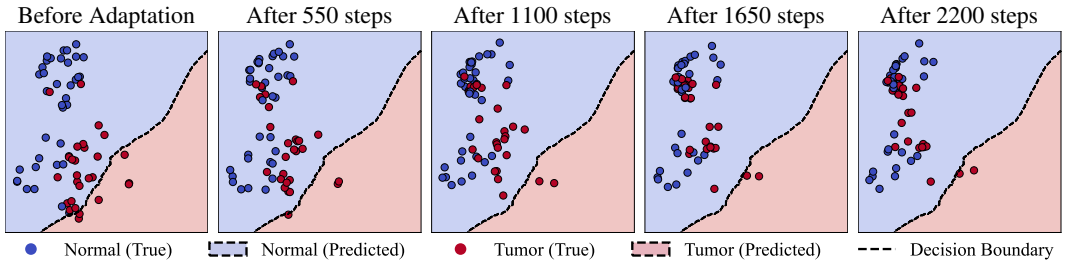

Figure 2: **Entropy minimization amplifies prediction bias and ultimately causes model collapse.** We show 2D UMAP embeddings of WILDS-Camelyon features from a ResNet50-GN pretrained on Hospital 2, before and during adaptation. Background colors denote predicted regions in feature space, obtained by fitting a k-nearest neighbors classifier to the embeddings and model predictions. As entropy minimization sharpens biased predictions, samples are increasingly reassigned to the dominant class, leading to cluster merging and collapse.

Table 1 illustrates how prediction bias leads to model collapse. Under mild shift, the pre-adaptation predicted distribution is only moderately distorted, and entropy minimization does not collapse the model. Under severe shift, however, the model is already strongly biased before adaptation begins, and entropy minimization amplifies this skew into a nearly single-class solution. Table 1 therefore shows that collapse does not emerge from an initially balanced state, but from a sufficiently biased initialization.

Entropy minimization does not create this bias, but amplifies it. By sharpening already dominant predictions and tightening entangled clusters, it pulls ambiguous samples further into overrepresented classes, thereby reinforcing the existing skew. As this feedback loop intensifies, the model collapses to a trivial dominant-class solution. In this sense, prediction bias is not a byproduct of collapse, but its root cause. From Table 1, the ImageNet-C Fog example illustrates this mechanism concretely. Before adaptation, the model already assigns 25.5% of samples to *Geyser*, despite its true frequency being only 0.2%, and after entropy minimization, this class absorbs nearly all probability mass. A plausible explanation is that fog introduces appearance statistics resembling those of geysers, displacing many samples toward this region in feature space before adaptation.

3.2 Why Entropy Minimization Amplifies Prediction Bias

In the failure case shown in Fig. 2, more than half of the tumor samples have crossed the decision boundary into the *Normal* cluster. EM treats this merged region as a single cluster and refines it on one side of the boundary, which is precisely how prediction bias amplifies into collapse. This behavior follows from the geometry of entropy minimization. Because EM-based test-time adaptation typically updates only the affine parameters of normalization layers, the classifier h and the decision boundary $\mathcal{H}$ remain fixed throughout adaptation. EM therefore operates entirely within a fixed decision geometry, pushing representations $\phi(x)$ toward more confident predictions. Since the entropy loss is continuous, nearby samples receive similar gradients, so EM refines groups of similarly predicted samples rather than correcting each sample independently.

When the distribution shift has already caused true-class clusters to overlap across $\mathcal{H}$, EM tightens this merged region instead of separating it. Ambiguous samples are pulled further toward the dominant predicted class, $p_{\hat{Y}}(k)$ becomes increasingly skewed, and each subsequent update starts from an even more biased state. Entropy minimization and prediction bias thus form a positive feedback loop, with model collapse as its endpoint.

4 Proposed Method: Distribution Shift Bias Reduction (DSBR)

4.1 Balanced Target Distribution

To counteract the effect of prediction bias leading to model collapse during EM, we define a *balanced target distribution* that assigns equal probability mass to each predicted cluster while preserving the

within-cluster relative density:

$$p_X^*(x) \;=\; \frac{p_X(x)}{p_{\hat{Y}}(\hat{y}(x))} \cdot \frac{1}{K} \tag{1}$$

which is derived from $p_X(x) = p_{X,\hat{Y}}(x \mid \hat{y}(x)) \cdot p_{\hat{Y}}(\hat{y}(x))$ by replacing the potentially biased $p_{\hat{Y}}(\hat{y}(x))$ with the uniform distribution $1/K$. Minimizing the expected loss under p_X^* rather than p_X prevents any single cluster from dominating the gradient signal. Because EM's dynamics are gradient-driven, redistributing gradient mass away from dominant clusters provides a corrective force that directly opposes the merging dynamic: samples within an overrepresented cluster, many of which were incorrectly absorbed from other classes, contribute less to each update, while samples in underrepresented clusters, precisely those that resisted incorrect reassignment, contribute more. This rebalancing actively encourages the separation of merged clusters rather than their further consolidation.

At test-time, we cannot sample from p_X^*, only from p_X. We recover the target loss $\mathcal{L}^*$ via importance reweighting [73]:

$$\mathcal{L}^* := \mathbb{E}_{X\sim p_X^*}[l(X)] = \mathbb{E}_{X\sim p_X}\left[\frac{p_X^*(X)}{p_X(X)}\, l(X)\right] = \mathbb{E}_{X\sim p_X}\left[\frac{l(X)}{K \cdot p_{\hat{Y}}(\hat{y}(X))}\right] \tag{2}$$

The importance weight is large for samples in underrepresented clusters and small for those in overrepresented ones.

4.2 Approximating the Cluster Probabilities at Test-Time

The weights in Eq. (2) require knowledge of $p_{\hat{Y}}(k)$, which is unknown at test-time. Given a batch $\mathcal{B}_t = \{X_1, \ldots, X_B\} \sim p_X(x)$ of size B and sampled at adaptation step t, a natural approximation for $p_{\hat{Y}}(k)$ at adaptation step t, which we refer to as $P_t^{(k)}$, is the empirical cluster frequency

$$\tilde{P}_t^{(k)} = \frac{B_t^{(k)}}{B}, \qquad B_t^{(k)} := \big|\{X \in \mathcal{B}_t \mid \hat{y}(X) = k\}\big|. \tag{3}$$

For datasets with fewer classes, a single-batch estimate is sufficient. For datasets with a high number of classes, such as ImageNet-C (K=1000), however, batches with $B \ll K$ omit most classes, making per-batch estimates degenerate. Increasing the batch size would address this, but is inefficient. Instead, assuming adaptation steps are small so that $P_t^{(k)} \approx P_{t-1}^{(k)}$, we maintain an exponential moving average (EMA) of cluster frequencies across batches:

$$\hat{P}_t^{(k)} = \alpha\, \hat{P}_{t-1}^{(k)} + (1-\alpha)\, \frac{B_t^{(k)}}{B}, \tag{4}$$

initialized at $\hat{P}_0^{(k)} = 1/K$ and with decay $\alpha = 0.9$. The initialization reflects the prior that clusters are balanced before any evidence is observed. Substituting into Eq. (2), the DSBR loss estimate over batch $\mathcal{B}_t$ is

$$\hat{\mathcal{L}}_t^* = \frac{1}{B} \sum_{X\in\mathcal{B}_t} \frac{l(X)}{K \cdot \hat{P}_t^{(\hat{y}(X))}}. \tag{5}$$

We provide the detailed algorithm in Appendix A.

5 Experiments

Datasets. We evaluate on four medical datasets and one natural-vision dataset each with multiple domains and spanning microscopy, X-ray, and MRI, with both 2D and 3D inputs and diverse anatomical sites. *WILDSCamelyon* [42], derived from Camelyon17 [78], is a breast histopathology dataset with 455,954 binary classification patches from five hospital domains differing in staining, scanners, and lab procedures. *Histopantum* [94] is a binary tumor histopathology dataset on TCGA-derived patches [88], with four domains corresponding to colorectal, ovarian, stomach, and uterine cancers, yielding biological rather than instrumental shift. *MammoBench* [24] unifies six mammography datasets (domains) [2, 13, 38, 43, 55, 60] into a Normal-vs-Abnormal dataset; excluding CMMD [13], which lacks Normal cases, leaves 14,529 images across the remaining domains. *GliomaMRI*

Table 2: **Balanced accuracy (%) on the four medical benchmarks.** The results are averaged over all source-target configurations of each dataset. We re-implement the baselines on the medical datasets and report the average improvement over the corresponding non-adapted baseline (ResNet, ViT) as Δ Average. **Bold** marks the best result per column, and underlined the second-best. DSBR is either the top performer or performs competitively

Model & Method	**Histopantum**	**MammoBench**	**WILDSCamelyon**	**GliomaMRI**	**Δ Average**
ResNet	$86.6_{\pm 0.0}$	$59.4_{\pm 0.0}$	$87.4_{\pm 0.0}$	$42.9_{\pm 0.0}$	▲ $0.0_{\pm 0.0}$
• Tent [82]	$88.5_{\pm 0.1}$	$58.8_{\pm 0.1}$	$87.9_{\pm 0.0}$	$44.3_{\pm 1.3}$	▲ $0.8_{\pm 0.4}$
• SAR [45]	$86.6_{\pm 0.0}$	$59.4_{\pm 0.0}$	$87.4_{\pm 0.0}$	$44.2_{\pm 1.4}$	▲ $0.3_{\pm 0.4}$
• ROID [54]	$87.2_{\pm 1.7}$	$56.6_{\pm 0.5}$	$94.8_{\pm 0.6}$	**44.4**$_{\pm 1.3}$*	▲ $1.7_{\pm 1.0}$
• DeYO [46]	$76.1_{\pm 2.8}$	$60.9_{\pm 0.2}$	$92.7_{\pm 0.5}$	n/a†	▼ $0.9_{\pm 0.9}$
• ***DSBR*** (ours)	**89.9**$_{\pm 0.4}$	**62.7**$_{\pm 0.5}$	**96.5**$_{\pm 0.1}$	$44.0_{\pm 1.1}$	▲ **4.2**$_{\pm 0.5}$
ViT	$87.1_{\pm 0.0}$	$54.4_{\pm 0.0}$	$87.0_{\pm 0.0}$	$37.2_{\pm 0.0}$	▲ $0.0_{\pm 0.0}$
• Tent [82]	$87.4_{\pm 0.0}$	$54.4_{\pm 0.0}$	$87.4_{\pm 0.1}$	$37.4_{\pm 0.2}$	▲ $0.2_{\pm 0.1}$
• SAR [45]	$87.1_{\pm 0.0}$	$54.4_{\pm 0.0}$	$87.0_{\pm 0.0}$	$37.2_{\pm 0.0}$	▲ $0.0_{\pm 0.0}$
• ROID [54]	$88.0_{\pm 0.3}$	$54.5_{\pm 0.8}$	$94.3_{\pm 0.2}$	**38.2**$_{\pm 5.0}$*	▲ $2.3_{\pm 1.6}$
• DeYO [46]	$87.9_{\pm 0.2}$	**55.2**$_{\pm 0.1}$	$93.5_{\pm 0.3}$	n/a†	▲ $2.0_{\pm 0.2}$
• ***DSBR*** (ours)	**88.8**$_{\pm 0.4}$	**55.2**$_{\pm 0.5}$	**95.6**$_{\pm 0.3}$	$36.5_{\pm 0.9}$	▲ **2.6**$_{\pm 0.5}$

follows Scholz et al. [68] and combines UCSF-PDGM [8], EGD [81], and TCGA [4] for three-class glioma subtype classification under the 2021 WHO CNS taxonomy [52]; with 1,146 volumes, it is the smallest dataset and the only 3D one.
We use *ImageNet-C* [32], which comprises 75 corruption domains across 15 corruption types and 5 severity levels on ImageNet [16], to study the high-K regime. Moreover, we use this dataset with multiple challenging scenarios. We provide additional dataset details in Appendix B.

