A REASSESSMENT OF GLOBAL WEIGHT-SPACE SAMPLING VIA DETERMINISTIC FLAT MINIMA

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Ensembles remain the empirical upper bound for Uncertainty Quantification, successfully capturing global epistemic diversity across disjoint loss valleys. To circumvent their prohibitive computational overhead, modern frameworks frequently rely on local weight-space sampling (e.g., SWAG, Laplace Approximations) to approximate an ensemble within a single basin. In this paper, we analyze the scaling issues that arise when global weight-space sampling is applied to overparameterized architectures, formalizing this as the weight-space variance dilemma. We highlight a geometric tension: strictly bounding stochastic weight perturbations within a certified optimal-transport safety radius leads to dimensionality dilution, where noise merges with the high-dimensional null space, leaving predictions functionally identical to the base checkpoint. Consequently, practical single-basin samplers are forced to inject unconstrained variance to achieve ensemble diversity. Such macroscopic noise can disrupt internal feature manifolds, necessitating computationally expensive deployment heuristics such as data-dependent batch normalization recalibration. Based on recent theoretical insights that flat minima can implicitly regularize predictive entropy, extensive evaluations on ResNet-50, ViT-B/16, and YOLOv8x demonstrate that much of the out-of-distribution logit smoothing attributed to global sampling can be natively embedded in the deterministic base optimizer. While unable to model true multi-modal posteriors, a single deterministic Sharpness-Aware Minimization (SAM) checkpoint serves as a highly efficient, zero-overhead baseline that frequently matches the local calibration of global ensembles. Furthermore, we show that while uniform Euclidean bounds stabilize homogeneous classifiers, complex dense prediction tasks require scale-aware Mahalanobis-Wasserstein spaces (ASAM) to prevent structural regression collapse.

1 INTRODUCTION

Deep Ensembles (Lakshminarayanan et al., 2017) are established as a standard benchmark for Uncertainty Quantification (UQ). By independently training multiple models from distinct initializations, they can successfully navigate the non-convex topography of deep networks to capture true global epistemic diversity. However, their computational and memory demands limit their practical use in resource-constrained environments (Wen et al., 2020). To bridge this efficiency gap, the field has turned to local weight-space sampling methods, including SWAG (Maddox et al., 2019), Variational Inference (VI) (Hoffman et al., 2013), and Laplace approximations (Deng et al., 2022). These methods aim to approximate an ensemble efficiently by stochastically mapping the local geometry of a single loss basin.

Despite their widespread adoption, scaling global single-basin sampling to modern, overparameterized architectures introduces theoretical and geometric complexities. To derive a mathematically safe variance budget for stochastic inference, perturbations can be constrained within an optimal-transport bound, such as the robust Wasserstein radius (ρ) certified by Sharpness-Aware Minimization (SAM) (Foret et al., 2021; Nguyen et al., 2023). However, applying global geometric bounds in high-dimensional spaces reveals a fundamental limitation, which we formalize as the weight-space variance dilemma:

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

- **Dimensionality Dilution.** If a finite macroscopic radius ρ is distributed isotropically across a global dimension (e.g., of $D \approx 25 \times 10^6$ parameters), theoretically safe noise is completely absorbed by the network's functional null-space. To remain mathematically safe, sampling passes become functional clones of the deterministic checkpoint.
- **Functional Disruption.** To bypass this dilution and extract measurable epistemic diversity, practical samplers must inject unconstrained macroscopic variance. However, abandoning safety bounds drives parameters into steep regions, intensifying likelihood misspecification and causing activation shifts that mandate computationally heavy batch normalization (BN) recalibration prior to inference.

This dilemma highlights an assumption of current global single-basin methods; that unconstrained sampling on a converged flat minimum in overparameterized systems generates reliable epistemic diversity. Recent theoretical work has formally proven that deterministic SAM inherently regularizes the negative entropy of the predictive distribution, systematically dampening overconfidence, and yielding calibrated models (Tan et al., 2026). While a deterministic point-estimate inherently abandons the mapping of true multi-modal Bayesian posteriors, flat minima are highly regularized. If the deterministic minimum is theoretically calibrated natively, it can serve as a zero-overhead baseline that forces us to question when the computational burden of single-basin sampling is truly justified.

The Bayesian deep learning community has addressed high-dimensionality issues through targeted subspace methods, such as subnetwork inference and last-layer methods, which bypass dimensionality dilution by reducing the parameter space (Izmailov et al., 2020). Therefore, our reassessment is strictly isolated to the systematic limitations of global, unconstrained sampling.

Through extensive evaluations on ImageNet-C (Hendrycks & Dietterich, 2019), ImageNet-O (Hendrycks et al., 2021), and COCO-O (Mao et al., 2023) domain shifts, we demonstrate that the flat topography created by a Sharpness-Aware Minimization (SAM) optimizer natively yields soft, high-entropy predictive distributions on anomalous data, inherently resisting overconfidence without the need for stochastic passes. The main contributions of this paper are as follows: **(1)** We formalize the weight-space variance dilemma, demonstrating that globally bounding variance in overparameterized networks yields functionally invisible noise (*Dimensionality Dilution*), while unconstrained macroscopic variance shatters internal manifolds (*Functional Disruption*). **(2)** We establish that a single deterministic forward pass of a SAM/ASAM-optimized network effectively matches the Out-of-Distribution (OOD) robustness typically attributed to local sampling, allowing for zero-recalibration deployment ($M = 0$). **(3)** Building on the optimization findings of Kwon et al. (2021), we provide empirical evidence that Euclidean bounds (SAM) structurally collapse dense regression heads under OOD shifts. We prove that scale-aware Mahalanobis-Wasserstein spaces (ASAM) are a required geometric prerequisite for single-basin UQ stability in heterogeneous models.

2 Related Work

Single-Basin Samplers and High-Dimensional Geometry. To avoid the computational burden of Deep Ensembles, methods like Subspace Inference (Izmailov et al., 2020), SWAG (Maddox et al., 2019), and Laplace Approximations (Deng et al., 2022) seek to characterize the local geometry of a single loss basin. Global Laplace methods construct a Gaussian posterior centered at the mode, relying on the local Hessian to dictate variance. However, estimating this curvature in modern overparameterized models can collapse under dimensionality constraints. Approximating the Hessian requires massive structural concessions (e.g., diagonal or Kronecker-factored matrices (Ritter et al., 2018)), which severely degrade fidelity. To counteract high dimensionality, recent literature developed Subnetwork Inference (Daxberger et al., 2021) and Last-Layer Laplace (Kristiadi et al., 2020). By restricting the probabilistic formulation to a small subset of weights ($S \ll D$) while freezing the remaining parameters, these methods yield expressive, tractable posteriors. Such localized approaches elegantly bypass the dimensionality scaling limits analyzed in this work; thus, our work is strictly focused on global, full-parameter sampling architectures.

Optimal Transport and Wasserstein Safety Boundaries. The mathematical connection between optimal transport and Bayesian inference provides a strong framework to define geometric safety.

Blanchet et al. (2019) and Nguyen et al. (2023) established Wasserstein Distributionally Robust Optimization (W-DRO), proving that minimax robustness over a Wasserstein ball is equivalent to penalized empirical risk minimization. Crucially, Möllenhoff & Khan (2023) demonstrated that SAM acts precisely as this optimal convex relaxation. Rather than utilizing this Kantorovich duality to iteratively sample an approximate posterior (a process hindered by unstable Sinkhorn iterations (Ambrogioni et al., 2018) or prohibitive Hessian oracles (Lambert et al., 2022)) we repurpose it. We utilize the optimal transport radius ρ strictly as a heuristic safety boundary that conservatively dictates how much variance any global sampling method can inject before severe generalization failure occurs.

The Implicit Calibration of Flat Minima. While traditional UQ methodologies attempt to manually inject predictive entropy via parameter perturbations (Gal & Ghahramani, 2016), a recent work by Tan et al. (2026) has demonstrated that SAM natively achieves this objective. Standard optimization frameworks, such as AdamW and SGD, exhibit a strong propensity for predictive overconfidence on OOD data. However, Tan et al. formally prove that the explicit gradient-norm penalization within SAM acts as an implicit regularizer on the negative entropy of the predictive distribution, operating similarly to Focal Loss but maintaining higher accuracy on corrupted data. By suppressing overconfident logits natively, a deterministic SAM can yield highly calibrated models without requiring downstream stochastic integration. This discovery requires a practical reassessment: if a deterministic single-basin optimizer is natively calibrated, imposing secondary global stochastic sampling on top of it introduces significant computational overhead that must be carefully justified against its actual epistemic yield.

The Cold Posterior Effect and Likelihood Misspecification. Injecting unconstrained variance into deep networks frequently causes structural degradation, a phenomenon historically linked to the *Cold Posterior Effect* (CPE) (Wenzel et al., 2020). Initially viewed as a theoretical discrepancy where shattered internal representations demanded heuristic temperature tuning, recent work by Zhang et al. (2024) mathematically proved that CPE is primarily an indicator of underfitting and likelihood misspecification. As noted by Tran et al. (2022) and Rudner et al. (2023), mapping uninformative weight-space noise to functional behavior in highly non-linear architectures is inherently chaotic. Unconstrained global sampling can drive the model into suboptimal regions, which directly exacerbates such misspecification. The resulting activation shift mandates computationally expensive data-dependent BN recalibrations (Ioffe & Szegedy, 2015). Our work maps this phenomenon as the second constraint in the variance dilemma, asserting that safely navigating single basins requires abandoning unconstrained global sampling in favor of determinism.

3 The Weight-Space Variance Dilemma

To understand why global, full-parameter weight-space sampling hits theoretical scaling limits in modern architectures, we must formalize the geometric constraints of overparameterized spaces. We highlight a structural tension in global stochastic sampling: to theoretically ensure the preservation of learned features, perturbations would need to remain within a mathematically certified safety boundary. However, mapping this theoretical safety bound onto high-dimensional parameter spaces exposes a fundamental geometric contradiction.

3.1 Geometric Boundaries via Heuristic Optimal Transport

While true bayesian model averaging marginalizes probabilities and does not strictly adhere to adversarial minimax bounds, W-DRO provides a useful heuristic safety boundary. W-DRO minimizes the expected worst-case loss for any approximate posterior bounded within a 1-Wasserstein ball of radius ρ around θ^*. Applying Kantorovich duality and a first-order Taylor approximation collapses this exactly into the empirical SAM optimizer step (we provide the complete mathematical reduction in Appendix A.2).

Heuristically, for global sampling, if a perturbation is to remain within the certified flat and safe space created during training, the 2-Wasserstein distance of the continuous probability distribution to the Dirac delta must not exceed ρ. This distance equates to the trace of the covariance matrix $\mathbf{\Sigma}$:

$$Tr(\mathbf{\Sigma}) \leq \rho^2 \tag{1}$$

This geometric bound defines the maximum conservative variance budget available to a global sampler.