Implementation details. We use two backbones for all 2D datasets: ResNet50-GN [31, 90] and ViT-B/16 [20]. Following seminal works [46, 59], we use ResNet's Group Normalization model to avoid known issues of entropy minimization, such as limited batch sizes, and use the same model for all our ablations. For GliomaMRI, we use volumetric architectures from the MONAI framework [9]: a 3D ResNet10 and a 3D ViT-T/16.

For source-target, i.e., training and adaptation data pairing on each medical dataset, every pair of distinct domains forms one configuration, except for MammoBench and GliomaMRI, where the small training sets motivate leave-one-domain-out pre-training instead. We use a batch size of 32 for pre-training and adaptation on the medical datasets, and 64 on ImageNet-C, matching prior work [46, 59, 82]. In line with [64], for all baselines and our method, we hold out a 20% split that is excluded from adaptation and used solely for evaluation in each medical domain since online accuracy can mask late-stage collapse and the final adapted state is what matters for clinical deployment. For ImageNet-C we follow prior work and adapt on the full corruption domain without a holdout split. We compare DSBR against four state-of-the-art EM-based baselines, all re-implemented for the medical datasets. Throughout, all reported accuracies are balanced accuracies; ROC-AUC results are reported in Appendix H. DSBR uses an EMA decay of $\alpha = 0.9$ unless specified otherwise. We provide additional details in the appendix.

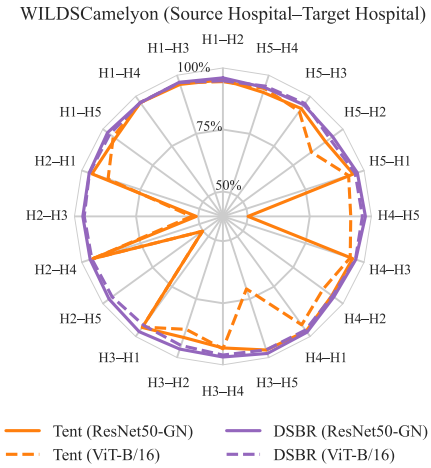

Figure 3: **DSBR stabilizes Entropy Minimization** where 'H' denotes the Hospital domain id. Balanced accuracy results on all *WILDSCamelyon* configurations for Tent (EM) and DSBR.

Table 3: **Accuracy (%) on ImageNet-C on default and wild settings [59] at severity level 5.** Results are averaged over 15 corruption types, except *Mixed Corruptions* where one run comprises a mixture of all 15 corruptions. We report the average improvement over the corresponding non-adapted baseline as Δ Average. DSBR is best **(bold)** or competitive across settings and is the top performer on average.

Model & Method	**Default**	**Label Shifts**	**Mixed Corruptions**	**Batch Size 1**	**Δ Average**
ResNet50-GN	$31.4_{\pm 0.0}$	$31.4_{\pm 0.0}$	$31.4_{\pm 0.0}$	$31.4_{\pm 0.0}$	▲ $0.0_{\pm 0.0}$
• Tent [82]	$32.2_{\pm 0.3}$	$30.3_{\pm 0.7}$	$32.1_{\pm 1.7}$	$30.8_{\pm 0.3}$	▲ $0.0_{+0.7}$
• SAR [59]	$36.7_{\pm 3.0}$	$37.8_{\pm 0.5}$	$38.5_{\pm 0.1}$	$35.8_{\pm 0.3}$	▲ $5.8_{\pm 1.0}$
• ROID [54]	$46.8_{\pm 0.1}$	$2.7_{\pm 0.1}$	$37.6_{\pm 0.1}$	$0.1_{\pm 0.0}$	▼ $9.6_{\pm 0.1}$
• DeYO [46]	$44.1_{\pm 4.9}$	$44.1_{\pm 5.6}$	$39.4_{\pm 1.1}$	**43.9**$_{\pm 4.3}$	▲ $11.4_{\pm 4.0}$
• ***DSBR***($\alpha = 0.99$)	**49.7**$_{\pm 0.1}$	**46.8**$_{\pm 0.2}$	**43.0**$_{\pm 0.2}$	$41.6_{\pm 0.1}$	▲ **13.9**$_{\pm 0.1}$
• ***DSBR***($\alpha = 0.9$)	$48.8_{\pm 0.1}$	$38.2_{\pm 0.3}$	$41.4_{\pm 0.2}$	$38.3_{\pm 0.1}$	▲ $10.2_{\pm 0.2}$
ViT-B/16	$39.8_{\pm 0.0}$	$39.8_{\pm 0.0}$	$39.8_{\pm 0.0}$	$39.8_{\pm 0.0}$	▲ $0.0_{\pm 0.0}$
• Tent [82]	$51.9_{\pm 0.5}$	$53.0_{\pm 0.3}$	$33.1_{\pm 4.5}$	$51.7_{\pm 0.3}$	▲ $7.6_{\pm 1.4}$
• SAR [59]	$54.2_{\pm 0.3}$	$56.2_{\pm 0.1}$	$53.7_{\pm 0.2}$	$56.2_{\pm 0.3}$	▲ $15.3_{\pm 0.2}$
• ROID [54]	$64.6_{\pm 0.1}$	**86.4**$_{\pm 4.9}$	$57.2_{\pm 0.1}$	$0.1_{\pm 0.0}$	▲ $12.2_{\pm 1.3}$
• DeYO [46]	$61.8_{\pm 2.0}$	$62.4_{\pm 1.1}$	$53.6_{\pm 0.8}$	**61.8**$_{\pm 0.9}$	▲ $20.1_{\pm 1.2}$
• ***DSBR***($\alpha = 0.99$)	**64.9**$_{\pm 0.3}$	$56.2_{\pm 0.2}$	**60.9**$_{\pm 0.7}$	$58.2_{\pm 0.1}$	▲ **20.2**$_{\pm 0.3}$
• ***DSBR***($\alpha = 0.9$)	$64.4_{\pm 0.2}$	$42.2_{\pm 0.2}$	$59.5_{\pm 1.0}$	$55.9_{\pm 0.4}$	▲ $15.7_{\pm 0.5}$

5.1 Eliminating Model Collapse for Medical Imaging

In Tab. 2, we provide the balanced accuracy on the medical datasets averaged over all source-target configurations. DSBR achieves the highest average improvement on both backbones, $+4.2\%$ for ResNet and $+2.6\%$ for ViT, considerably more than the next-best baseline (ROID, $+1.7\%$ and $+2.3\%$). On individual datasets, DSBR is the top performer on Histopantum, MammoBench, and WILDSCamelyon for both backbones, and within 0.4% of the best on GliomaMRI. The only non-leading entries are on GliomaMRI, where ROID is best with its consistency component disabled, and DeYO is inapplicable, since both depend on 2D augmentations that do not transfer to 3D MRI volumes. We attribute DSBR's limited gains on the GliomaMRI dataset to the low performance of the source model, which results from the small training set and also limits the gains of all baseline methods. Nevertheless, despite these small gains, DSBR remains within the error margins of the best-performing methods for both backbones.

From Tab. 2, the dataset-level averages do not reveal the instability of EM. Therefore, in figure Figure 3 we report per-configuration accuracy on WILDSCamelyon for Tent and DSBR, which reveals Tent's tendency towards model collapse. In our single-source pairing, Tent collapses to chance-level accuracy ($\approx 50\%$) on a substantial fraction of hospital pairs for both backbones, while DSBR shows no such collapses: every configuration attains near-optimal accuracy (visible as the outer ring in Fig. 3). Targeting prediction bias does not just improve average accuracy; it eliminates the configurations in which Tent fails outright.

5.2 Performance on natural imaging by mitigating bias

Table 3 reports results on ImageNet-C in four settings: standard adaptation, online imbalanced label shifts with imbalance ratio ∞, mixed corruptions across all 15 types, and batch size 1, all at severity 5 [59]. We report DSBR with two EMA decays: the default $\alpha = 0.9$ and a higher $\alpha = 0.99$, motivated by Sec. 4.2: when $K \gg B$ ($K = 1000$, $B = 64$), the per-batch frequency is sharply peaked on the few classes that appear in the batch, and a higher decay softens it across batches before reweighting. On the standard and mixed-corruption settings, DSBR achieves the highest accuracy on both backbones.

For label shifts and batch size 1, DSBR with $\alpha = 0.99$ performs better than DSBR with $\alpha = 0.9$ since each batch is dominated by a single class. Moreover, this DSBR $\alpha = 0.99$ performs best on label shifts for ResNet50-GN ($+2.7\%$ over the next-best baseline) and second-best at batch size 1 on both backbones, demonstrating that the EMA preserves cluster-frequency information across batches even when a single batch is uninformative.

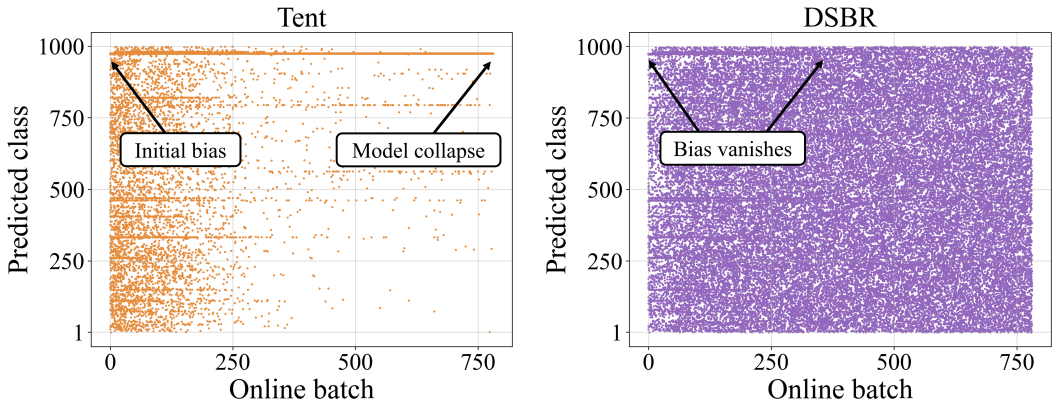

Figure 4: **Model collapse in Tent (EM) and its mitigation by DSBR.** Initial bias is present for both models. **Left:** Tent amplifies bias until the model converges to a single class for almost every input (*Model collapse*) by the end of the episode. **Right:** DSBR progressively dampens bias until it vanishes by approximately batch 400.

Figure 4 visualizes the mechanism behind these gains on a single ImageNet-C Fog (severity 5) episode. Both Tent and DSBR start from the same source model and inherit the same noticeable initial prediction bias toward a small set of dominant classes induced by the corruption. Under Tent, this bias is amplified rather than corrected: within a few hundred batches, predictions concentrate on a single class, and by the end of the episode, the model has collapsed. Under DSBR, the same initial bias is progressively dampened by the importance reweighting; by the second half of the episode, the predicted distribution has spread across the full label space and remains spread thereafter.

5.3 EMA stabilizes DSBR for limited batch sizes

We demonstrate the benefits of using EMA for DSBR in Figure 5. DSBR's effectiveness depends on the quality of the running estimate $\hat{P}_t^{(k)}$ of the predicted class distribution, which in turn depends on the batch size B, the number of classes K, and the EMA decay α. We ablate these by reporting DSBR with $\alpha = 0.9$ and $\alpha = 0$ across batch sizes from 1 to 128, on representative configurations from each end of the K spectrum: ImageNet-C Gaussian Noise at severity 5 ($K = 1000$) and WILDSCamelyon Hospital Nr. $2 \to 5$ ($K = 2$). At batch size 1 and $\alpha = 0$, without EMA, DSBR degrades to standard Tent for both datasets, since the loss reduces to Tent's up to a constant factor of $1/K$. Enabling the EMA recovers near-optimal performance immediately on WILDSCamelyon. For low-K benchmarks, the per-batch estimator alone matches the EMA estimator at moderate batch sizes ($B \geq 4$); for high-K benchmarks, the EMA remains necessary throughout. DSBR's reweighting works straightforwardly when B is on the order of K or larger, and the EMA is what extends it to the regime $B \ll K$.

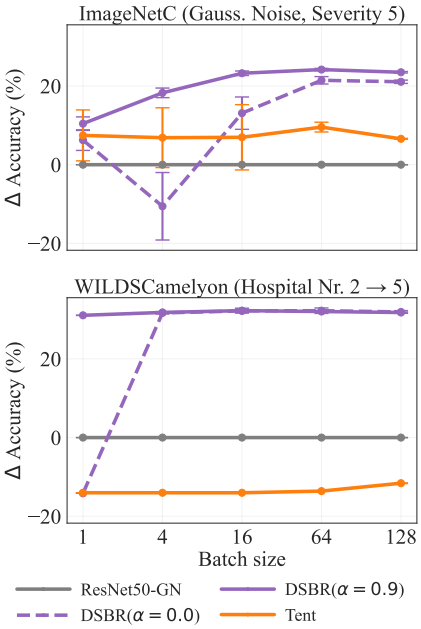

Figure 5: **EMA stabilizes DSBR for limited batch sizes.** Balanced accuracy across batch sizes on ImageNet-C and WILDSCamelyon (ResNet50-GN), evaluated on a held-out 20% split. DSBR remains effective for limited batch sizes.

6 Conclusion

We identify the root cause of entropy-minimization (EM) failure under distribution shift: the merging of feature-space clusters. The merging induces prediction bias, and EM, which refines whatever

clusters are present in feature space, treats the merged region as a single dominant cluster and further compresses it. This reinforces the bias and ultimately results in a single-class collapse. To address this failure, we propose Distribution Shift Bias Reduction (DSBR), which balances predicted-class contributions during adaptation. In contrast to other EM state-of-the-art methods, DSBR does not require sample filtering or domain-specific augmentations and uses only a single hyperparameter (the EMA decay). In addition, DSBR operates solely at test time.

We extensively evaluate DSBR across medical-imaging and natural-vision datasets. Across all settings, DSBR consistently stabilizes test-time adaptation, prevents model collapse, and performs competitively or outperforms current state-of-the-art methods. DSBR depends on an accurate estimate of the predicted-class distribution. This estimate may deteriorate in a few specific cases, such as imbalanced label shifts or extremely small batch sizes relative to the number of classes. However, our experiments demonstrate that correcting a single, mechanistically grounded quantity, the prediction bias that arises from cluster merging is sufficient to stabilize entropy minimization on both medical-imaging and natural-image benchmarks. This provides strong empirical evidence that cluster merging (as reflected by prediction bias) is the primary objective to address when stabilizing EM. More broadly, our findings highlight the importance of a principled mechanistic understanding for developing robust and reliable test-time adaptation methods under distribution shift.

Acknowledgments

This work is supported by DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, Munich Center for Machine Learning (MCML), both sponsored by the Federal Ministry of Research, Technology, Bavaria California Technology Center (BaCaTeC, Bayerisch-Kalifornischen Hochschulzentrum) under BaCaTeC-Förderprojekt Nr. 20 [2026-1], Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 515279324 / SPP 2177, and the German Federal Ministry of Research, Technology and Space (DECIPHER-M, 01KD2420G). DML and JAS received funding from HELMHOLTZ IMAGING, a platform of the Helmholtz Information and Data Science Incubator.

References

[1] S. Agarwal, Z. Zhang, L. Yuan, J. Han, and H. Peng. The unreasonable effectiveness of entropy minimization in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=UfFTBEsLgI.

[2] A. S. Alsolami, W. Shalash, W. Alsaggaf, S. Ashoor, H. Refaat, and M. Elmogy. King abdulaziz university breast cancer mammogram dataset (kau-bcmd). *Data*, 6(11):111, 2021.

[3] S. Ambekar, R. Nasirigerdeh, P. J. Schuffler, L. Felsner, D. M. Lang, and J. A. Schnabel. The mean is the mirage: Entropy-adaptive model merging under heterogeneous domain shifts in medical imaging. *arXiv preprint arXiv:2602.21372*, 2026.

[4] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):170117, 2017.

[5] A. Balendran, C. Beji, F. Bouvier, O. Khalifa, T. Evgeniou, P. Ravaud, and R. Porcher. A scoping review of robustness concepts for machine learning in healthcare. *npj Digital Medicine*, 8(1):38, 2025.

[6] Y. Bar, S. Shaer, and Y. Romano. Protected test-time adaptation via online entropy matching: A betting approach. In *Advances in Neural Information Processing Systems*, 2024.

[7] M. Bernhardt, F. D. S. Ribeiro, and B. Glocker. Failure detection in medical image classification: A reality check and benchmarking testbed. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=VBHuLfnOMf.

[8] E. Calabrese, J. E. Villanueva-Meyer, J. D. Rudie, A. M. Rauschecker, U. Baid, S. Bakas, S. Cha, J. T. Mongan, and C. P. Hess. The university of california san francisco preoperative diffuse glioma mri dataset. *Radiology: Artificial Intelligence*, 4(6):e220058, 2022.

[9] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[10] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi. Contrastive test-time adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.

[11] H. Chen, G. Zheng, A. Awadallah, and Y. Ji. Pathologies of pre-trained language models in few-shot fine-tuning. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 144–153, 2022.

[12] I. Chung, K. Hwang, J. Yoo, and N. Kwak. Mitigating the bias in the model for continual test-time adaptation, 2024. URL https://arxiv.org/abs/2403.01344.

[13] C. Cui, L. Li, H. Cai, Z. Fan, L. Zhang, T. Dan, J. Li, and J. Wang. The chinese mammography database (cmmd): An online mammography database with biopsy confirmed types for machine diagnosis of breast. *The Cancer Imaging Archive*, 1, 2021.

[14] H. Dang, T. Tran, S. Osher, H. Tran-The, N. Ho, and T. Nguyen. Neural collapse in deep linear networks: from balanced to imbalanced data. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[15] M. Z. Darestani, J. Liu, and R. Heckel. Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing. In *International Conference on Machine Learning*, 2022.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[17] R. Deng, W. Bao, T. Wei, and J. He. Panda: Test-time adaptation with negative data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 3551–3559, 2026.

[18] E. Dohmatob, Y. Feng, and J. Kempe. Model collapse demystified: The case of regression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=bioHNTRnQk.

[19] E. Dohmatob, Y. Feng, A. Subramonian, and J. Kempe. Strong model collapse. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=et5l9qPUhm.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[21] H. Evans and D. Snead. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *NPJ Digital Medicine*, 7(1):89, 2024.

[22] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *International Conference on Learning Representations (ICLR)*, 2021.

[23] D. Gambetta, G. Gezici, F. Giannotti, D. Pedreschi, A. Knott, and L. Pappalardo. Learning by surprise: Surplexity for mitigating model collapse in generative ai, 2025. URL https://arxiv.org/abs/2410.12341.

[24] N. P. Gaurav Bhole, S Suba. Mammo-bench: A large-scale benchmark dataset of mammography images. 2025.

[25] S. Goyal, M. Sun, A. Raghunathan, and J. Z. Kolter. Test time adaptation via conjugate pseudo-labels. In *Advances in Neural Information Processing Systems*, 2022.

[26] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, B. Schölkopf, et al. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[27] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

[28] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[29] Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, and S. S. Liu. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*, 2024.

[30] J. Han, J. Na, and W. Hwang. Ranked entropy minimization for continual test-time adaptation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=lHaGLJ65J9.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[32] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*, 2019.

[33] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, 2018.

[34] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

[35] Y. Iwasawa and Y. Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[36] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, 2021.

[37] J. Kazdan, R. Schaeffer, A. Dey, M. Gerstgrasser, R. Rafailov, D. L. Donoho, and S. Koyejo. Collapse or thrive? perils and promises of synthetic data in a self-generating world. *arXiv preprint arXiv:2410.16713*, 2024.

[38] R. Khaled, M. Helal, O. Alfarghaly, O. Mokhtar, A. Elkorany, H. El Kassas, and A. Fahmy. Categorized digital database for low energy and subtracted contrast enhanced spectral mammography images. *The Cancer Imaging Archive*, 16, 2021.

[39] O. Kilim, A. Olar, T. Joó, T. Palicz, P. Pollner, and I. Csabai. Physical imaging parameter variation drives domain shift. *Scientific Reports*, 12(1):21302, 2022.

[40] E. Kim, M. Sun, C. Baek, A. Raghunathan, and J. Z. Kolter. Test-time adaptation induces stronger accuracy and agreement-on-the-line. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=giXUx4VH9t`.

[41] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[42] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

[43] D. Kopans, R. Moore, K. Bowyer, P. Kegelmeyer, and P. Shile. Ddsm: digital database for screening mammography, 2020.

[44] D.-H. Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, International Conference on Machine Learning (ICML)*, page 896, 2013.

[45] J. Lee, D. Das, J. Choo, and S. Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *IEEE International Conference on Computer Vision*, 2023.

[46] J. Lee, D. Jung, S. Lee, J. Park, J. Shin, U. Hwang, and S. Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, 2024.