3.2 HORN 1: DIMENSIONALITY DILUTION

Applying this conservative budget to modern neural networks reveals *Dimensionality Dilution*. For a global, isotropic predictive distribution $Q = \mathcal{N}(\theta^*, \beta\mathbf{I})$ of dimension D, the trace evaluates to $Tr(\beta\mathbf{I}) = \beta D$. Adhering to the geometric bound dictates:

$$\beta \leq \rho^2/D \tag{2}$$

In modern architectures (e.g., $D \approx 25 \times 10^6$ for ResNet-50), the maximum permissible per-parameter variance shrinks to functionally irrelevant magnitudes ($\beta \approx 10^{-8}$). Such minimal noise would be almost entirely absorbed by the network's vast functional null-space. Model outputs thus shift only marginally, rendering the sampling passes functional clones of the deterministic checkpoint. This demonstrates that theoretically safe, bounded sampling is functionally invisible, showing why practical UQ methods cannot rely on mathematically certified geometric bounds.

3.3 HORN 2: FUNCTIONAL DISRUPTION

As theoretically safe variance yields no functional diversity, practical single-basin samplers (such as SWAG or Laplace) are forced to violate this geometric boundary. To escape dimensionality dilution and extract measurable epistemic diversity, they must inject unconstrained macroscopic variance. Furthermore, advanced methods abandon isotropy; Laplace approximations utilize the inverse Hessian ($\mathbf{\Sigma} \approx H^{-1}$), structurally maximizing variance in flat, low-curvature directions while minimizing it in steep, high-curvature directions ($1/\lambda_{max} \ll 1/\lambda_{min}$).

However, this mathematically correct behavior creates a practical contradiction in global parameter spaces. As established by analyses of the Hessian eigenspectrum (Sagun et al., 2017; Papyan, 2020), overparameterized networks possess a vast, rank-deficient null-space. The flat directions where the Laplace approximation safely places its variance align almost entirely with this null-space. Consequently, moving the weights along these flat axes produces no measurable change in the network's actual predictions (Singh et al., 2021). To generate meaningful functional ensemble diversity, the sampler must inject variance into the highly curved directions that actually govern functional behavior.

Even advanced structural approximations like Kronecker-factored Approximate Curvature (KFAC) force structural concessions across layers that inevitably misalign with the true flat manifolds of largely heterogeneous architectures (Ritter et al., 2018). Furthermore, as demonstrated by Antoran et al. (2022), linearized Laplace methods collide with BN layers unless extensive reparameterizations are applied. This demonstrates the practical consequence of breaking the safety boundary; attempting to force measurable global variance post-hoc drives the model into unoptimized loss walls, intensifying likelihood misspecification (triggering the CPE) and disrupting internal feature representations. Consequently, activation shift occurs, forcing practitioners to run computationally heavy BN recalibration prior to inference.

3.4 DETERMINISTIC BASELINE AND SCALE-AWARENESS

With this weight-space variance dilemma we show the hurdles of unconstrained, post-hoc global sampling. We acknowledge that variance can be actively learned if it is integrated directly into the optimization trajectory as a dedicated parameter (e.g., Variational-SAM or OT-MDR (Möllenhoff & Khan, 2023; Nguyen et al., 2023)). However, for post-hoc samplers attempting to extract epistemic diversity from a converged deterministic model, this geometric tension dictates that safe boundaries dilute predictions, while practical unconstrained variance disrupts them.

As the flat minima excavated by SAM inherently act as entropy regularizers that suppress overconfidence (Tan et al., 2026), the deterministic checkpoint (θ^*) itself serves as a local UQ estimator. By removing post-hoc sampling entirely, native feature distributions remain intact, enabling instant inference without recalibration ($M = 0$). Deploying deterministic flat minima, however, requires topological awareness:

- **Homogeneous Topographies.** Image classifiers (ResNet-50, ViT-B/16) possess uniform structural scaling across layers. Consequently, such architectures can be stabilized by standard SAM, which enforces a uniform Euclidean space ($T = I$).
- **Heterogeneous Topographies.** Dense multi-objective models like YOLOv8x pair large backbone extractors with shallow Feature Pyramid Networks (FPN). Enforcing a uniform Euclidean variance budget across such multi-scale layers over-perturbs the small-magnitude weights within the detection heads, collapsing bounding-box regression. Here, we mathematically require a transition to Mahalanobis-Wasserstein spaces via adaptive SAM (ASAM) (Kwon et al., 2021). By integrating a scale-aware operator ($T \propto |\theta|$) directly into the sharpness calculation, ASAM ensures that local geometry is constrained proportionately to the parameter magnitude. This strictly preserves structural integrity for dense prediction tasks without resorting to multi-model ensembles.

4 Experiments

We evaluate deterministic flat minima across image classification and dense object detection to empirically evaluate the efficiency-calibration tradeoffs. We investigate whether a single deterministic SAM/ASAM checkpoint can serve as a competitive, zero-overhead baseline capable of matching the OOD logit smoothing typically attributed to stochastic sampling.

4.1 Experimental Setup & Implementation Details

Architectures and Datasets. We benchmark across three architectures: ResNet-50 (He et al., 2016), ViT-B/16 (Dosovitskiy et al., 2021) (classification), and YOLOv8x (Redmon et al., 2016) (object detection). Classification is evaluated on ImageNet-1k (Russakovsky et al., 2015) (in-distribution (ID)), ImageNet-O (Hendrycks et al., 2021) (anomalies), and ImageNet-C (Hendrycks & Dietterich, 2019) (15 corruptions $\times$ 5 severities). Dense object detection is evaluated on COCO (Lin et al., 2014) (ID) and the COCO-O domain-shift benchmark (Mao et al., 2023). In total, this forms an evaluation matrix of over 80 distinct benchmark configurations.

Baselines and Metrics. We compare against established methods (Deep Ensembles (Lakshminarayanan et al., 2017), SWAG (Maddox et al., 2019)) as well as stochastic activation sampling techniques, including Monte Carlo Dropout (MCD) (Gal & Ghahramani, 2016), DropBlock (MCDB) (Deepshikha et al., 2021), and Stochastic Depth (MCSD) (Müller et al., 2026). Performance is quantified via Top-1 Accuracy and mAP50-95. Uncertainty calibration and anomaly rejection are measured using Expected Calibration Error (ECE) (Guo et al., 2017), Negative Log-Likelihood (NLL), Predictive Entropy, and the Area Under the Accuracy-Rejection Curve (AUARC) (Geifman et al., 2019).

Implementation and Geometric Bounds. The deterministic baseline introduces no learnable parameters; robustness is governed entirely by the Wasserstein radius ρ utilized by the SAM/ASAM optimizer during training. As our deterministic approach natively preserves activation map stability, it completely eliminates the data-dependent BN recalibration bottleneck required by baselines like SWAG ($M = 0$). Exact training budgets and ρ configurations are detailed in Appendix A.1.

4.2 Results

4.2.1 In-Distribution Predictive Performance and Calibration

Image Classification. On clean ImageNet-1k classification, standard Cross-Entropy loss landscapes can contain numerous disjoint, valid local minima. Consequently, Deep Ensembles ($N = 5$) set the upper bound for predictive diversity, achieving the highest Top-1 accuracy (80.26%) on ResNet-50 (see Table 1). However, the deterministic SAM model functions as an exceptionally efficient 1x-parameter alternative. It achieves 79.85% (ResNet-50) and 80.37% (ViT-B/16) Top-1 accuracy, effectively matching the ensemble's performance without the prohibitive memory overhead. Furthermore, aligning with recent theoretical findings by Tan et al. (2026), deterministic SAM natively suppresses overconfident logits, yielding a best-in-class ECE of 0.015 on ResNet-50, effectively halving the calibration error of the multi-modal Deep Ensemble (0.028).

Table 1: In-distribution predictive performance and calibration (on ImageNet-1k (classification) and COCO (object detection)). Deep Ensembles utilize $N = 5$, SWAG utilizes $K = 20$. Acc refers to Top-1 accuracy, while mAP refers to mAP50-95.

	ResNet-50			**ViT-B/16**			**YOLOv8x**		
Method	Acc ↑	ECE ↓	AUARC ↑	Acc ↑	ECE ↓	AUARC ↑	mAP ↑	ECE ↓	AUARC ↑
MCD	0.7896	0.016	0.939	0.7908	0.058	0.945	0.5368	0.076	0.909
MCDB	0.7924	0.016	0.940	0.7799	0.096	0.940	0.5183	0.075	0.901
MCSD	0.7887	0.031	0.933	0.7926	**0.021**	0.937	0.5300	0.067	0.911
SWAG	0.7985	0.040	0.944	0.8014	0.070	0.948	0.5276	0.209	0.760
Deep Ensemble	**0.8026**	0.028	**0.948**	**0.8065**	0.049	**0.949**	0.5400	**0.037**	0.915
Det. SAM	0.7985	**0.015**	0.945	0.8037	0.031	0.942	0.2998	0.194	**0.946**
Det. ASAM	0.7934	0.032	0.944	**0.8065**	0.046	**0.949**	**0.5433**	0.070	0.919

Table 2: Out-of-distribution evaluations on ImageNet-C (Severity 5) and ImageNet-O anomaly detection. Acc refers to Top-1 accuracy.

	ImageNet-C (ResNet) Gaussian Noise Sev. 5		**ImageNet-C (ViT)** Zoom Blur Sev. 5		**ImageNet-O (ResNet)** Anomaly Detection
Method	Acc (↑)	ECE (↓)	Acc (↑)	ECE (↓)	Entropy (↑)
MCD	0.086	0.076	0.291	0.156	**1.368**
SWAG	0.086	0.113	0.324	0.172	1.035
Deep Ensemble	0.105	0.143	0.321	0.130	1.037
Det. SAM	0.094	**0.055**	0.321	**0.028**	1.179
Det. ASAM	**0.128**	0.060	**0.325**	0.119	1.055

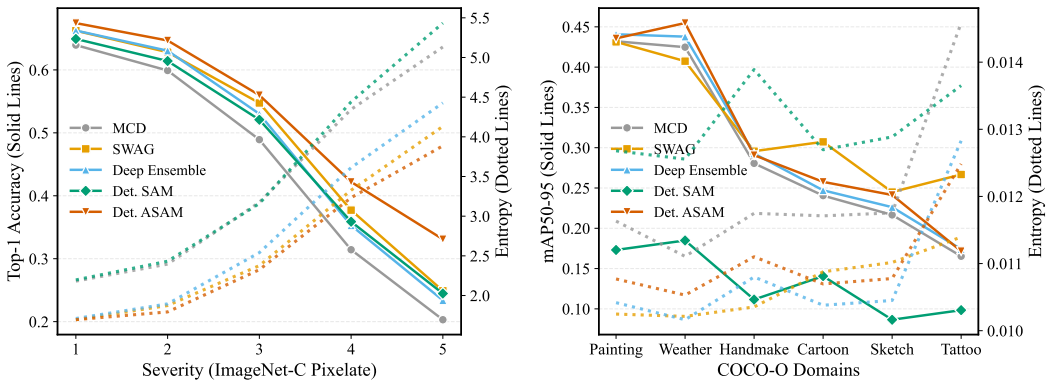
Dense Object Detection. The empirical narrative inverts on the more complex multi-objective landscape of YOLOv8x. In dense object detection, mode-hopping across disjoint valleys (the core mechanic of an ensemble) can degrade spatial coordinate consistency, such as bounding-box alignment (also see Appendix A.4.3). By explicitly forcing the network to optimize a single wide minimum, deterministic ASAM natively stabilizes these regression heads. On the COCO benchmark, deterministic ASAM outperforms the 5-model Deep Ensemble across all primary metrics, achieving a superior mAP50-95 of 0.5433 (vs. 0.5400) and an AUARC of 0.919 (vs. 0.915).