[47] K. Lekadir, A. F. Frangi, A. R. Porras, B. Glocker, C. Cintas, C. P. Langlotz, E. Weicken, F. W. Asselbergs, F. Prior, G. S. Collins, et al. Future-ai: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *bmj*, 388, 2025.

[48] J. Liang, R. He, and T. Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

[49] R. Lin and Y. You. Optimizing class-level probability reweighting coefficients for equitable prompting accuracy. *arXiv preprint arXiv:2405.07623*, 2024.

[50] Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[51] Y. Liu, P. Kothari, B. van Delft, B. Bellot-Gurlet, T. Mordan, and A. Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[52] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. Ng, S. M. Pfister, G. Reifenberger, et al. The 2021 who classification of tumors of the central nervous system: a summary. *Neuro-oncology*, 23(8):1231–1251, 2021.

[53] S. K. Maharana, S. Mishra, Y. Zhang, S. Niu, T. H. Rafi, J. Hamm, M. Pedersoli, J. Dolz, and Y. Guo. Continual test-time adaptation in computer vision: Methods, benchmarks, and future directions. *arXiv preprint arXiv:2607.08164*, 2026.

[54] R. A. Marsden, M. Döbler, and B. Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2555–2565, 2024.

[55] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

[56] S. Mounsaveng, F. Chiaroni, M. Boudiaf, M. Pedersoli, and I. Ben Ayed. Bag of tricks for fully test-time adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1936–1945, 2024.

[57] Z. Nado, S. Padhy, D. Sculley, A. D'Amour, B. Lakshminarayanan, and J. Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

[58] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, 2022.

[59] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

[60] P. Oza, R. Oza, U. Oza, P. Sharma, S. Patel, P. Kumar, and B. Gohel. Digital mammography Dataset for Breast Cancer Diagnosis Research (DMID). 11 2023. doi: 10.6084/m9.figshare.24522883.v2. URL `https://figshare.com/articles/dataset/_b_Digital_mammography_Dataset_for_Breast_Cancer_Diagnosis_Research_DMID_b_DMID_rar/24522883`.

[61] P. Pandey, A. K. Tyagi, S. Ambekar, and A. Prathosh. Unsupervised domain adaptation for semantic segmentation of nir images through generative latent search. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 413–429. Springer, 2020.

[62] M. Park, H. Won, W. W. Ro, and S. Kim. Rethinking entropy in test-time adaptation: The missing piece from energy duality. *Advances in Neural Information Processing Systems*, 38: 47123–47140, 2026.

[63] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[64] O. Press, R. Shwartz-Ziv, Y. LeCun, and M. Bethge. The entropy enigma: Success and failure of entropy minimization. *arXiv preprint arXiv:2405.05012*, 2024.

[65] M. Roschewitz, G. Khara, J. Yearsley, N. Sharma, J. J. James, É. Ambrózay, A. Heroux, P. Kecskemethy, T. Rijken, and B. Glocker. Automatic correction of performance drift under acquisition shift in medical image classification. *Nature Communications*, 14(1):6608, 2023.

[66] R. Schaeffer, J. Kazdan, A. C. Arulandu, and S. Koyejo. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2025.

[67] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, pages 11539–11551, 2020.

[68] D. Scholz, A. C. Erdur, J. A. Buchner, J. C. Peeken, D. Rueckert, and B. Wiestler. Imbalance-aware loss functions improve medical image classification. In *Medical imaging with deep learning*, 2024.

[69] S. Seto, B.-J. Theobald, F. Danieli, N. Jaitly, and D. Busbridge. Realm: Robust entropy adaptive loss minimization for improved single-sample test-time adaptation, 2023. URL `https://arxiv.org/abs/2309.03964`.

[70] D. S. Shah, H. A. Schwartz, and D. Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5248–5264, 2020.

[71] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[72] L. Shi, M. Wu, H. Zhang, Z. Zhang, M. Tao, and Q. Qu. A closer look at model collapse: From a generalization-to-memorization perspective. *arXiv preprint arXiv:2509.16499*, 2025.

[73] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[74] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020.

[75] Y. Su, Y. Ji, J. Li, H. Ye, and M. Zhang. Beware of model collapse! fast and stable test-time adaptation for robust question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12998–13011, 2023.

[76] Y. Su, X. Xu, and K. Jia. Towards real-world test-time adaptation: Tri-net self-training with balanced normalization. In *AAAI Conference on Artificial Intelligence*, 2024.

[77] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

[78] T. Sun, T. Meng, and Y. Liu. Camelyon 17 challenge: A comparison of traditional machine learning (svm) with the deep learning method. *Wireless Communications and Mobile Computing*, 2022(1):9910471, 2022.

[79] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 2020.

[80] J. Van der Laak, G. Litjens, and F. Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[81] S. R. van der Voort, F. Incekara, M. M. Wijnenga, G. Kapsas, R. Gahrmann, J. W. Schouten, H. J. Dubbink, A. J. Vincent, M. J. van den Bent, P. J. French, et al. The erasmus glioma database (egd): Structural mri scans, who 2016 subtypes, and segmentations of 774 patients with glioma. *Data in brief*, 37:107191, 2021.

[82] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

[83] G. Wang, F. Lyu, and C. Ding. Partition-then-adapt: Combating prediction bias for reliable multi-modal test-time adaptation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=T6RkYsuoMW`.

[84] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 2018.

[85] Q. Wang, O. Fink, L. Van Gool, and D. Dai. Continual test-time domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.

[86] Z. Wang, Y. Luo, L. Zheng, Z. Chen, S. Wang, and Z. Huang. In search of lost online test-time adaptation: A survey. *arXiv preprint arXiv:2310.20199*, 2023.

[87] T. Weerasinghe, R. Tennakoon, W. Chuah, and A. Bab-Hadiashar. Logit-adjusted test-time adaptation under partial class imbalance. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5977–5985. IEEE, 2026.

[88] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[89] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

[90] Y. Wu and K. He. Group normalization. In *European Conference on Computer Vision*, pages 3–19, 2018.

[91] Z. Xiao and C. G. Snoek. Beyond model adaptation at test time: A survey. *arXiv preprint arXiv:2411.03687*, 2024.

[92] J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, and H.-I. Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609, 2024.

[93] Y. Yu, L. Sheng, R. He, and J. Liang. Benchmarking test-time adaptation against distribution shifts in image classification. *arXiv preprint arXiv:2307.03133*, 2023.

[94] N. Zamanitajeddin, M. Jahanifar, K. Xu, F. Siraj, and N. Rajpoot. Benchmarking domain generalization algorithms in computational pathology. *arXiv preprint arXiv:2409.17063*, 2024.

[95] M. Zhang, S. Levine, and C. Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, pages 38629–38642, 2022.

[96] Q. Zhang, Y. Bian, X. Kong, P. Zhao, and C. Zhang. Come: Test-time adaption by conservatively minimizing entropy. *arXiv preprint arXiv:2410.10894*, 2024.

[97] H. Zhao, Y. Liu, A. Alahi, and T. Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning*, 2023.

[98] Z. Zhong, J. Cui, Y. Yang, X. Wu, X. Qi, X. Zhang, and J. Jia. Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19550–19560, 2023.

Appendix

A Additional Implementation Details **18**

B Additional Datasets Details **18**

C Pre-training Results **19**

D Compute Resources **21**

E High Gradient Norm is an Effect of Model Collapse rather than its Cause **22**

F Ability to Recover Rare Classes and Calibrate Uncertainty **23**

G Entropy filtering as one solution for cluster separation **25**

H ROC-AUC reveals a structural limit of EM **26**

A Additional Implementation Details

Our method's code is based on Domainbed [27] and is efficient. We pre-train all source models using a unified DomainBed-based pipeline [27], modified to allow for early stopping when validation losses stop decreasing and gradient accumulation for 3D volumes where no full batch fits into VRAM. Each domain is split 80/20 into IN and OUT subsets. Pre-training uses source-IN for training and source-OUT for validation, while adaptation uses target-IN as the stream of inputs the model adapts on and target-OUT as the held-out evaluation set on which we report performance. Evaluating on a held-out target split rather than online during adaptation is a deliberate choice, because online accuracy can mask late-stage collapse: a model can produce reasonable predictions for the first several thousand samples and then collapse to a trivial constant predictor in the final thousand, leaving the running average of online accuracy looking acceptable while the model is in fact unusable. In clinical settings, where the input stream has no natural endpoint, we need to be confident that the model does not degrade at any point during deployment, which makes the *final* adapted state more informative than the online accuracy. We use batch size 32 on the medical datasets (constrained by VRAM for the volumetric inputs) and 64 on ImageNet-C, matching prior work. For ImageNet-C we follow prior-work convention instead [46, 58, 59, 82]: we use publicly available ImageNet-pretrained checkpoints as source models and adapt on the full target corruption domain without a held-out split. All reported numbers are averaged over 5 seeds for the medical datasets and 3 seeds for ImageNet-C; 2 standard deviations across seeds are calculated using *pandas* library std function and reported alongside each mean.

For the re-implementation of related work's baselines [46, 54, 59, 82] we stick closely to the official repositories and use the default hyperparameters and optimizers of each method.

Algorithm 1 Distribution **S**hift **B**ias **R**eduction (DSBR)

Input: Test samples $\mathcal{X} = \{x_i\}_{i=1}^{N^{\text{test}}}$, model $f_\theta(x)$ with trainable parameters $\tilde{\theta} \subset \theta$, batch size B, optimizer $\mathcal{O}_\eta(\cdot)$ with learning rate $\eta > 0$, EMA decay α

Output: Predictions $\{\hat{y}_i\}_{i=1}^{N^{\text{test}}}$, adapted parameters $\tilde{\theta}$.

Initialize model parameters $\tilde{\theta}$ with a chosen scheme.
Initialize $\hat{P}^{(k)} \leftarrow 1/K$ for all $k \in \{1, \dots, K\}$
for $b = 1$ to N^{test} step B **do**
  $\mathcal{B} \leftarrow \{b, \dots, \min(b + B - 1, N)\}$
  $\hat{y}_i \leftarrow \arg\max_j \; f_\theta(x_i)_j$ **for** every $i \in \mathcal{B}$
  $B^{(k)} \leftarrow |\{i \in \mathcal{B} \mid \hat{y}_i = k\}|$ **for** every $k \in \{1, \dots, K\}$
  $\hat{P}^{(k)} \leftarrow \alpha\, \hat{P}^{(k)} + (1 - \alpha)\, B^{(k)}/|\mathcal{B}|$ **for** every $k \in \{1, \dots, K\}$ ▷ Eq. (4)
  $l_i \leftarrow H(f_\theta(x_i))$ **for** every $i \in \mathcal{B}$
  $\mathcal{L} \leftarrow \dfrac{1}{B} \displaystyle\sum_{i \in \mathcal{B}} \dfrac{l_i}{K \cdot \hat{P}^{(\hat{y}_i)}}$ ▷ Eq. (5)
  Update $\tilde{\theta} \leftarrow \mathcal{O}_\eta\Big(\tilde{\theta},\, \nabla_{\tilde{\theta}} \mathcal{L}\Big)$
end for

Full Algorithm and Optimizer. Algorithm 1 provides the pseudo-code for our method. As optimizer $\mathcal{O}_\eta$ we use Adam [41] with default hyperparameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and $\lambda = 0$ (no weight decay). For the learning rate η we use the same per batch size defaults as Wang et al. [82].