Crucially, standard Euclidean SAM experiences a severe collapse on YOLOv8x, with accuracy falling to 0.2998 mAP50-95. Kwon et al. (2021) originally identified that uniform Euclidean bounds collapse on multi-scale architectures during training due to scale variance. Our results reveal that the proportional scaling operator of Mahalanobis-Wasserstein spaces (ASAM) is not merely an optimization convenience, but a strict geometric prerequisite for single-basin UQ stability in dense prediction tasks.

4.2.2 Out-of-Distribution Robustness

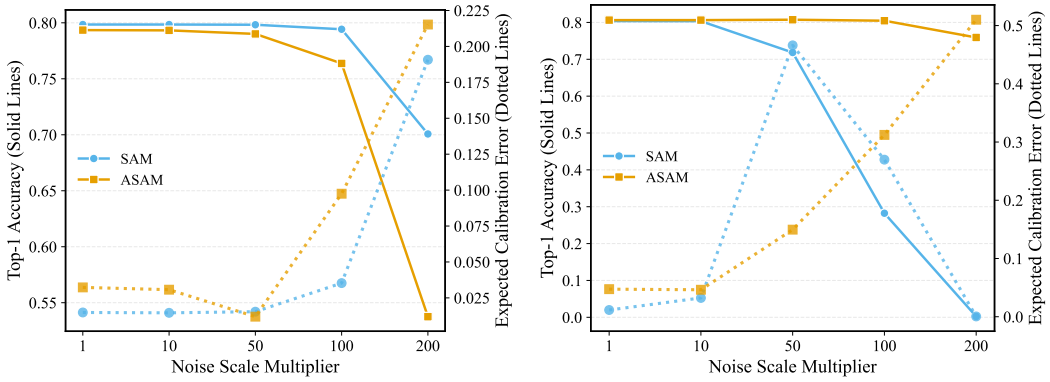
High-Frequency Noise Immunity. On ImageNet-C, deterministic flat minima demonstrate a significant resilience to high-frequency corruptions. As degradation severity scales to level 5 (Table 2), the scale-aware ASAM formulation outperforms the Deep Ensemble baseline (0.128 vs. 0.105 accuracy on Gaussian Noise). This degradation is visually confirmed in Figure 1a, where deterministic (Det.) ASAM maintains the highest Top-1 accuracy across all severity levels of the Pixelate corruption. Concurrently, SAM natively yields the steepest, most monotonic entropy scaling. This shows, that the deterministic flat basin inherently regularizes the negative entropy of the predictive distribution under distribution shift. This prevents the network from confidently committing to sharp, spurious OOD artifacts without requiring any explicit post-hoc sampling. (*Complete evaluation tables for all 15 corruptions are provided in Appendix A.4.1*).

Anomaly Detection and Entropy Scaling. When confronted with completely anomalous data (ImageNet-O), uncertainty scaling mechanisms diverge. Stochastic activation sampling (e.g., MCD) produces the highest predictive entropy (1.368), a pronounced logit spread that naturally arises from the stochastic masking of intermediate features during the forward pass. On the other hand, de-



(a) Classification on ResNet-50 (b) Object Detection on YOLOv8x

Figure 1: Out of Distribution Evaluation. Showing (a) ImageNet-C *Pixelate* corruption with varying severity and (b) COCO-O domain shift examples.



(a) Scale Sweep on ResNet-50 (b) Scale Sweep on ViT-B/16

Figure 2: Inference noise scale sweeps at increasing magnitude multipliers.

terministic SAM does not alter the base activation graphs. However, by utilizing the implicitly regularized geometry of the flat optimum, it still yields competitive entropy scaling (1.179 vs. 1.037 for Deep Ensembles). This demonstrates that meaningful OOD anomaly rejection can be achieved natively by the base optimizer, without necessarily requiring stochastic interventions.

Domain Shift and Regression Stability. This geometric resilience extends to the object detection task under visual domain shifts (COCO-O). As illustrated in Figure 1b, standard SAM again shows a collapse in accuracy, with mAP50-95 falling below 0.20 across all evaluated domains. The scale-aware ASAM successfully stabilizes the detection head, matching or exceeding the Deep Ensemble's mAP50-95 on shifts like Painting, Weather, and Handmake. Importantly, performance under domain shift remains nuanced: global covariance sampling (SWAG) explicitly yields superior mAP on highly stylized domains such as Cartoon and Tattoo, while stochastic activation sampling (MCD) again generates the most responsive entropy spikes. This confirms that while ASAM provides a latency-free calibration baseline, computationally heavy samplers retain specific, localized advantages.

4.2.3 Empirical Evidence of Dimensionality Dilution

To empirically validate Horn 1 of the dilemma (Dimensionality Dilution), we subjected our optimized SAM/ASAM checkpoints to an artificial inference-scale sweep. We synthetically injected isotropic stochastic noise at escalating magnitude multipliers (k). At conservative, geometrically safe bounds (e.g., up to $50\times$ for ResNet-50, and $10\times$ for ViT-B/16), the stochastic noise completely

Table 3: Inference saturation across sample sizes (N) for stochastically sampled SAM/ASAM inference (ViT-B/16 on ImageNet-1k).

Samples (N)	**1**	**3**	**5**	**10**	**20**	**30**	**50**
Top-1 Acc (↑)	0.8058	0.8063	0.8063	0.8063	**0.8065**	0.8063	0.8064
ECE (↓)	0.0474	0.0470	0.0471	0.0470	**0.0467**	0.0470	0.0469
Am. Throughput (ms/img)	**1.26**	3.10	4.98	9.70	19.08	28.48	47.29

dilutes into the high-dimensional null-space. The predictive accuracy and ECE remain virtually flat, confirming that safely bounded single-basin sampling in modern architectures is mathematically invisible, rendering the computation functionally redundant.

However, when forced to uncertified extremes to induce measurable logit variance, the models predictably degrade (Horn 2: Functional Disruption). As illustrated in Figure 2 (and in Appendix A.5), while the exact phase transition depends on architectural topography (e.g., SAM on ViT-B/16 collapsing at $50\times$, and highly parametrized networks like YOLOv8x at $> 500\times$) all models ultimately suffer predictive and calibration degradation when unconstrained variance is forced.

As mathematically safe noise is functionally invisible (dilution) and macroscopic noise triggers structural collapse (disruption), the precise magnitude of global noise required to generate meaningful epistemic diversity simultaneously destroys the network's learned feature representations.

Synthetic W-DRO Boundary Ablation. To empirically isolate and prove Horn 1 (Dimensionality Dilution), we conducted a synthetic ablation. We instantiated a stochastic surrogate that is strictly constrained to the mathematically certified optimal-transport radius ρ (see Appendix A.3.1). This represents a theoretical boundary that preserves internal feature manifolds.

The empirical evaluation confirms our hypothesis (see Table 3), as when bounded safely to ρ, sweeping the sample size from $N = 1$ to $N = 50$ on ViT-B/16 yields virtually no change in accuracy or ECE. The noise is entirely absorbed by the null-space. As theoretically safe noise is functionally invisible, practical samplers are mathematically forced to inject macroscopic variance (thus leaving safe geometric bounds, see Figure 2), which in turn triggers Horn 2 (Functional Disruption).

4.2.4 Zero Recalibration Deployment

As practical weight-sampling techniques (like SWAG) inject unconstrained weight variance to overcome dimensionality dilution, they invariably degrade internal activation manifolds (Horn 2). This triggers an activation shift bottleneck, mandating a computationally expensive BN recalibration pass over $M > 0$ images before evaluation. By relying exclusively on the centered deterministic optimum, SAM/ASAM bypasses this disruption, enabling zero-overhead deployment ($M = 0$). We further show SAM/ASAM maintains its peak performance natively at $M = 0$ via sampling based evaluations (see Section 4.2.3, as well as Appendix A.3.1 for the full analysis on stochastic sampling).

5 Limitations

While deterministic single-basin optimization provides an elegant, highly efficient bypass to the computational overhead of local weight sampling, it introduces several critical theoretical and operational constraints:

Dimensionality Explosion and the Laplace Approximation Limit. A core premise of this work is questioning the default reliance on post-hoc global weight-space sampling, a framework historically grounded in the Laplace approximation. The fundamental mathematical barrier driving this skepticism is that, as the parameter dimension D scales up, the volume of a high-dimensional Gaussian distribution becomes sensitive to local curvature. In such extensive spaces, minimal inaccuracies in curvature estimation can compound exponentially, causing the Laplace approximation to break down and yield degenerate probability densities. This structural fragility reinforces our conclusion that modeling full-parameter probabilistic distributions via single-basin matrices is mathematically unreliable, necessitating the shift to deterministic bounds.

Probabilistic Void and Hallucinated Minimizers. Our deterministic approach formally abandons the pursuit of a true Bayesian posterior. Recent theoretical analyses demonstrate that SAM algorithms frequently fail to converge to strict stationary minima, instead settling in *hallucinated minimizers*, flat regions that act as attractors but possess non-zero gradients (Andriushchenko & Flammarion, 2022; Park et al., 2026). Because the basin lacks strict geometric stationarity, treating it as the anchor of a valid probabilistic surrogate is mathematically unsound, restricting its interpretation purely to empirical robust optimization.

Reparameterization Invariance and Geometric Mismatch. The theoretical Kantorovich duality justifying SAM's optimal transport bound relies on a Euclidean norm-ball. However, Euclidean geometries in weight space are notoriously not invariant to parameterization transformations; a fundamentally equivalent neural network with rescaled layers possesses a drastically different Euclidean Hessian. While ASAM's Mahalanobis correction mitigates scaling mismatches across heterogeneous layers, neither method perfectly maps onto the true, reparameterization-invariant Fisher information manifold necessary for optimal statistical boundaries, as neural network parameter spaces are fundamentally non-Euclidean and Fisher-dominated (Kim et al., 2022).

Global Diversity vs. Local Calibration. By relying exclusively on a single wide, flat minimum, deterministic optimization inherently prevents mode-hopping. This optimally stabilizes multi-objective regression tasks (e.g., YOLOv8x) and provides exceptional local calibration by implicitly suppressing overconfidence. However, it also abandons the ability to model true global epistemic uncertainty. For capturing the distinct, multi-modal disjoint valleys associated with complex, unconstrained hypothesis spaces, multi-run Deep Ensembles remain indispensable.

Spatial Stability Tradeoff. While deterministic flat minima optimally stabilize classification confidence under domain shift, our evaluations reveal a strict limitation for all weight-space uncertainty estimators regarding absolute spatial consistency. As detailed in Appendix A.4.3, global weight perturbations inherently induce higher bounding-box jitter in sensitive regression heads. As stochastic activation sampling (e.g., MCD, MCSD) solely masks intermediate features and leaves the deterministic regression weights physically intact, it remains the superior paradigm when pure spatial coordinate consistency is the primary deployment constraint.

6 Conclusion

In this work, we empirically evaluate the efficiency-calibration tradeoffs of global single-basin sampling versus deterministic flat minima. We formalize the weight-space variance dilemma as a geometric tension: mathematically safe, bounded sampling dilutes entirely into the high-dimensional null-space. Consequently, practical samplers are forced to inject unconstrained macroscopic variance, which drives weights into highly curved axes, increasing likelihood misspecification and necessitating computationally heavy heuristics like BN recalibration.

While targeted Subnetwork Inference elegantly bypasses such algebraic bottlenecks, we demonstrate that much of the predictive calibration and OOD logit smoothing often sought via complex global sampling are intrinsic properties of flat minima themselves. A single deterministic SAM checkpoint provides a highly competitive, zero-overhead baseline that frequently matches the local calibration of global ensembles.

Crucially, topological awareness is mandatory, as while Euclidean SAM collapses on multi-scale detection tasks, Mahalanobis-Wasserstein spaces (ASAM) natively stabilize these regression heads. We emphasize that sampling is not obsolete; global samplers (SWAG) retain dominance on highly stylized domain shifts, and activation sampling (MCD) remains superior for strict spatial coordinate stability. However, when instant, zero-recalibration deployment is the priority, optimally centered, entropy-regularized deterministic point-estimates provide a highly efficient and formidable baseline.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual frameworks, refine hypotheses, and provide feedback on research methodology. We have not used generative AI tools for generating datasets, formulating proofs, analyzing data, interpreting results, or other required disclosure tasks. Additionally, we used generative AI tools to edit the manuscript to improve readability and to edit/format software code for our supplementary repository. We have reviewed all AI-assisted work: theoretical refinements were manually verified by the authors, and all minimized code examples were tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our experiments, we provide the exemplary source code for our SAM and ASAM implementations, as well as the diagnostic stochastic surrogate, via an anonymous GitHub repository. A comprehensive reproducibility statement, which includes details on our optimization and fine-tuning budgets, along with the specific Wasserstein radius configurations used for each architecture, is located in Appendix A.1.

REFERENCES

Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Marcel A. J. van Gerven, and Eric Maris. Wasserstein Variational Inference. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/2c89109d42178de8a367c0228f169bf8-Abstract.html.

Maksym Andriushchenko and Nicolas Flammarion. Towards Understanding Sharpness-Aware Minimization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 639–668. PMLR, June 2022. URL https://proceedings.mlr.press/v162/andriushchenko22a.html.

Javier Antoran, David Janz, James U. Allingham, Erik Daxberger, Riccardo Rb Barbano, Eric Nalisnick, and Jose Miguel Hernandez-Lobato. Adapting the Linearised Laplace Model Evidence for Modern Deep Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 796–821. PMLR, June 2022. URL https://proceedings.mlr.press/v162/antoran22a.html.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, September 2019. ISSN 0021-9002, 1475-6072. doi: 10.1017/jpr.2019.49. URL https://www.cambridge.org/core/journals/journal-of-applied-probability/article/robust-wasserstein-profile-inference-and-applications-to-machine-learning/4024D05DE4681E67334E45D039295527.

Erik Daxberger, Eric Nalisnick, James U. Allingham, Javier Antoran, and Jose Miguel Hernandez-Lobato. Bayesian Deep Learning via Subnetwork Inference. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2510–2521. PMLR, July 2021. URL https://proceedings.mlr.press/v139/daxberger21a.html.

Kumari Deepshikha, Sai Harsha Yelleni, P. K. Srijith, and C. Krishna Mohan. Monte Carlo DropBlock for Modelling Uncertainty in Object Detection, August 2021. URL http://arxiv.org/abs/2108.03614. arXiv:2108.03614 [cs.CV].

Zhijie Deng, Feng Zhou, and Jun Zhu. Accelerated Linearized Laplace Approximation for Bayesian Deep Learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2695–2708. Curran Associates, Inc., 2022. doi: 10.52202/068431-0195. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/12143893d9d37c3569dda800b95cabd9-Abstract-Conference.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs.CV].

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-Aware Minimization for Efficiently Improving Generalization, April 2021. URL http://arxiv.org/abs/2010.01412. arXiv:2010.01412 [cs.LG].

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1050–1059. PMLR, June 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers, April 2019. URL http://arxiv.org/abs/1805.08206. arXiv:1805.08206 [cs.LG].

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330. PMLR, July 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. pp. 770–778, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, March 2019. URL http://arxiv.org/abs/1903.12261. arXiv:1903.12261 [cs.LG].

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. pp. 15262–15271, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456. PMLR, June 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

Pavel Izmailov, Wesley J. Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace Inference for Bayesian Deep Learning. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pp. 1169–1179. PMLR, August 2020. URL https://proceedings.mlr.press/v115/izmailov20a.html.

Minyoung Kim, Da Li, Shell X. Hu, and Timothy Hospedales. Fisher SAM: Information Geometry and Sharpness Aware Minimisation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 11148–11161. PMLR, June 2022. URL https://proceedings.mlr.press/v162/kim22f.html.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5436–5446. PMLR, November 2020. URL https://proceedings.mlr.press/v119/kristiadi20a.html.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5905–5914. PMLR, July 2021. URL https://proceedings.mlr.press/v139/kwon21b.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14434–14447. Curran Associates, Inc., 2022. doi: 10.52202/068431-1049. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5d087955ee13fe9a7402eedec879b9c3-Abstract-Conference.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html.

Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. COCO-O: A Benchmark for Object Detectors under Natural Distribution Shifts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6316–6327, October 2023. doi: 10.1109/ICCV51070.2023.00583. URL https://ieeexplore.ieee.org/abstract/document/10376670. ISSN: 2380-7504.

Thomas Möllenhoff and Mohammad Emtiyaz Khan. SAM as an Optimal Relaxation of Bayes, December 2023. URL http://arxiv.org/abs/2210.01620. arXiv:2210.01620 [cs.LG].

Adam T. Müller, Tobias Rögelein, and Nicolaj C. Stache. Monte Carlo Stochastic Depth for Uncertainty Estimation in Deep Learning. pp. 4250–4260, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/SAIAD/html/Muller_Monte_Carlo_Stochastic_Depth_for_Uncertainty_Estimation_in_Deep_Learning_CVPRW_2026_paper.html.

Van-Anh Nguyen, Trung Le, Anh Bui, Thanh-Toan Do, and Dinh Phung. Optimal Transport Model Distributional Robustness. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24074–24087. Curran Associates, Inc., 2023. doi: 10.52202/075280-1045. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4b91825aec2ed35150f1d3e8fb195556-Abstract-Conference.html.

Vardan Papyan. Traces of Class/Cross-Class Structure Pervade Deep Learning Spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020. ISSN 1533-7928. URL http://jmlr.org/papers/v21/20-933.html.

Chanwoong Park, Uijeong Jang, Ernest K. Ryu, and Insoon Yang. Sharpness-Aware Minimization Can Hallucinate Minimizers, February 2026. URL http://arxiv.org/abs/2509.21818. arXiv:2509.21818 [cs.LG].

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. pp. 779–788, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A Scalable Laplace Approximation for Neural Networks. February 2018. URL https://openreview.net/forum?id=Skdvd2xAZ.

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Function-Space Regularization in Neural Networks: A Probabilistic Perspective. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 29275–29290. PMLR, July 2023. URL https://proceedings.mlr.press/v202/rudner23a.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks, June 2017. URL https://arxiv.org/abs/1706.04454v3.

Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic Insights into Structure and Rank of Neural Network Hessian Maps. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23914–23927. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/c900ced7451da79502d29aa37ebb7b60-Abstract.html.

Chengli Tan, Yubo Zhou, Haishan Ye, Guang Dai, Junmin Liu, Zengjie Song, Jiangshe Zhang, Zixiang Zhao, Yunda Hao, and Yong Xu. Towards Understanding The Calibration Benefits of Sharpness-Aware Minimization. *International Conference on Learning Representations*, 2026: 87728–87748, April 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8ce0b141fb61bb847898d0b6258f9acf-Abstract-Conference.html.

Ba-Hien Tran, Simone Rossi, Dimitrios Milios, and Maurizio Filippone. All You Need is a Good Functional Prior for Bayesian Deep Learning. *Journal of Machine Learning Research*, 23(74): 1–56, 2022. ISSN 1533-7928. URL http://jmlr.org/papers/v23/20-1340.html.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: An alternative approach to efficient ensemble and lifelong learning. *Eighth International Conference on Learning Representations (ICLR 2020)*, 2020. doi: 10.48550/arxiv.2002.06715.

Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How Good is the Bayes Posterior in Deep Neural Networks Really?, July 2020. URL http://arxiv.org/abs/2002.02405. arXiv:2002.02405 [stat.ML].

Yijie Zhang, Yi-Shan Wu, Luis A. Ortega, and Andres R. Masegosa. The Cold Posterior Effect Indicates Underfitting, and Cold Posteriors Represent a Fully Bayesian Method to Mitigate It. *Transactions on Machine Learning Research*, April 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=GZORXGxHHT.

A Appendix

A.1 Reproducibility Statement

Exemplary source code for the implementation of SAM/ASAM, as well as the diagnostic stochastic surrogate (shown in Algorithm 1) is provided in the GitHub[1] repository.

Optimization and Fine-Tuning Budget. All networks were fine-tuned for 10 epochs using a Cosine Annealing learning rate schedule, where SWAG was trained for additional 10 epochs to accumulate trajectory variance. To establish the base training trajectory, we utilized SGD (momentum = 0.9) for ResNet-50, and AdamW (weight decay = 1×10^{-4}) for ViT-B/16 and YOLOv8x. Crucially, the baselines are evaluated using their native optimal training trajectories (SGD or AdamW). SWAG fundamentally relies on the unguided, noisy trajectory of standard SGD to populate its covariance matrix; applying heavily regularized flat-minima optimizers (SAM/ASAM) might actively suppress such temporal variance, collapsing the SWAG covariance. Crucially, while Automatic Mixed Precision (AMP) can be used for standard SGD training, the geometry extraction phase of SAM/ASAM for stochastic sampling tests required full FP32 precision to prevent gradient underflow when calculating the adversarial step vector (ϵ^*).

Wasserstein Radius (ρ) Configurations. The geometric boundary of the predictive posterior is governed by the neighborhood parameter ρ utilized during the SAM/ASAM training phase. Because different architectures exhibit drastically varying topographical sensitivities, the optimal ρ is architecture-dependent. We utilized the following configurations across our experiments: ResNet-50: $\rho = 0.5$, ViT-B/16: $\rho = 0.2$, YOLOv8x: $\rho = 0.001$ (for both SAM and ASAM).

The small ρ required for YOLOv8x empirically validates our finding that dense, multi-scale detection heads operate on scale-sensitive landscapes, thus requiring ASAM's localized p^2 safety brake to prevent structural collapse.

A.2 Mathematical Reduction of W-DRO to SAM

To ensure the theoretical foundation given in the main text is fully transparent, we detail the mathematical reduction proving that SAM is formally equivalent to W-DRO in weight space.