Comparison with results from other works. For the ImageNet-C datasets, our re-implementation obtains higher results than the results reported in baseline methods. We verified that this is due to the usage of pip packages instead of conda. We encourage the community to adopt pip to implement Tent for baselines.

B Additional Datasets Details

We evaluate on four medical datasets and one natural vision dataset. We do not cover the full breadth of medical imaging in this evaluation, but the four datasets below span a substantial cross-section of it: three modalities (microscopy, X-ray, and MRI), two- and three-dimensional inputs, and several anatomical sites (breast, brain, and a range of organs covered by the histopathology datasets). Notably, these datasets exhibit real-world distribution shifts and the class imbalance typical of the medical domain. We quantify the latter using the imbalance ratio $\text{IR} = N_{\text{majority}}/N_{\text{minority}}$. We

report, for each dataset, the maximum IR across its domains, reflecting the most skewed setting a model may encounter. GliomaMRI is the most imbalanced ($\mathrm{IR} = 28.31$), followed by MammoBench ($\mathrm{IR} = 5.15$) and Histopantum ($\mathrm{IR} = 1.65$), whereas WILDS-Camelyon is the only balanced dataset ($\mathrm{IR} = 1.00$).

WILDSCamelyon [42] is a histopathology dataset for breast cancer metastasis detection, derived from Camelyon17 [78] by extracting patches from whole-slide images for binary cancerous-vs-noncancerous classification; the five domains correspond to five hospitals, each with its own staining protocol, scanner, and lab procedure, totaling 455,954 patches. License: Creative Commons Zero (CC0)

Histopantum [94] is a histopathology dataset for binary tumor classification on patches derived from TCGA [88], where the four domains correspond to four cancer types: colorectal, ovarian, stomach, and uterus; adaptation across these domains is biological rather than instrumental. The dataset comprises 281,142 patches sourced from whole-slide images annotated by experienced pathologists License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

MammoBench [24] aggregates six mammography datasets [2, 13, 38, 43, 55, 60] into a unified Normal-vs-Abnormal dataset; we exclude CMMD [13] because it does not contain Normal cases, leaving 14,529 images across the remaining datasets, each treated as a domain. The dataset underwent rigorous preprocessing including breast segmentation, pectoral muscle removal, and intelligent cropping to ensure consistency across images from diverse sources while preserving clinically relevant features. License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

GliomaMRI follows Scholz et al. [68] and combines three 3D brain MRI datasets (UCSF-PDGM [8], EGD [81], and TCGA [4]) for three-class glioma subtype classification under the 2021 WHO classification of CNS tumors [52], with the three classes being IDH-wildtype glioblastoma, IDH-mutant 1p/19q-intact astrocytoma, and IDH-mutant 1p/19q-codeleted oligodendroglioma; with 1,146 volumes it is the smallest dataset and the only one with volumetric inputs. The UCSF-PDGM dataset includes 501 subjects with histopathologically-proven diffuse gliomas imaged with a standardized 3 Tesla preoperative brain tumor MRI protocol featuring predominantly 3D imaging and advanced diffusion and perfusion imaging techniques. License: TCIA Data Usage Policy (open access with attribution); TCGA: NIH Genomic Data Sharing Policy (open and controlled access tiers); EGD: Open access

ImageNet-C [32] provides 75 corruption domains, each consisting of one of 15 corruption types at one of 5 severity levels, applied to the ImageNet [16] validation set, with 50,000 images and 1,000 classes per domain; we use it to compare DSBR against prior TTA work calibrated on this dataset and to study the high-K regime. The corruption types span noise, blur, weather, digital, and additional degradation categories to comprehensively evaluate classifier robustness. License: Creative Commons Attribution 4.0 International (CC BY 4.0) (for corruptions); ImageNet base dataset: Non-commercial research and educational use

C Pre-training Results

We perform pre-training on four multi-domain medical imaging datasets containing 5, 4, 3, and 5 domains for **WILDSCamelyon**, **Histopantum**, **GliomaMRI**, and **MammoBench**, respectively. The strategy for selecting training domains differs by dataset. For **WILDSCamelyon** and **Histopantum**, we use each individual domain as a separate training set, since these datasets are large enough to produce reasonable models from a single source. This also matches the realistic scenario in which a practitioner has access to data from only one institution. For **GliomaMRI** and **MammoBench**, the per-domain data is too small for stand-alone training, so we instead train on every leave-one-out combination of domains. With two backbone architectures evaluated per dataset, this yields 10, 8, 6, and 10 pretrained models per dataset, for a total of **34 pretrained models**.

Per-dataset results are reported in Tables 4, 5, 6, and 7. The histopathology datasets, **WILDSCamelyon** and **Histopantum**, yield strong pretrained models: their Avg. Source OUT scores approach 100%, indicating that the models reliably fit the source distribution. In contrast, **GliomaMRI** and **MammoBench** exhibit substantially weaker in-domain generalization, with Avg. Source OUT scores

well below this level, suggesting that the available training data is insufficient to even fit the source distribution well, let alone generalize to held-out domains.

Table 4: Balanced accuracy (%) after pre-training on the **WILDSCamelyon** benchmark across five hospitals used as separate domains. Each row represents a hospital dataset which was used for training; per-hospital columns report performance on the respective *IN* and *OUT* splits. **Avg. Source** averages over training hospitals, **Avg. Target** over out-of-distribution dataset.

(a) WILDSCamelyon – ViT-B/16

Domain	Hospital 1		Hospital 2		Hospital 3		Hospital 4		Hospital 5		Avg. Source		Avg. Target	
Split	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT
Training Domains (Source)														
Hospital 1	99.0	98.9	94.6	94.1	96.3	96.5	96.5	96.6	94.1	94.1	99.0	98.9	95.4	95.3
Hospital 2	81.9	82.5	98.5	98.1	66.7	67.0	89.6	89.6	58.9	58.8	98.5	98.1	74.3	74.5
Hospital 3	95.4	95.3	86.3	86.2	98.6	98.6	91.7	91.5	79.0	79.5	98.6	98.6	88.1	88.1
Hospital 4	92.6	92.8	89.0	89.0	91.7	91.6	98.8	98.6	76.6	76.7	98.8	98.6	87.5	87.5
Hospital 5	92.2	92.2	85.0	84.5	90.0	90.1	91.7	91.8	98.8	98.8	98.8	98.8	89.7	89.6

(b) WILDSCamelyon – ResNet50-GN

Domain	Hospital 1		Hospital 2		Hospital 3		Hospital 4		Hospital 5		Avg. Source		Avg. Target	
Split	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT
Training Domains (Source)														
Hospital 1	99.2	99.0	94.9	94.8	95.3	95.4	96.4	96.5	79.6	79.7	99.2	99.0	91.6	91.6
Hospital 2	90.7	90.5	99.5	98.9	62.4	63.0	92.1	92.2	64.3	64.2	99.5	98.9	77.4	77.5
Hospital 3	95.4	95.3	88.0	87.3	99.1	98.9	90.2	90.4	96.1	96.4	99.1	98.9	92.4	92.4
Hospital 4	94.2	94.6	93.7	93.6	84.2	84.4	98.9	98.8	62.1	62.1	98.9	98.8	83.6	83.7
Hospital 5	93.3	93.5	90.1	90.0	92.1	92.3	91.3	91.3	98.6	98.6	98.6	98.6	91.7	91.8

Table 5: Balanced accuracy (%) after pre-training on the **Histopantum** benchmark across four organ datasets used as separate domains. Each row represents the set of datasets which were used for training; per-dataset columns report performance on the respective *IN* and *OUT* splits. **Avg. Source** averages over training datasets, **Avg. Target** over out-of-distribution datasets.

(a) Histopantum – ViT-B/16

Domain	colon		ovarian		stomach		uterus		Avg. Source		Avg. Target	
Split	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT
Training Domains (Source)												
colon	98.8	98.2	92.3	91.9	86.1	85.3	84.0	84.1	98.8	98.2	87.5	87.1
ovarian	87.3	86.6	99.1	98.9	81.2	79.9	93.9	94.2	99.1	98.9	87.5	86.9
stomach	89.1	88.8	85.5	85.3	98.4	97.0	86.2	86.5	98.4	97.0	86.9	86.9
uterus	83.9	83.2	96.1	96.0	84.8	83.4	99.0	98.9	99.0	98.9	88.3	87.5

(b) Histopantum – ResNet50-GN

Domain	colon		ovarian		stomach		uterus		Avg. Source		Avg. Target	
Split	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT
Training Domains (Source)												
colon	99.3	98.5	91.3	91.6	83.9	82.8	85.1	85.4	99.3	98.5	86.8	86.6
ovarian	87.3	86.9	99.1	98.6	85.0	83.6	91.9	92.1	99.1	98.6	88.1	87.5
stomach	90.8	90.5	86.7	86.4	97.7	96.7	79.3	79.2	97.7	96.7	85.6	85.4
uterus	83.1	82.7	96.5	96.3	84.0	82.3	98.9	99.0	98.9	99.0	87.9	87.1

Table 6: Balanced accuracy (%) after pre-training on the **GliomaMRI** benchmark across three dataset used as separate domains. Each row represents the set of datasets which were used for training; per-dataset columns report performance on the respective *IN* and *OUT* splits. **Avg. Source** averages over training datasets, **Avg. Target** over out-of-distribution datasets.

(a) GliomaMRI – MONAI ViT-T/16

Domain	erasmus		tcga		ucsf		Avg. Source		Avg. Target	
Split	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT
Training Domains (Source)										
erasmus, tcga	88.7	47.4	98.5	45.4	43.6	33.0	93.6	46.4	43.6	33.0
erasmus, ucsf	100.0	50.8	37.6	41.1	100.0	33.1	100.0	42.0	37.6	41.1
tcga, ucsf	37.0	37.6	100.0	43.7	100.0	36.4	100.0	40.0	37.0	37.6

(b) GliomaMRI – MONAI ResNet10

Domain	erasmus		tcga		ucsf		Avg. Source		Avg. Target	
Split	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT
Training Domains (Source)										
erasmus, tcga	100.0	64.0	100.0	40.7	54.4	41.1	100.0	52.4	54.4	41.1
erasmus, ucsf	100.0	55.2	42.9	41.7	100.0	44.9	100.0	50.0	42.9	41.7
tcga, ucsf	42.9	45.8	100.0	50.0	100.0	40.2	100.0	45.1	42.9	45.8

Table 7: Balanced accuracy (%) after pre-training on the **MammoBench** benchmark across five dataset used as separate domains. Each row represents the set of datasets which were used for training; per-dataset columns report performance on the respective *IN* and *OUT* splits. **Avg. Source** averages over training datasets, **Avg. Target** over out-of-distribution datasets.