Step 1: Kantorovich Strong Duality. We begin with the infinite-dimensional inner maximization of the W-DRO objective over the approximate posterior q. By utilizing the strong duality of Optimal Transport (specifically the robust optimization duality formulated by Blanchet et al. (2019)), we can convert this inner maximization into a finite-dimensional minimization over a scalar dual variable $\lambda \geq 0$:

$$\max_{q:\theta_1(q,\delta_{\theta^*})\leq\rho} \mathbb{E}_q[L(\theta)] = \min_{\lambda\geq 0} \left\{ \lambda\rho + \mathbb{E}_{\theta\sim\delta_{\theta^*}} \left[\max_{\theta'} (L(\theta') - \lambda\|\theta' - \theta\|_p) \right] \right\} \tag{3}$$

As the expectation is evaluated strictly over the Dirac delta δ_{θ^*}, the equation collapses to evaluating directly at the point θ^*:

$$= \min_{\lambda\geq 0} \left\{ \lambda\rho + \max_{\theta'} (L(\theta') - \lambda\|\theta' - \theta^*\|_p) \right\} \tag{4}$$

Step 2: First-Order Taylor Approximation. To solve the new inner maximization, we assume the perturbed weights θ' remain close to θ^* (bounded by a small Wasserstein radius ρ). We apply a first-order Taylor expansion to the loss function around θ^*:

$$L(\theta') \approx L(\theta^*) + \nabla_\theta L(\theta^*)^T(\theta' - \theta^*) \tag{5}$$

Let $\Delta\theta = \theta' - \theta^*$. Substituting this expansion back into the maximization yields:

[1] https://github.com/anon-code-supplement-27/ICLR_supplementary_code

Table 4: Activation Stability for ResNet-50 on ImageNet-1k. Stochastically sampled SAM variant evaluated across varying batch normalization recalibration sizes (M).

Recalib. Images (M)	Top-1 Acc (↑)	ECE (↓)	NLL (↓)	AUARC (↑)
0 (Instant Deploy)	**0.7986**	0.0152	0.7696	0.9441
128	0.7981	0.0143	**0.7685**	0.9443
256	0.7985	0.0148	0.7686	0.9446
640	0.7979	**0.0137**	0.7699	**0.9447**
1,280	0.7974	0.0149	0.7754	0.9447
2,560	0.7961	0.0177	0.7816	0.9445

Table 5: Activation Stability for YOLOv8x on COCO. Stochastically sampled ASAM variant evaluated under dense bounding-box regression dynamics. (Note: AmT measures Amortized Throughput inclusive of the recalibration pass overhead).

Recalib. Images (M)	mAP50-95 (↑)	ECE (↓)	AUARC (↑)	AmT (ms/img)
0 (Instant Deploy)	0.5433	0.070	**0.919**	**21.82**
32	0.5442	0.068	**0.919**	22.77
64	**0.5447**	0.066	0.918	23.46
160	0.5446	0.063	0.917	26.15
320	0.5432	**0.062**	0.915	30.53

$$L(\theta^*) + \max_{\Delta\theta}(\nabla_\theta L(\theta^*)^T \Delta\theta - \lambda\|\Delta\theta\|_p) \tag{6}$$

Step 3: Dual Norm Constraint. By the definition of the dual norm $\|\cdot\|_q$ (where $1/p + 1/q = 1$), the maximum value over $\Delta\theta$ evaluates exactly to 0 if the condition $\|\nabla_\theta L(\theta^*)\|_q \leq \lambda$ holds, and explodes to $+\infty$ otherwise. To minimize the outer equation with respect to λ, we must select the absolute smallest λ that avoids positive infinity. Therefore:

$$\lambda^* = \|\nabla_\theta L(\theta^*)\|_q \tag{7}$$

Step 4: The Final SAM Objective. Plugging λ^* back into the dual equation, the complex W-DRO problem collapses into a simple regularized loss function:

$$\min_{\theta^*} \{L(\theta^*) + \rho\|\nabla_\theta L(\theta^*)\|_q\} \tag{8}$$

This confirms that perturbing weights along the gradient vector is mathematically equivalent to minimizing the expected loss over a distribution strictly bounded by the 1-Wasserstein distance. *(Note: This derivation assumes a uniform Euclidean distance metric. By substituting the standard distance with a scale-aware Mahalanobis distance metric $\|T^{-1}\Delta\theta\|_p$, the dual norm constraint organically yields the ASAM objective, thereby preserving the W-DRO duality for heterogeneous architectures.)*

A.3 Further Ablations

A.3.1 Batch Normalization Activation Stability

A fundamental limitation of unconstrained weight-sampling methods (such as SWAG) is the disruption of BN running statistics. As unconstrained Gaussian noise perturbs intermediate feature activations, such methods require a mandatory gradient-free recalibration pass over M training images prior to evaluating the test set. This data-dependent step can bottleneck deployment in edge and streaming scenarios.

To isolate this effect, we formalized the Wasserstein-Bounded Stochastic Surrogate (Algorithm 1), which applies optimally scaled isotropic variance directly to the deterministic SAM/ASAM checkpoint. As shown in Tables 4 and 5, applying this geometrically safe-bounded sampling to ResNet-50

Table 6: Geometric Validation (ResNet-50). Evaluation of the true adversarial vector (ϵ^*) against a Random Gaussian vector of identical geometric norm.

Dataset	Noise Geometry	Acc (↑)	ECE (↓)	Entropy (↑)	AUARC (↑)
ImageNet-1k (ID)	Adversarial Vector (ϵ^*)	**0.8065**	**0.0469**	**0.9735**	**0.9489**
	Random Vector (Same Norm)	0.8062	0.0474	0.9711	**0.9489**
ImageNet-O	Adversarial Vector (ϵ^*)	0.0000	**0.7611**	**1.8330**	0.0000
	Random Vector (Same Norm)	0.0000	0.7617	1.8282	0.0000
ImageNet-C (Gauss 5)	Adversarial Vector (ϵ^*)	0.3218	**0.0478**	**5.1476**	0.5761
	Random Vector (Same Norm)	**0.3219**	0.0485	5.1405	**0.5762**
ImageNet-C (Zoom 5)	Adversarial Vector (ϵ^*)	**0.3254**	**0.1210**	**4.5135**	**0.5840**
	Random Vector (Same Norm)	0.3252	0.1222	4.5050	**0.5840**

and YOLOv8x circumvents the Activation Shift bottleneck. The safe variance budget natively preserves activation distributions, allowing for stable zero-recalibration ($M = 0$) deployment, retaining $> 99.7\%$ of peak mAP50-95 on YOLOv8x.

This ablation highlights the geometric tension of single-basin sampling: if the variance is strictly bounded enough to preserve internal activations ($M = 0$), it is too small to alter the predictive logits. Therefore, to force measurable logit disagreement, practical samplers are forced to artificially inflate the variance beyond the safe Wasserstein boundary, thereby degrading the activations and triggering Functional Disruption.

Algorithm 1 Synthetic W-DRO Boundary Ablation

Require: Base model f_θ, Training data $\mathcal{D}$, Wasserstein radius ρ, Samples N, Parameter dim. D

Phase 1: Deterministic Geometry Extraction (Training)
for each batch $(X, y) \in \mathcal{D}$ **do**
  Compute base gradient: $g = \nabla_\theta \mathcal{L}(f_\theta(X), y)$ ▷ First forward-backward pass
  Define scaling operator T ▷ $T = \mathbf{I}$ for SAM, $T = |\theta|$ for ASAM
  Extract adversarial boundary: $\epsilon^* = \rho \frac{T^2 g}{\|Tg\|_2}$
  Compute adversarial gradient: $g_{adv} = \nabla_\theta \mathcal{L}(f_{\theta+\epsilon^*}(X), y)$ ▷ Second pass
  Update weights $\theta \leftarrow \theta - \eta \cdot g_{adv}$
end for
Save deterministic weights θ^* and geometric radius ρ.

Phase 2: Bounded Isotropic Sampling (Inference Ablation)
Calculate maximum permissible variance: $\beta = \frac{\rho^2}{Tr(T^2)}$ ▷ Defaults to ρ^2/D for standard SAM
for each test input x_{test} **do**
  for $i = 1$ **to** N **do**
    Sample isotropic background vector: $z \sim \mathcal{N}(0, \mathbf{I})$
    Apply bounded perturbation: $\theta_i = \theta^* + \sqrt{\beta} \cdot z \cdot T$
    Compute stochastic forward pass: $p_i = f_{\theta_i}(x_{test})$
  end for
  return Predictive Mean $\mu = \frac{1}{N} \sum p_i$ and Predictive Entropy $\mathcal{H}(p_1, \ldots, p_N)$
end for

A.3.2 Extended Geometric Validation

In Section 3.2 (Horn 1: Dimensionality Dilution), the mathematical formulation relies on the premise of an isotropic predictive distribution $Q = \mathcal{N}(\theta^*, \beta I)$. To empirically validate this isotropic assumption, Table 6 compares the deterministic ResNet-50 checkpoint perturbed by its exact final-epoch adversarial vector against an unstructured random Gaussian vector scaled to the identical geometric norm.

The results demonstrate parity between the two geometries. Not only does the unstructured random vector match the Top-1 Accuracy and ECE of the true ϵ^* vector on ImageNet-1k, but it also perfectly mirrors the network's predictive entropy on OOD ImageNet-O anomalies (1.828 vs. 1.833). Furthermore, under high-frequency corruptions (Gaussian Noise) and structural corruptions (Zoom Blur), the AUARC remains virtually identical.

This suggests that the local geometry optimized by SAM exhibits isotropic flatness, rather than being robust solely in the specific adversarial direction of the final training batch. These findings support the premise of our synthetic ablation: safely bounded isotropic sampling locally mirrors the deterministic optimum, triggering Dimensionality Dilution.

We evaluated the deterministic ResNet-50 checkpoint perturbed by the exact adversarial vector (ϵ^*) stored during the final training epoch, and compared it against an unstructured random Gaussian vector scaled to the same geometric norm.

Table 7: Full ImageNet-C Evaluation for ResNet-50. Top-1 Accuracy (↑), Expected Calibration Error (↓), and Predictive Entropy (↑) across all 15 corruptions and 5 severities.

	Severity 1			Severity 2			Severity 3			Severity 4			Severity 5		
Method	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)
Gaussian Noise															
MCD	0.649	0.035	1.839	0.552	0.031	2.276	0.392	0.006	2.915	0.232	0.037	3.619	0.086	0.076	4.673
MCDB	0.655	0.012	1.579	0.567	0.026	1.942	0.414	0.065	2.474	0.255	0.108	3.090	0.103	0.141	3.992
MCSD	0.589	0.007	2.073	0.490	0.048	2.463	0.334	0.035	3.033	0.198	0.064	3.659	0.066	0.098	4.664
SWAG	0.657	0.018	1.536	0.573	0.012	1.879	0.428	0.031	2.441	0.276	0.052	3.123	0.086	0.113	4.316
Deep Ensemble	0.686	0.022	1.564	0.612	0.011	1.890	0.470	0.023	2.454	0.301	0.053	3.204	0.105	0.143	4.414
Det. SAM	0.648	0.047	1.802	0.553	0.031	2.247	0.392	0.003	2.923	0.237	0.031	3.731	0.094	0.055	5.148
Det. ASAM	0.672	0.051	1.636	0.584	0.034	2.052	0.435	0.012	2.709	0.275	0.020	3.513	0.128	0.060	4.654
Shot Noise															
MCD	0.642	0.038	1.884	0.536	0.025	2.378	0.370	0.012	3.033	0.207	0.051	3.782	0.085	0.087	4.693
MCDB	0.651	0.009	1.614	0.548	0.034	2.029	0.389	0.076	2.580	0.224	0.120	3.243	0.100	0.148	4.025
MCSD	0.582	0.008	2.112	0.473	0.044	2.548	0.315	0.045	3.123	0.176	0.074	3.791	0.064	0.107	4.707
SWAG	0.651	0.021	1.571	0.556	0.004	1.954	0.406	0.039	2.546	0.244	0.068	3.300	0.086	0.117	4.331
Deep Ensemble	0.678	0.020	1.597	0.597	0.004	1.961	0.449	0.033	2.553	0.269	0.067	3.376	0.102	0.148	4.437
Det. SAM	0.642	0.044	1.843	0.538	0.026	2.341	0.370	0.009	3.037	0.210	0.038	3.910	0.093	0.056	5.143
Det. ASAM	0.665	0.047	1.674	0.569	0.028	2.130	0.412	0.003	2.825	0.245	0.028	3.693	0.126	0.062	4.671
Impulse Noise															
MCD	0.608	0.033	1.942	0.505	0.030	2.327	0.370	0.016	2.923	0.205	0.034	3.659	0.069	0.091	4.636
MCDB	0.612	0.011	1.693	0.515	0.035	2.030	0.395	0.063	2.520	0.231	0.111	3.161	0.081	0.158	4.018
MCSD	0.556	0.014	2.164	0.449	0.045	2.502	0.317	0.041	3.023	0.170	0.071	3.695	0.052	0.118	4.664
SWAG	0.610	0.011	1.666	0.528	0.010	1.966	0.407	0.029	2.457	0.248	0.056	3.160	0.065	0.115	4.298
Deep Ensemble	0.639	0.010	1.684	0.559	0.013	1.989	0.446	0.018	2.463	0.278	0.054	3.204	0.083	0.146	4.412
Det. SAM	0.605	0.037	1.921	0.506	0.032	2.309	0.371	0.014	2.916	0.210	0.027	3.738	0.073	0.051	5.151
Det. ASAM	0.627	0.041	1.761	0.536	0.032	2.132	0.413	0.010	2.686	0.247	0.018	3.486	0.096	0.070	4.686
Defocus Blur															
MCD	0.628	0.060	1.572	0.527	0.048	2.062	0.395	0.038	2.671	0.277	0.050	3.250	0.163	0.091	4.015
MCDB	0.640	0.028	1.340	0.551	0.040	1.748	0.419	0.086	2.259	0.301	0.126	2.776	0.190	0.174	3.473
MCSD	0.569	0.037	1.831	0.468	0.067	2.298	0.337	0.069	2.859	0.228	0.083	3.376	0.127	0.111	4.093
SWAG	0.642	0.048	1.328	0.548	0.022	1.706	0.435	0.023	2.193	0.320	0.053	2.766	0.186	0.121	3.567
Deep Ensemble	0.672	0.050	1.332	0.579	0.018	1.714	0.470	0.019	2.203	0.354	0.054	2.793	0.210	0.134	3.655
Det. SAM	0.630	0.073	1.542	0.523	0.050	2.036	0.388	0.028	2.659	0.266	0.024	3.298	0.155	0.072	4.225
Det. ASAM	0.655	0.067	1.405	0.558	0.046	1.874	0.428	0.020	2.456	0.309	0.014	3.064	0.187	0.076	3.882
Glass Blur															
MCD	0.574	0.046	1.942	0.439	0.032	2.593	0.215	0.040	3.593	0.134	0.080	4.195	0.093	0.076	4.502
MCDB	0.582	0.023	1.688	0.457	0.031	2.261	0.240	0.106	3.123	0.157	0.155	3.649	0.111	0.147	3.907
MCSD	0.520	0.024	2.179	0.386	0.044	2.802	0.181	0.054	3.738	0.108	0.085	4.254	0.073	0.095	4.549
SWAG	0.584	0.019	1.666	0.459	0.016	2.203	0.255	0.043	3.136	0.155	0.099	3.791	0.096	0.104	4.185
Deep Ensemble	0.608	0.021	1.683	0.485	0.019	2.228	0.279	0.040	3.197	0.177	0.105	3.886	0.111	0.122	4.304
Det. SAM	0.573	0.050	1.921	0.438	0.027	2.576	0.218	0.023	3.626	0.138	0.052	4.382	0.097	0.057	4.887
Det. ASAM	0.596	0.050	1.765	0.468	0.027	2.378	0.245	0.012	3.409	0.160	0.061	4.150	0.113	0.080	4.597
Motion Blur															
MCD	0.607	0.051	1.670	0.505	0.038	2.115	0.360	0.033	2.766	0.229	0.058	3.447	0.125	0.085	4.208
MCDB	0.615	0.020	1.428	0.524	0.040	1.821	0.383	0.094	2.375	0.252	0.134	2.975	0.147	0.170	3.666
MCSD	0.548	0.028	1.916	0.446	0.056	2.338	0.306	0.057	2.934	0.191	0.076	3.541	0.101	0.102	4.254
SWAG	0.617	0.036	1.421	0.526	0.016	1.790	0.400	0.038	2.298	0.266	0.065	2.923	0.144	0.115	3.738
Deep Ensemble	0.648	0.042	1.426	0.559	0.011	1.802	0.434	0.025	2.316	0.300	0.063	2.966	0.165	0.134	3.842
Det. SAM	0.606	0.063	1.644	0.505	0.042	2.091	0.354	0.016	2.762	0.222	0.027	3.504	0.125	0.076	4.417
Det. ASAM	0.631	0.058	1.503	0.533	0.035	1.934	0.386	0.007	2.569	0.254	0.021	3.275	0.150	0.079	4.103
Zoom Blur															
MCD	0.573	0.050	1.838	0.506	0.039	2.148	0.428	0.032	2.502	0.316	0.046	3.037	0.207	0.066	3.657

Continued on next page

Table 7 – continued from previous page

	Severity 1			Severity 2			Severity 3			Severity 4			Severity 5		
Method	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)
MCDB	0.584	0.016	1.579	0.523	0.034	1.854	0.449	0.074	2.164	0.342	0.115	2.624	0.231	0.143	3.176
MCSD	0.517	0.027	2.062	0.450	0.052	2.361	0.372	0.048	2.696	0.266	0.060	3.204	0.167	0.082	3.793
SWAG	0.586	0.029	1.583	0.528	0.010	1.831	0.454	0.030	2.138	0.347	0.045	2.636	0.237	0.087	3.250
Deep Ensemble	0.613	0.034	1.594	0.558	0.010	1.848	0.485	0.015	2.158	0.380	0.035	2.668	0.267	0.093	3.324
Det. SAM	0.571	0.063	1.815	0.502	0.045	2.126	0.419	0.021	2.503	0.311	0.013	3.078	0.198	0.034	3.821
Det. ASAM	0.597	0.056	1.666	0.531	0.035	1.961	0.452	0.005	2.316	0.341	0.007	2.859	0.230	0.031	3.512
Snow															
MCD	0.536	0.040	2.091	0.434	0.029	2.565	0.344	0.017	3.013	0.203	0.040	3.821	0.126	0.076	4.390
MCDB	0.548	0.020	1.815	0.452	0.045	2.235	0.364	0.072	2.618	0.222	0.134	3.313	0.143	0.158	3.811
MCSD	0.481	0.017	2.309	0.378	0.039	2.766	0.293	0.039	3.176	0.165	0.066	3.937	0.100	0.093	4.475
SWAG	0.549	0.014	1.777	0.457	0.015	2.179	0.373	0.030	2.541	0.233	0.075	3.284	0.145	0.104	3.921
Deep Ensemble	0.574	0.016	1.796	0.483	0.012	2.193	0.404	0.022	2.547	0.260	0.078	3.344	0.164	0.125	4.025
Det. SAM	0.532	0.048	2.057	0.428	0.031	2.534	0.340	0.010	2.997	0.204	0.034	3.905	0.131	0.062	4.603
Det. ASAM	0.559	0.048	1.884	0.459	0.027	2.327	0.369	0.006	2.766	0.236	0.021	3.633	0.159	0.071	4.254
Frost															
MCD	0.569	0.042	1.990	0.436	0.032	2.624	0.294	0.027	3.266	0.218	0.039	3.731	0.133	0.070	4.314
MCDB	0.581	0.019	1.714	0.457	0.042	2.292	0.316	0.089	2.852	0.237	0.127	3.250	0.152	0.160	3.754
MCSD	0.512	0.017	2.213	0.380	0.043	2.825	0.250	0.048	3.424	0.180	0.066	3.864	0.106	0.093	4.414
SWAG	0.584	0.015	1.684	0.461	0.016	2.228	0.322	0.041	2.782	0.244	0.076	3.220	0.150	0.111	3.821
Deep Ensemble	0.609	0.016	1.706	0.490	0.016	2.247	0.347	0.041	2.802	0.272	0.077	3.284	0.170	0.131	3.921
Det. SAM	0.565	0.048	1.961	0.432	0.030	2.601	0.292	0.016	3.253	0.215	0.031	3.805	0.135	0.061	4.502
Det. ASAM	0.593	0.049	1.790	0.462	0.026	2.404	0.318	0.012	3.013	0.246	0.019	3.551	0.162	0.070	4.168
Fog															
MCD	0.578	0.033	1.956	0.533	0.037	2.204	0.472	0.034	2.502	0.412	0.035	2.793	0.273	0.043	3.493
MCDB	0.590	0.012	1.684	0.547	0.023	1.905	0.491	0.053	2.158	0.434	0.074	2.410	0.296	0.114	3.033
MCSD	0.521	0.011	2.179	0.478	0.045	2.420	0.419	0.046	2.709	0.358	0.050	3.003	0.231	0.066	3.659
SWAG	0.591	0.010	1.660	0.551	0.012	1.868	0.493	0.024	2.120	0.436	0.032	2.375	0.298	0.062	3.037
Deep Ensemble	0.619	0.012	1.670	0.581	0.010	1.884	0.523	0.015	2.138	0.466	0.020	2.404	0.326	0.060	3.103
Det. SAM	0.575	0.038	1.934	0.530	0.040	2.174	0.467	0.033	2.484	0.407	0.024	2.802	0.272	0.021	3.619
Det. ASAM	0.601	0.038	1.771	0.557	0.039	1.998	0.496	0.029	2.298	0.437	0.017	2.601	0.301	0.023	3.324
Brightness															
MCD	0.722	0.031	1.346	0.686	0.035	1.543	0.638	0.040	1.796	0.579	0.041	2.099	0.495	0.044	2.527
MCDB	0.726	0.019	1.144	0.693	0.013	1.319	0.648	0.013	1.542	0.590	0.030	1.815	0.509	0.052	2.203
MCSD	0.665	0.023	1.565	0.627	0.036	1.777	0.580	0.042	2.036	0.518	0.045	2.341	0.439	0.045	2.766
SWAG	0.730	0.035	1.137	0.697	0.019	1.309	0.650	0.013	1.528	0.595	0.012	1.785	0.513	0.022	2.164
Deep Ensemble	0.751	0.036	1.157	0.718	0.023	1.328	0.672	0.014	1.551	0.621	0.012	1.802	0.540	0.015	2.193
Det. SAM	0.725	0.036	1.322	0.688	0.041	1.522	0.637	0.045	1.796	0.577	0.043	2.126	0.492	0.036	2.611
Det. ASAM	0.742	0.036	1.218	0.707	0.041	1.405	0.660	0.044	1.654	0.603	0.040	1.954	0.522	0.027	2.404
Contrast															
MCD	0.528	0.052	2.185	0.334	0.023	3.129	0.169	0.027	3.864	0.063	0.076	4.549	0.019	0.129	5.161
MCDB	0.545	0.031	1.884	0.354	0.077	2.766	0.187	0.122	3.447	0.076	0.147	4.103	0.026	0.170	4.707
MCSD	0.472	0.042	2.435	0.287	0.047	3.324	0.140	0.068	4.025	0.052	0.093	4.636	0.016	0.120	5.213
SWAG	0.547	0.021	1.874	0.358	0.038	2.724	0.187	0.068	3.447	0.073	0.111	4.137	0.022	0.115	4.786
Deep Ensemble	0.573	0.024	1.896	0.384	0.036	2.754	0.207	0.067	3.486	0.084	0.118	4.185	0.026	0.118	4.831
Det. SAM	0.525	0.054	2.179	0.329	0.016	3.176	0.164	0.022	4.025	0.061	0.058	4.821	0.018	0.088	5.539
Det. ASAM	0.551	0.054	2.015	0.356	0.013	2.955	0.182	0.012	3.791	0.070	0.064	4.549	0.022	0.100	5.231
Elastic Transform															
MCD	0.655	0.032	1.706	0.449	0.028	2.502	0.364	0.022	2.873	0.279	0.038	3.250	0.204	0.046	3.659
MCDB	0.662	0.010	1.458	0.468	0.058	2.179	0.385	0.076	2.502	0.301	0.105	2.814	0.223	0.120	3.161
MCSD	0.597	0.009	1.942	0.395	0.044	2.716	0.313	0.048	3.064	0.237	0.066	3.424	0.170	0.076	3.821
SWAG	0.663	0.016	1.436	0.467	0.026	2.138	0.387	0.034	2.441	0.302	0.057	2.766	0.223	0.076	3.113
Deep Ensemble	0.692	0.021	1.447	0.499	0.019	2.148	0.418	0.024	2.454	0.332	0.047	2.793	0.248	0.075	3.160
Det. SAM	0.654	0.044	1.670	0.448	0.023	2.474	0.362	0.012	2.873	0.273	0.017	3.324	0.198	0.026	3.842
Det. ASAM	0.677	0.049	1.528	0.474	0.016	2.298	0.388	0.007	2.671	0.298	0.010	3.078	0.221	0.025	3.525
Pixelate															
MCD	0.638	0.036	1.831	0.601	0.035	1.990	0.489	0.024	2.553	0.313	0.013	3.424	0.203	0.034	4.103
MCDB	0.647	0.012	1.551	0.612	0.016	1.693	0.505	0.049	2.179	0.332	0.090	2.923	0.222	0.126	3.551
MCSD	0.579	0.011	2.052	0.542	0.025	2.203	0.428	0.047	2.735	0.264	0.054	3.541	0.165	0.074	4.185
SWAG	0.647	0.019	1.536	0.612	0.014	1.670	0.508	0.015	2.138	0.334	0.038	2.887	0.223	0.081	3.525
Deep Ensemble	0.677	0.021	1.551	0.643	0.015	1.684	0.539	0.006	2.158	0.364	0.029	2.915	0.248	0.080	3.567
Det. SAM	0.637	0.045	1.796	0.598	0.041	1.972	0.488	0.022	2.576	0.310	0.013	3.567	0.203	0.025	4.390
Det. ASAM	0.660	0.048	1.625	0.622	0.045	1.785	0.514	0.022	2.341	0.336	0.004	3.250	0.224	0.035	3.992
Jpeg Compression															
MCD	0.570	0.034	1.956	0.507	0.031	2.221	0.468	0.031	2.384	0.413	0.034	2.618	0.360	0.035	2.852
MCDB	0.582	0.011	1.688	0.521	0.021	1.921	0.486	0.033	2.062	0.435	0.052	2.261	0.383	0.076	2.474
MCSD	0.512	0.012	2.185	0.453	0.032	2.441	0.417	0.036	2.587	0.363	0.043	2.793	0.314	0.053	2.997
SWAG	0.583	0.012	1.666	0.523	0.008	1.884	0.488	0.013	2.029	0.437	0.022	2.221	0.385	0.039	2.420
Deep Ensemble	0.612	0.012	1.674	0.554	0.008	1.896	0.518	0.012	2.036	0.467	0.018	2.228	0.416	0.028	2.435
Det. SAM	0.567	0.038	1.934	0.504	0.036	2.203	0.465	0.032	2.378	0.408	0.029	2.645	0.354	0.023	2.923
Det. ASAM	0.592	0.041	1.777	0.530	0.037	2.029	0.493	0.032	2.185	0.437	0.023	2.420	0.384	0.015	2.659