(a) MammoBench – ViT-B/16

Domain	cdd-cesm		ddsm		dmid		inbreast		kau-bcmd		Avg. Source		Avg. Target	
Split	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT
Training Domains (Source)														
cdd-cesm, ddsm, dmid, inbreast	90.8	66.5	81.3	75.0	100.0	80.4	99.8	61.8	48.5	48.2	93.0	70.9	48.5	48.2
cdd-cesm, ddsm, dmid, kau-bcmd	90.0	71.7	74.3	72.4	99.7	75.1	60.7	65.0	62.3	61.8	81.6	70.2	60.7	65.0
cdd-cesm, ddsm, inbreast, kau-bcmd	82.6	60.1	73.8	70.9	52.7	52.4	99.8	50.9	72.4	68.4	82.1	62.6	52.7	52.4
cdd-cesm, dmid, inbreast, kau-bcmd	90.4	67.6	56.9	54.5	99.7	77.1	100.0	52.5	82.5	79.5	93.2	69.2	56.9	54.5
ddsm, dmid, inbreast, kau-bcmd	53.8	52.0	77.9	69.9	99.7	75.9	100.0	59.5	59.9	57.8	84.4	65.8	53.8	52.0

(b) MammoBench – ResNet50-GN

Domain	cdd-cesm		ddsm		dmid		inbreast		kau-bcmd		Avg. Source		Avg. Target	
Split	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT	IN	OUT
Training Domains (Source)														
cdd-cesm, ddsm, dmid, inbreast	86.0	73.7	79.0	76.1	100.0	82.0	100.0	58.6	50.1	51.0	91.2	72.6	50.1	51.0
cdd-cesm, ddsm, dmid, kau-bcmd	90.2	68.2	76.5	72.3	100.0	78.8	59.5	65.6	74.4	76.3	85.3	73.9	59.5	65.6
cdd-cesm, ddsm, inbreast, kau-bcmd	91.6	74.1	76.4	73.5	63.7	61.6	100.0	56.2	77.0	71.4	86.2	68.8	63.7	61.6
cdd-cesm, dmid, inbreast, kau-bcmd	80.6	76.8	54.6	53.7	88.7	81.6	94.6	57.1	90.5	88.8	88.6	76.1	54.6	53.7
ddsm, dmid, inbreast, kau-bcmd	62.8	65.2	78.1	76.0	97.9	77.5	98.8	59.5	79.9	78.3	88.7	72.8	62.8	65.2

D Compute Resources

For our experiments we utilize a SLURM cluster with varying hardware. GPUs used include NVIDIA RTX A5000 24GB, A100 40GB, H100 40GB, and H100 80GB. Except for the experiments on the 3D MRI dataset GliomaMRI, all experiments can be reproduced on a GPU with at least 24GB of VRAM. For pre-training the GliomaMRI models, we required the H100 80GB GPUs as the 3D volumes and models consumed more than 50GB of VRAM.

Even though we utilize different GPUs in our adaptation experiments, the total consumed time per experiment is almost equal for all GPUs. Tab. 8 and 9 report the average runtime of each method for every dataset. Note, that these numbers refer to a single adaptation run of a single configuration and we repeat each experiment for 5 and 3 seeds for medical datasets and ImageNet-C respectively.

Table 8: Average adaptation runtime in seconds per adaptation source–target configuration for medical datasets. **Bold** = lowest runtime per column; underline = second lowest. DSBR consistently ranks among the fastest methods across both architectures, matching the baseline with the lowest runtime in nearly every setting.

Model Type	Method	GliomaMRI	Histopantum	MammoBench	WILDSCamelyon
ResNet	Tent [82]	22	**82**	63	**149**
	SAR [59]	26	100	64	209
	ROID [54]	**21**	106	65	207
	DeYO [46]	n/a	85	**61**	167
	DSBR	23	**82**	63	**149**
ViT	Tent [82]	20	**162**	**70**	**395**
	SAR [59]	19	236	79	613
	ROID [54]	**18**	241	75	551
	DeYO [46]	n/a	184	74	426
	DSBR	19	167	**70**	396

Table 9: Average adaptation runtime in seconds per adaptation domain (corruption) for all ImageNet-C based experiments. **Bold** = lowest runtime per column; underline = second lowest. DSBR achieves the lowest or second-lowest runtime in the majority of settings, making it a computationally efficient method while remaining competitive in accuracy.

Model Type	Method	Batch Size 1	Default	Label Shifts	Mixed Corruptions
ResNet	Tent [82]	737	77	**111**	876
	SAR [59]	1349	101	175	1418
	ROID [54]	1632	98	166	1173
	DeYO [46]	**692**	81	131	896
	DSBR($\alpha = 0.99$)	734	82	117	**802**
	DSBR($\alpha = 0.9$)	717	**75**	115	817
ViT	Tent [82]	**555**	186	**428**	**2967**
	SAR [59]	1142	241	808	6118
	ROID [54]	1273	206	526	4456
	DeYO [46]	686	182	505	4058
	DSBR($\alpha = 0.99$)	612	**167**	478	3162
	DSBR($\alpha = 0.9$)	639	172	477	3362

For example, WILDSCamelyon has a total number of 20 source–target configurations, so the total runtime of each method needs to be multiplied by $20(\text{configurations}) \cdot 5(\text{seeds}) = 100$.

To pretrain all models using the codebase build on Domainbed [27], we require approximately 10 hours for each model on a dataset.

E High Gradient Norm is an Effect of Model Collapse rather than its Cause

Niu et al. [59] attribute model collapse to a small subset of samples that produce disproportionately large gradient updates and destabilize adaptation. Their argument is empirical: during runs that end in collapse, the per-batch gradient norm grows sharply just before predictions concentrate on a single class. The proposed remedy is to identify and discard high-gradient-norm samples before they contribute to the update.

We test this attribution directly. For each adaptation configuration we precompute every sample's *initial gradient norm* by running the model on it as a single-sample batch and recording the resulting gradient before any update is applied. This gives us a per-sample ranking of the gradient signal each input would contribute to the very first adaptation step. We then form subsets of the data corresponding to fixed percentiles of this ranking and run full adaptation on each subset separately. If the high-gradient-norm hypothesis is correct, subsets containing the highest-percentile samples should produce visibly worse outcomes than subsets containing only low-percentile samples.

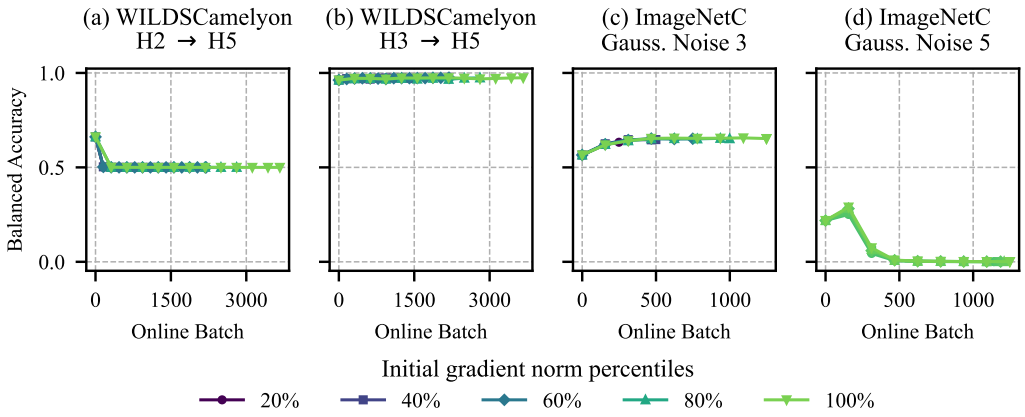

Figure 6: Initial gradient norm has no effect on entropy minimization. We compute the initial gradient norm for each sample by running adaptation on a pre-trained ResNet50-GN with batch size 1 and resetting the model after each step. Afterward, we run normal adaptation on varying percentiles of gradient norms.

Fig. 6 shows the result. Performance is essentially equal across gradient-norm percentiles, both in configurations where adaptation succeeds (b, c) and in configurations where it fails (a, d). The samples with the largest initial gradient norms are not, on their own, harmful to adaptation: removing them does not prevent collapse in the failure cases, and including them does not damage adaptation in the success cases.

We therefore argue that the temporal correlation observed by Niu et al. [59] is read in the wrong direction. The high gradient norms that precede collapse are an effect of the model already drifting toward a degenerate state, in which the loss surface becomes locally steep. Whatever drives collapse must therefore originate upstream of the gradient signal, in the geometry of the feature space. We note that the SAR algorithm proposed by Niu et al. [59] does not filter on gradient norms but on entropy in practice, motivated as a model- and dataset-independent proxy for gradient-norm filtering; we state in Sec. G that entropy filtering does work, but for a reason different from the one this proxy implies.

F Ability to Recover Rare Classes and Calibrate Uncertainty

Recovery of rare classes at test-time. As described in methodology, DSBR up-weights each sample's contribution to the entropy loss through inverse-frequency reweighting. Through the EMA estimate of cluster frequencies, this reweighting also persists across batches in which a class is absent. The predicted distribution therefore stays spread across classes rather than collapsing to one, allowing rare classes suppressed by the domain shift to re-emerge in DSBR's predictions.

To further highlight this behavior, which we term class recovery, we designed an additional experiment in Tab. 10. Class recovery refers to the reappearance, after adaptation, of classes that the source model rarely or never predicts under domain shift, with their predicted frequencies moving closer to their true frequencies. The experiment also extends the analysis in Tab. 1 of the main paper. For ImageNet-C, we track the classes with the lowest prediction probability before adaptation and measure whether Tent or DSBR are able to recover them. DSBR's predictions are closer to the true class frequency in 8 of 10 cases, Tent is closer in one case, and the remaining case is a tie. This shows that DSBR is able to recover suppressed classes more reliably.

Addressing uncertainty. To demonstrate the ability to handle uncertain scenarios, Tab. 11 also reports the Expected Calibration Error (ECE) [28] for all four medical datasets on both backbones, where lower ECE is better. With ResNet, DSBR achieves the best average ECE change among all TTA methods. It also ranks second on Histopathology and WILDSCamelyon, where it is the only method that reduces ECE on both. With ViT, SAR and Tent obtain the lowest average ECE change, while DSBR increases ECE only moderately, by less than ROID, whose calibration is unstable on

Table 10: **Recovery of rare classes at test-time** on ImageNet-C with ResNet50-GN. Predicted class distributions before and after Tent (Entropy Minimization) and DSBR (ours). DSBR recovers the rare classes more often than Tent.

Class from ImageNet	True	Before	DSBR	Tent	Method closer to True (Best)
African elephant	0.09%	0.00%	**0.04%**	0.01%	DSBR
Backpack	0.06%	0.00%	**0.03%**	0.01%	DSBR
Breakwater	0.08%	0.00%	**0.05%**	0.01%	DSBR
Chocolate sauce	0.11%	0.00%	0.16%	**0.09%**	Tent
Garden spider	0.16%	0.00%	**0.15%**	0.10%	DSBR
Groom	0.04%	0.00%	**0.04%**	0.02%	DSBR
Median (other)	0.10%	0.08%	**0.10%**	0.09%	DSBR
Pizza	0.11%	0.01%	**0.06%**	0.02%	DSBR
Typewriter kbd	0.07%	0.01%	**0.03%**	0.01%	DSBR
Unicycle	0.05%	0.01%	**0.06%**	**0.06%**	Both
Wolf spider	0.11%	0.00%	**0.14%**	0.05%	DSBR

GliomaMRI. Notably, DSBR relies solely on the proposed reweighting of the entropy loss, without the auxiliary tasks or models, smoothing, weight ensembling, or augmentation-based consistency used by other baselines. Such components are orthogonal to our method.

Table 11: **Addressing uncertainty** with ResNet and ViT backbones using ECE as the uncertainty metric (lower is better). The subscripts show the change relative to the source model, where a negative value indicates a decrease in ECE and therefore an improvement.

Model & Method	Histopantum	MammoBench	WILDSCamelyon	GliomaMRI	Δ Average
ResNet	9.9 $^{0.0}_{\pm 0.0}$	38.4 $^{0.0}_{\pm 0.0}$	8.4 $^{0.0}_{\pm 0.0}$	16.4 $^{0.0}_{\pm 0.0}$	$^{0.0}_{\pm 0.0}$
• Tent [82]	**8.5** $^{\mathbf{-1.4}}_{\pm 0.1}$	39.0 $^{0.6}_{\pm 0.1}$	9.9 $^{1.5}_{\pm 0.0}$	19.1 $^{2.7}_{\pm 1.3}$	$^{0.9}_{\pm 0.4}$
• SAR [45]	9.9 $^{0.0}_{\pm 0.0}$	**38.4** $^{\mathbf{0.0}}_{\pm 0.0}$	8.4 $^{0.0}_{\pm 0.0}$	19.2 $^{2.8}_{\pm 1.1}$	$^{0.7}_{\pm 0.3}$
• ROID [54]	11.4 $^{1.5}_{\pm 1.3}$	46.0 $^{7.6}_{\pm 2.8}$	**2.6** $^{\mathbf{-5.7}}_{\pm 0.3}$	**18.8** $^{\mathbf{2.4}}_{\pm 1.2}$	$^{1.4}_{\pm 1.4}$
• DeYO [46]	22.3 $^{12.4}_{\pm 3.1}$	44.8 $^{6.4}_{\pm 0.5}$	5.0 $^{-3.3}_{\pm 0.3}$	n/a	$^{5.2}_{\pm 1.3}$
• ***DSBR***($\alpha = 0.9$)	9.6 $^{-0.3}_{\pm 0.5}$	41.7 $^{3.2}_{\pm 0.8}$	2.9 $^{-5.5}_{\pm 0.1}$	19.6 $^{3.2}_{\pm 1.7}$	$^{\mathbf{0.2}}_{\pm 0.8}$
ViT	8.6 $^{0.0}_{\pm 0.0}$	41.5 $^{0.0}_{\pm 0.0}$	7.3 $^{0.0}_{\pm 0.0}$	34.7 $^{0.0}_{\pm 0.0}$	$^{0.0}_{\pm 0.0}$
• Tent [82]	**8.7** $^{\mathbf{0.1}}_{\pm 0.0}$	**41.5** $^{\mathbf{0.0}}_{\pm 0.2}$	8.8 $^{1.5}_{\pm 0.2}$	**34.6** $^{\mathbf{-0.1}}_{\pm 1.2}$	$^{0.4}_{\pm 0.4}$
• SAR [45]	**8.7** $^{\mathbf{0.0}}_{\pm 0.4}$	**41.5** $^{\mathbf{0.0}}_{\pm 0.0}$	7.3 $^{0.0}_{\pm 0.0}$	34.7 $^{0.0}_{\pm 0.0}$	$^{\mathbf{0.0}}_{\pm 0.1}$
• ROID [54]	8.8 $^{0.2}_{\pm 0.3}$	42.7 $^{1.2}_{\pm 0.7}$	**2.0** $^{\mathbf{-5.3}}_{\pm 0.2}$	45.8 $^{11.1}_{\pm 8.8}$	$^{1.8}_{\pm 2.5}$
• DeYO [46]	9.1 $^{0.4}_{\pm 0.5}$	43.1 $^{1.6}_{\pm 0.2}$	3.8 $^{-3.5}_{\pm 1.9}$	n/a	$^{-0.5}_{\pm 0.9}$
• ***DSBR***($\alpha = 0.9$)	11.3 $^{2.7}_{\pm 0.6}$	44.5 $^{3.1}_{\pm 0.4}$	4.3 $^{-3.0}_{\pm 0.3}$	36.6 $^{1.9}_{\pm 0.6}$	$^{1.2}_{\pm 0.5}$

Severe class imbalance. Beyond the datasets in the main paper, which already exhibit the class imbalance typical of the medical domain, we also evaluate under the imbalanced protocol of Weerasinghe et al. [87] (Table 1). We apply it to all ImageNet-C corruptions on both backbones and report balanced accuracy (BA), which we provide in Tab. 12. With ViT, DSBR improves average BA over the source model by $+14.2$ (54.0 ± 3.0), on par with the best method, ROID [54] ($+16.3$; 56.1 ± 3.0), within error margins, and well above Tent [82] ($+5.9$; 45.7 ± 3.3). ResNet leads to a similar conclusion. Together with the wild TTA scenarios in Tab. 3 of the main paper, this indicates that DSBR adapts stably under severe imbalance, whereas methods such as ROID collapse under batch size 1 and label shift on the ResNet backbone.

	Noise			Blur				Weather				Digital				
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
ResNet	22.1$_{\pm 0.0}$	23.0$_{\pm 0.0}$	22.0$_{\pm 0.0}$	19.8$_{\pm 0.0}$	11.4$_{\pm 0.0}$	21.5$_{\pm 0.0}$	25.0$_{\pm 0.0}$	40.2$_{\pm 0.0}$	**46.9**$_{\pm 0.0}$	34.0$_{\pm 0.0}$	**68.8**$_{\pm 0.0}$	36.2$_{\pm 0.0}$	18.5$_{\pm 0.0}$	29.3$_{\pm 0.0}$	52.6$_{\pm 0.0}$	31.4$_{\pm 0.0}$
• Tent	24.0$_{\pm 1.5}$	22.9$_{\pm 0.8}$	23.0$_{\pm 1.6}$	19.1$_{\pm 1.5}$	11.1$_{\pm 3.5}$	21.7$_{\pm 2.7}$	25.5$_{\pm 1.6}$	37.7$_{\pm 4.2}$	46.2$_{\pm 1.4}$	25.3$_{\pm 4.5}$	67.9$_{\pm 0.3}$	37.0$_{\pm 5.3}$	17.5$_{\pm 3.5}$	33.8$_{\pm 0.3}$	52.9$_{\pm 10.6}$	31.1$_{\pm 2.9}$
• SAR	24.1$_{\pm 1.6}$	23.4$_{\pm 1.3}$	23.4$_{\pm 1.4}$	**20.3**$_{\pm 1.9}$	11.4$_{\pm 3.8}$	22.0$_{\pm 3.3}$	25.6$_{\pm 1.6}$	37.9$_{\pm 3.7}$	45.7$_{\pm 1.6}$	29.4$_{\pm 8.4}$	68.0$_{\pm 0.4}$	36.8$_{\pm 4.6}$	17.7$_{\pm 3.6}$	32.5$_{\pm 1.5}$	52.7$_{\pm 10.4}$	31.4$_{\pm 3.3}$
• ROID	**31.8**$_{\pm 0.8}$	**31.1**$_{\pm 2.0}$	**31.3**$_{\pm 2.0}$	17.8$_{\pm 1.4}$	**16.3**$_{\pm 3.2}$	**27.3**$_{\pm 4.2}$	29.4$_{\pm 2.9}$	**42.4**$_{\pm 3.3}$	45.1$_{\pm 0.9}$	**46.1**$_{\pm 2.9}$	68.5$_{\pm 0.2}$	**42.3**$_{\pm 6.2}$	**27.0**$_{\pm 3.8}$	**43.4**$_{\pm 1.4}$	53.5$_{\pm 9.8}$	**36.9**$_{\pm 3.0}$
• DeYO	28.8$_{\pm 0.9}$	27.8$_{\pm 1.7}$	28.6$_{\pm 3.7}$	18.5$_{\pm 1.6}$	13.8$_{\pm 4.0}$	25.4$_{\pm 4.9}$	27.1$_{\pm 2.3}$	38.9$_{\pm 4.9}$	44.1$_{\pm 1.9}$	22.1$_{\pm 19.2}$	68.6$_{\pm 0.6}$	39.7$_{\pm 5.8}$	21.1$_{\pm 4.2}$	42.9$_{\pm 2.0}$	53.9$_{\pm 10.1}$	33.4$_{\pm 4.5}$
• ***DSBR*** (α=0.99)	27.8$_{\pm 1.3}$	27.8$_{\pm 0.5}$	27.2$_{\pm 1.6}$	18.1$_{\pm 1.6}$	14.8$_{\pm 3.4}$	26.4$_{\pm 4.4}$	29.2$_{\pm 2.5}$	40.1$_{\pm 3.5}$	45.1$_{\pm 1.4}$	33.9$_{\pm 2.9}$	68.6$_{\pm 0.3}$	39.7$_{\pm 6.2}$	24.3$_{\pm 3.9}$	39.6$_{\pm 0.6}$	**54.7**$_{\pm 10.2}$	34.5$_{\pm 3.0}$
• ***DSBR*** (α=0.9)	28.2$_{\pm 1.3}$	28.1$_{\pm 0.5}$	27.6$_{\pm 1.2}$	18.4$_{\pm 1.4}$	15.3$_{\pm 3.9}$	26.8$_{\pm 4.1}$	**29.4**$_{\pm 2.4}$	40.4$_{\pm 3.1}$	45.2$_{\pm 1.4}$	35.0$_{\pm 3.1}$	68.4$_{\pm 0.4}$	39.7$_{\pm 6.3}$	24.7$_{\pm 3.8}$	39.6$_{\pm 0.5}$	54.5$_{\pm 10.0}$	34.8$_{\pm 2.9}$
ViT	35.1$_{\pm 0.0}$	32.1$_{\pm 0.0}$	35.8$_{\pm 0.0}$	31.4$_{\pm 0.0}$	25.3$_{\pm 0.0}$	39.5$_{\pm 0.0}$	31.5$_{\pm 0.0}$	24.5$_{\pm 0.0}$	30.1$_{\pm 0.0}$	54.7$_{\pm 0.0}$	64.5$_{\pm 0.0}$	49.0$_{\pm 0.0}$	34.2$_{\pm 0.0}$	53.2$_{\pm 0.0}$	56.5$_{\pm 0.0}$	39.8$_{\pm 0.0}$
• Tent	45.1$_{\pm 2.3}$	38.3$_{\pm 2.2}$	45.1$_{\pm 0.8}$	37.6$_{\pm 2.8}$	33.8$_{\pm 5.0}$	45.6$_{\pm 4.6}$	37.5$_{\pm 4.5}$	28.6$_{\pm 4.7}$	41.2$_{\pm 0.5}$	55.8$_{\pm 1.7}$	67.2$_{\pm 1.6}$	57.2$_{\pm 3.3}$	37.6$_{\pm 3.0}$	56.2$_{\pm 2.4}$	59.1$_{\pm 9.7}$	45.7$_{\pm 3.3}$
• SAR	44.7$_{\pm 1.8}$	38.4$_{\pm 2.1}$	45.0$_{\pm 0.7}$	39.3$_{\pm 2.2}$	33.9$_{\pm 4.2}$	45.2$_{\pm 5.0}$	37.2$_{\pm 4.8}$	27.0$_{\pm 4.3}$	40.5$_{\pm 1.7}$	56.1$_{\pm 0.6}$	66.3$_{\pm 2.5}$	56.2$_{\pm 3.6}$	37.2$_{\pm 2.7}$	55.3$_{\pm 2.3}$	58.6$_{\pm 9.4}$	45.4$_{\pm 3.2}$
• ROID	53.4$_{\pm 1.0}$	**50.1**$_{\pm 1.4}$	53.5$_{\pm 1.3}$	**49.0**$_{\pm 4.5}$	**46.6**$_{\pm 3.7}$	**54.0**$_{\pm 4.0}$	**48.3**$_{\pm 5.6}$	**51.6**$_{\pm 3.1}$	**54.2**$_{\pm 2.7}$	**66.3**$_{\pm 0.6}$	70.4$_{\pm 1.3}$	**64.2**$_{\pm 2.6}$	**54.9**$_{\pm 3.0}$	**61.9**$_{\pm 1.5}$	62.9$_{\pm 8.7}$	**56.1**$_{\pm 3.0}$
• DeYO	50.9$_{\pm 2.1}$	47.4$_{\pm 2.2}$	51.9$_{\pm 1.2}$	46.4$_{\pm 4.8}$	41.9$_{\pm 3.8}$	51.1$_{\pm 6.1}$	42.2$_{\pm 6.5}$	41.3$_{\pm 3.9}$	51.0$_{\pm 0.6}$	63.0$_{\pm 1.4}$	**70.8**$_{\pm 2.0}$	61.6$_{\pm 4.0}$	47.1$_{\pm 2.7}$	59.2$_{\pm 2.5}$	60.8$_{\pm 8.5}$	52.4$_{\pm 3.5}$
• ***DSBR*** (α=0.99)	**53.5**$_{\pm 2.2}$	49.2$_{\pm 1.4}$	**53.7**$_{\pm 1.0}$	46.0$_{\pm 5.2}$	45.0$_{\pm 4.1}$	53.7$_{\pm 2.5}$	48.1$_{\pm 3.6}$	38.6$_{\pm 4.8}$	47.6$_{\pm 2.5}$	64.7$_{\pm 1.5}$	69.5$_{\pm 1.4}$	63.0$_{\pm 2.0}$	52.7$_{\pm 2.0}$	61.4$_{\pm 0.8}$	**62.9**$_{\pm 9.8}$	54.0$_{\pm 3.0}$
• ***DSBR*** (α=0.9)	53.1$_{\pm 1.7}$	48.5$_{\pm 2.0}$	53.4$_{\pm 1.1}$	46.5$_{\pm 3.9}$	45.2$_{\pm 3.8}$	53.8$_{\pm 3.0}$	48.2$_{\pm 3.6}$	39.2$_{\pm 4.6}$	47.2$_{\pm 1.8}$	64.3$_{\pm 1.4}$	69.0$_{\pm 2.1}$	62.7$_{\pm 2.2}$	53.4$_{\pm 4.1}$	60.8$_{\pm 2.3}$	62.8$_{\pm 9.7}$	53.9$_{\pm 3.1}$