Table 8: Full ImageNet-C Evaluation for ViT-B/16. Top-1 Accuracy (↑), Expected Calibration Error (↓), and Predictive Entropy (↑) across all 15 corruptions and 5 severities.

	Severity 1			Severity 2			Severity 3			Severity 4			Severity 5		
Method	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)
Gaussian Noise															
MCD	0.702	0.054	1.588	0.589	0.063	1.944	0.407	0.076	2.457	0.287	0.072	2.809	0.169	0.072	3.232
MCDB	0.686	0.088	1.455	0.573	0.106	1.776	0.395	0.141	2.235	0.279	0.148	2.548	0.165	0.166	2.924
MCSD	0.697	0.038	1.834	0.573	0.061	2.259	0.380	0.076	2.822	0.259	0.072	3.203	0.146	0.068	3.655
SWAG	0.718	0.063	1.439	0.612	0.064	1.782	0.435	0.066	2.292	0.312	0.053	2.658	0.187	0.048	3.111
Deep Ensemble	0.725	0.047	1.437	0.622	0.049	1.775	0.449	0.053	2.274	0.329	0.045	2.631	0.205	0.041	3.071
Det. SAM	0.724	0.042	1.637	0.622	0.049	1.996	0.450	0.057	2.518	0.329	0.048	2.894	0.203	0.046	3.342
Det. ASAM	0.724	0.053	1.543	0.621	0.058	1.884	0.449	0.066	2.399	0.328	0.057	2.766	0.203	0.053	3.209
Shot Noise															
MCD	0.686	0.056	1.619	0.569	0.066	1.998	0.428	0.077	2.428	0.286	0.074	2.868	0.222	0.068	3.125
MCDB	0.670	0.091	1.485	0.552	0.111	1.826	0.413	0.136	2.213	0.276	0.147	2.607	0.215	0.155	2.836
MCSD	0.679	0.041	1.868	0.551	0.065	2.316	0.403	0.079	2.793	0.258	0.076	3.275	0.198	0.069	3.551
SWAG	0.702	0.064	1.469	0.592	0.064	1.834	0.456	0.063	2.261	0.312	0.053	2.716	0.245	0.046	2.997
Deep Ensemble	0.709	0.049	1.466	0.603	0.050	1.826	0.471	0.049	2.247	0.328	0.045	2.686	0.262	0.038	2.966
Det. SAM	0.708	0.044	1.666	0.603	0.051	2.052	0.470	0.053	2.484	0.328	0.049	2.955	0.261	0.042	3.232
Det. ASAM	0.708	0.055	1.571	0.603	0.061	1.938	0.470	0.063	2.361	0.327	0.057	2.825	0.261	0.051	3.103
Impulse Noise															
MCD	0.697	0.056	1.614	0.612	0.063	1.890	0.505	0.066	2.247	0.354	0.070	2.716	0.264	0.068	3.049
MCDB	0.683	0.089	1.479	0.597	0.103	1.730	0.490	0.118	2.046	0.342	0.137	2.474	0.255	0.150	2.766
MCSD	0.691	0.040	1.864	0.597	0.056	2.193	0.482	0.064	2.580	0.324	0.068	3.103	0.237	0.067	3.473
SWAG	0.714	0.063	1.464	0.635	0.063	1.734	0.531	0.057	2.083	0.380	0.053	2.576	0.288	0.049	2.934
Deep Ensemble	0.722	0.048	1.462	0.645	0.048	1.727	0.547	0.043	2.062	0.400	0.041	2.541	0.306	0.040	2.894
Det. SAM	0.721	0.043	1.666	0.645	0.048	1.942	0.546	0.047	2.309	0.399	0.048	2.814	0.306	0.047	3.189
Det. ASAM	0.721	0.054	1.571	0.645	0.057	1.831	0.546	0.055	2.179	0.399	0.055	2.671	0.306	0.053	3.033
Defocus Blur															
MCD	0.696	0.058	1.583	0.617	0.064	1.838	0.490	0.068	2.259	0.380	0.071	2.618	0.334	0.069	2.809
MCDB	0.684	0.090	1.455	0.603	0.103	1.691	0.476	0.118	2.064	0.367	0.129	2.384	0.320	0.134	2.553
MCSD	0.691	0.041	1.831	0.604	0.052	2.138	0.468	0.063	2.593	0.354	0.066	2.997	0.306	0.066	3.204
SWAG	0.713	0.064	1.436	0.640	0.062	1.688	0.518	0.055	2.091	0.407	0.050	2.474	0.360	0.047	2.686
Deep Ensemble	0.719	0.049	1.437	0.650	0.046	1.684	0.533	0.040	2.073	0.426	0.038	2.441	0.379	0.037	2.645
Det. SAM	0.718	0.045	1.636	0.649	0.045	1.890	0.532	0.044	2.316	0.426	0.044	2.709	0.379	0.044	2.934
Det. ASAM	0.718	0.055	1.542	0.649	0.055	1.785	0.531	0.053	2.193	0.426	0.051	2.565	0.378	0.049	2.782
Glass Blur															
MCD	0.638	0.063	1.777	0.551	0.070	2.083	0.400	0.079	2.553	0.287	0.078	2.975	0.230	0.071	3.250
MCDB	0.622	0.100	1.636	0.535	0.113	1.914	0.384	0.141	2.338	0.274	0.150	2.724	0.218	0.155	2.975
MCSD	0.629	0.052	2.052	0.535	0.066	2.404	0.373	0.082	2.923	0.258	0.081	3.389	0.203	0.073	3.693
SWAG	0.659	0.065	1.619	0.575	0.064	1.916	0.427	0.064	2.392	0.313	0.053	2.836	0.255	0.047	3.129
Deep Ensemble	0.669	0.050	1.614	0.588	0.048	1.905	0.443	0.048	2.375	0.330	0.042	2.802	0.272	0.038	3.090
Det. SAM	0.668	0.049	1.826	0.587	0.050	2.138	0.442	0.053	2.636	0.329	0.049	3.090	0.271	0.045	3.398
Det. ASAM	0.667	0.059	1.727	0.586	0.059	2.029	0.441	0.061	2.502	0.329	0.056	2.944	0.271	0.051	3.243
Motion Blur															
MCD	0.663	0.059	1.691	0.581	0.065	1.954	0.442	0.075	2.404	0.321	0.075	2.814	0.247	0.069	3.113
MCDB	0.648	0.096	1.551	0.565	0.106	1.796	0.426	0.129	2.203	0.308	0.138	2.569	0.237	0.147	2.844
MCSD	0.655	0.046	1.954	0.565	0.061	2.259	0.415	0.078	2.766	0.292	0.081	3.220	0.219	0.074	3.538
SWAG	0.683	0.064	1.536	0.605	0.063	1.790	0.468	0.061	2.235	0.347	0.053	2.668	0.271	0.047	2.986
Deep Ensemble	0.691	0.049	1.532	0.615	0.047	1.785	0.483	0.045	2.221	0.364	0.042	2.636	0.288	0.038	2.944
Det. SAM	0.690	0.047	1.738	0.614	0.049	2.007	0.482	0.051	2.463	0.363	0.048	2.915	0.287	0.044	3.243
Det. ASAM	0.690	0.057	1.644	0.614	0.058	1.896	0.481	0.059	2.338	0.362	0.055	2.766	0.286	0.050	3.078
Zoom Blur															
MCD	0.656	0.061	1.734	0.584	0.068	1.972	0.485	0.073	2.298	0.395	0.075	2.611	0.291	0.156	2.873
MCDB	0.640	0.099	1.588	0.567	0.111	1.815	0.469	0.125	2.112	0.380	0.134	2.404	0.278	0.126	2.636
MCSD	0.647	0.049	2.007	0.568	0.065	2.276	0.460	0.078	2.636	0.368	0.082	2.986	0.261	0.100	3.284
SWAG	0.678	0.064	1.565	0.608	0.064	1.802	0.513	0.061	2.126	0.423	0.057	2.441	0.324	0.172	2.744
Deep Ensemble	0.686	0.049	1.564	0.618	0.049	1.796	0.528	0.045	2.115	0.440	0.043	2.420	0.321	0.130	2.716
Det. SAM	0.685	0.048	1.777	0.617	0.051	2.029	0.527	0.049	2.378	0.439	0.047	2.716	0.321	0.028	3.033
Det. ASAM	0.685	0.058	1.674	0.617	0.060	1.916	0.526	0.058	2.247	0.439	0.055	2.569	0.325	0.119	2.868
Snow															
MCD	0.638	0.060	1.751	0.536	0.069	2.103	0.446	0.076	2.404	0.366	0.076	2.668	0.297	0.072	2.934
MCDB	0.622	0.098	1.603	0.520	0.114	1.921	0.430	0.132	2.193	0.352	0.139	2.441	0.284	0.144	2.686
MCSD	0.629	0.048	2.015	0.520	0.067	2.420	0.419	0.082	2.766	0.339	0.085	3.064	0.267	0.082	3.366
SWAG	0.658	0.065	1.583	0.562	0.064	1.934	0.472	0.061	2.228	0.392	0.056	2.493	0.324	0.049	2.766
Deep Ensemble	0.668	0.050	1.588	0.573	0.048	1.938	0.487	0.046	2.221	0.408	0.043	2.474	0.340	0.038	2.744
Det. SAM	0.666	0.048	1.802	0.572	0.050	2.164	0.485	0.052	2.474	0.407	0.048	2.744	0.340	0.045	3.033
Det. ASAM	0.666	0.058	1.693	0.571	0.059	2.046	0.485	0.059	2.341	0.406	0.055	2.607	0.340	0.051	2.873
Frost															
MCD	0.660	0.057	1.706	0.536	0.071	2.115	0.400	0.080	2.553	0.321	0.081	2.802	0.228	0.073	3.103
MCDB	0.644	0.093	1.564	0.518	0.117	1.942	0.384	0.141	2.341	0.307	0.149	2.576	0.216	0.158	2.836
MCSD	0.651	0.043	1.972	0.519	0.071	2.441	0.373	0.085	2.944	0.292	0.087	3.232	0.200	0.078	3.551
SWAG	0.680	0.065	1.551	0.562	0.065	1.954	0.428	0.065	2.399	0.347	0.058	2.636	0.252	0.051	2.955
Deep Ensemble	0.689	0.050	1.542	0.573	0.050	1.942	0.443	0.050	2.384	0.363	0.046	2.618	0.269	0.041	2.923