Table 12: **Severe class imbalance** on ImageNet-C, following the imbalanced protocol of Weerasinghe et al. [87] for ResNet and ViT backbones.

Table 13: ROC-AUC (%) on the four medical benchmarks, in the same configuration as Tab. 2.*Disabled *consistency* component because it relies on 2D augmentations. †Not available because whole method relies on 2D augmentations.

Model & Method	Histopantum	MammoBench	WILDSCamelyon	GliomaMRI	Δ Average
ResNet	94.1$_{\pm 0.0}$	61.3$_{\pm 0.0}$	97.0$_{\pm 0.0}$	**70.0**$_{\pm 0.0}$	▲ 0.0$_{\pm 0.0}$
• Tent [82]	**94.5**$_{\pm 0.0}$	61.1$_{\pm 0.1}$	93.9$_{\pm 0.0}$	67.8$_{\pm 0.7}$	▼ 1.3$_{\pm 0.2}$
• SAR [45]	94.1$_{\pm 0.0}$	61.3$_{\pm 0.0}$	97.0$_{\pm 0.0}$	67.8$_{\pm 0.7}$	▼ 0.5$_{\pm 0.2}$
• ROID [54]	94.4$_{\pm 0.4}$	60.3$_{\pm 0.9}$	98.6$_{\pm 0.1}$	68.0$_{\pm 0.5}$*	▼ 0.3$_{\pm 0.5}$
• DeYO [46]	89.8$_{\pm 1.8}$	**63.7**$_{\pm 0.1}$	97.1$_{\pm 0.4}$	n/a†	▼ 0.5$_{\pm 0.6}$
• ***DSBR*** (ours)	93.9$_{\pm 0.4}$	62.6$_{\pm 0.4}$	**98.8**$_{\pm 0.1}$	68.1$_{\pm 0.6}$	▲ **0.2**$_{\pm 0.4}$
ViT	94.5$_{\pm 0.0}$	58.9$_{\pm 0.0}$	97.3$_{\pm 0.0}$	57.4$_{\pm 0.0}$	▲ 0.0$_{\pm 0.0}$
• Tent [82]	94.5$_{\pm 0.0}$	58.9$_{\pm 0.0}$	96.1$_{\pm 0.2}$	57.9$_{\pm 0.5}$	▼ 0.2$_{\pm 0.2}$
• SAR [45]	94.5$_{\pm 0.0}$	58.9$_{\pm 0.0}$	97.3$_{\pm 0.0}$	57.4$_{\pm 0.0}$	▲ 0.0$_{\pm 0.0}$
• ROID [54]	**94.6**$_{\pm 0.0}$	59.1$_{\pm 0.3}$	**98.4**$_{\pm 0.0}$	**59.9**$_{\pm 1.1}$*	▲ **1.0**$_{\pm 0.4}$
• DeYO [46]	94.5$_{\pm 0.1}$	59.0$_{\pm 0.1}$	98.3$_{\pm 0.0}$	n/a†	▲ 0.3$_{\pm 0.1}$
• ***DSBR*** (ours)	93.0$_{\pm 0.3}$	**59.9**$_{\pm 0.4}$	97.9$_{\pm 0.2}$	58.0$_{\pm 0.4}$	▲ 0.2$_{\pm 0.3}$

G Entropy filtering as one solution for cluster separation

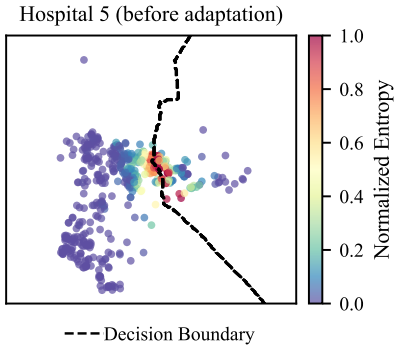

Figure 7: Entropy peaks near the decision boundary with the setup as in Fig. 1, samples are colored by prediction entropy normalized by $\log n$, with $n = 2$.

The cluster-refinement view also explains why entropy filtering, despite our argument in Section E that the gradient-norm motivation for it is false, can still work. Samples close to the decision boundary have high entropy, since the model's confidence falls off with distance from the boundary, as illustrated in Fig. 7. Under cluster merging, the region near the boundary is densely populated with samples that have been incorrectly reassigned across it. Filtering high-entropy samples [46, 58, 59] therefore removes a disproportionate share of these reassigned samples from the gradient update, and additionally cuts an artificial trench through the boundary region that helps the merged clusters separate again. This is, in our view, the actual mechanism by which entropy-based filtering stabilizes Tent, rather than the gradient-norm proxy argument given in Niu et al. [59]. We do not pursue filtering in our work: it is wasteful when data are scarce, and the samples it discards are possibly very informative about how the model is misaligned with the target domain. The next section develops an alternative that addresses prediction bias directly while keeping every sample in the adaptation signal.

H ROC-AUC reveals a structural limit of EM

Table 13 reports ROC-AUC scores for the same configurations. Across the board, the gains in ROC-AUC are negligible: no method, DSBR included, exceeds $+1.0\%$ average improvement on either backbone, and several methods slightly degrade the source model's ROC-AUC despite improving its balanced accuracy substantially. The reason is that by construction, EM sharpens the model's predictions toward the simplex corners. Balanced accuracy depends only on the argmax and benefits from sharper peaks; ROC-AUC depends on the relative ordering of the underlying scores and is degraded when high confidence is assigned indiscriminately, including to misclassified samples. This is a property of the entropy objective, not of any specific algorithm built on it.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the proposed method. The method's main contributions are located in Section 3 and Section 4 which have also been summarized in abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It's included in the conclusion (Sec. 6). Furthermore, we show and discuss scenarios where the default hyperparameter of our method produces limited results. (Sec. 5.2)

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not contain theoretical results/proofs. The mathematical notations however, are defined in Sec. 2, and the derivation of our algorithm is described in Sec. 4.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We explain implementation details in Sec. 5 with additional information in Appendix A and B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have provided an anonymized link to our code in the abstract, which includes clear instructions for reproducing our results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Included in implementation details in Sec. 5 and further details in Appendix A and B. Especially, in Appendix A we give a thorough explanation of the full pre-training and adaptation pipeline. Furthermore, in Sec. 5.2 we give an explanation on why we report results for an additional value of the EMA hyperparameter. Finally, we provide all remaining details with our code base.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiment tables include 2 standard deviations for all numbers (see Tab. 2, Tab. 3). Furthermore, Fig. 5 contains error bars of 2 standard deviations as well. Section A provides these details.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D details all required and used compute resources as well as time of execution for our main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We comply.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: This work has a minor positive societal impact because it improves the stability and prevents model collapse during test-time adaptation under distribution shift without labels, and states it in the Section 1: Introduction section

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our work doesn't involve these.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Sec. 5 properly cites the creators of all datasets, models and adaptation methods. Sec. B provides further details including the licenses and terms of use, all of which we respect.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable to us.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable to us.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Not used for core method development.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.