Continued on next page

Table 8 – continued from previous page

	Severity 1			Severity 2			Severity 3			Severity 4			Severity 5		
Method	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)
Det. SAM	0.687	0.046	1.751	0.571	0.052	2.179	0.442	0.055	2.659	0.362	0.052	2.903	0.268	0.047	3.220
Det. ASAM	0.687	0.056	1.654	0.570	0.061	2.062	0.441	0.062	2.527	0.361	0.058	2.766	0.267	0.053	3.064
Fog															
MCD	0.640	0.061	1.751	0.601	0.063	1.884	0.551	0.064	2.036	0.511	0.064	2.164	0.428	0.069	2.457
MCDB	0.625	0.098	1.603	0.585	0.103	1.727	0.536	0.106	1.874	0.496	0.111	1.989	0.413	0.128	2.259
MCSD	0.632	0.049	2.015	0.591	0.052	2.185	0.540	0.055	2.351	0.499	0.056	2.502	0.413	0.065	2.825
SWAG	0.661	0.065	1.588	0.623	0.064	1.730	0.575	0.060	1.879	0.536	0.056	2.007	0.455	0.052	2.298
Deep Ensemble	0.670	0.050	1.583	0.633	0.048	1.714	0.586	0.045	1.868	0.548	0.043	1.996	0.468	0.042	2.276
Det. SAM	0.669	0.048	1.802	0.632	0.046	1.934	0.585	0.046	2.099	0.547	0.046	2.228	0.467	0.047	2.527
Det. ASAM	0.668	0.058	1.693	0.631	0.056	1.815	0.584	0.054	1.972	0.546	0.054	2.103	0.467	0.053	2.384
Brightness															
MCD	0.748	0.050	1.455	0.728	0.051	1.512	0.701	0.053	1.597	0.672	0.054	1.688	0.627	0.058	1.831
MCDB	0.732	0.082	1.319	0.713	0.084	1.378	0.686	0.087	1.464	0.657	0.090	1.551	0.612	0.097	1.684
MCSD	0.743	0.033	1.674	0.724	0.034	1.748	0.697	0.037	1.848	0.667	0.039	1.961	0.621	0.045	2.138
SWAG	0.764	0.061	1.309	0.745	0.061	1.362	0.718	0.061	1.442	0.690	0.062	1.528	0.647	0.062	1.666
Deep Ensemble	0.771	0.046	1.300	0.753	0.046	1.353	0.727	0.045	1.436	0.699	0.046	1.512	0.657	0.047	1.660
Det. SAM	0.770	0.038	1.485	0.752	0.039	1.551	0.725	0.040	1.654	0.698	0.041	1.748	0.656	0.044	1.921
Det. ASAM	0.769	0.049	1.405	0.751	0.050	1.458	0.725	0.050	1.564	0.698	0.052	1.644	0.655	0.054	1.802
Contrast															
MCD	0.609	0.062	1.868	0.505	0.070	2.213	0.354	0.076	2.744	0.170	0.078	3.447	0.047	0.079	4.304
MCDB	0.593	0.103	1.727	0.490	0.117	2.052	0.342	0.141	2.534	0.161	0.168	3.204	0.043	0.200	3.992
MCSD	0.600	0.053	2.164	0.493	0.066	2.553	0.340	0.084	3.129	0.158	0.096	3.886	0.041	0.096	4.773
SWAG	0.631	0.063	1.706	0.528	0.063	2.036	0.375	0.063	2.553	0.187	0.051	3.266	0.054	0.024	4.113
Deep Ensemble	0.641	0.048	1.706	0.539	0.047	2.036	0.387	0.046	2.541	0.198	0.038	3.220	0.060	0.021	4.093
Det. SAM	0.640	0.047	1.934	0.538	0.051	2.276	0.386	0.056	2.825	0.197	0.055	3.567	0.059	0.037	4.475
Det. ASAM	0.639	0.058	1.826	0.537	0.060	2.158	0.385	0.062	2.686	0.196	0.057	3.424	0.059	0.033	4.314
Elastic Transform															
MCD	0.697	0.055	1.603	0.639	0.061	1.790	0.584	0.065	1.954	0.531	0.068	2.112	0.449	0.073	2.378
MCDB	0.681	0.089	1.469	0.623	0.098	1.636	0.569	0.104	1.785	0.516	0.111	1.934	0.435	0.125	2.185
MCSD	0.691	0.039	1.848	0.631	0.051	2.083	0.575	0.057	2.261	0.521	0.062	2.441	0.439	0.070	2.735
SWAG	0.713	0.064	1.455	0.658	0.062	1.614	0.605	0.062	1.751	0.554	0.061	1.890	0.472	0.061	2.126
Deep Ensemble	0.720	0.048	1.455	0.667	0.047	1.603	0.615	0.046	1.748	0.565	0.045	1.884	0.485	0.045	2.115
Det. SAM	0.719	0.043	1.654	0.666	0.045	1.826	0.614	0.048	1.989	0.563	0.049	2.148	0.484	0.052	2.410
Det. ASAM	0.719	0.054	1.564	0.666	0.055	1.727	0.614	0.057	1.884	0.563	0.058	2.029	0.483	0.061	2.276
Pixelate															
MCD	0.741	0.051	1.479	0.728	0.051	1.512	0.686	0.054	1.644	0.627	0.059	1.821	0.578	0.061	1.989
MCDB	0.725	0.084	1.353	0.712	0.085	1.378	0.670	0.090	1.503	0.612	0.100	1.666	0.564	0.106	1.826
MCSD	0.736	0.034	1.706	0.722	0.035	1.748	0.679	0.039	1.905	0.619	0.047	2.126	0.569	0.052	2.327
SWAG	0.757	0.062	1.332	0.744	0.062	1.362	0.703	0.061	1.485	0.648	0.062	1.644	0.601	0.062	1.796
Deep Ensemble	0.764	0.046	1.324	0.752	0.046	1.353	0.711	0.046	1.479	0.658	0.046	1.636	0.613	0.047	1.785
Det. SAM	0.763	0.039	1.512	0.751	0.039	1.551	0.710	0.042	1.706	0.657	0.045	1.896	0.612	0.048	2.073
Det. ASAM	0.762	0.050	1.428	0.751	0.050	1.458	0.710	0.052	1.603	0.656	0.055	1.785	0.612	0.057	1.942
Jpeg Compression															
MCD	0.706	0.054	1.579	0.655	0.060	1.734	0.627	0.063	1.826	0.589	0.066	1.954	0.528	0.071	2.179
MCDB	0.690	0.088	1.447	0.639	0.097	1.588	0.612	0.103	1.670	0.573	0.108	1.790	0.514	0.119	2.007
MCSD	0.700	0.038	1.821	0.648	0.046	2.015	0.619	0.050	2.138	0.581	0.054	2.298	0.519	0.062	2.565
SWAG	0.722	0.063	1.432	0.674	0.062	1.564	0.647	0.061	1.644	0.610	0.061	1.751	0.551	0.062	1.961
Deep Ensemble	0.729	0.048	1.428	0.683	0.047	1.551	0.657	0.047	1.636	0.620	0.046	1.738	0.562	0.046	1.942
Det. SAM	0.728	0.042	1.625	0.682	0.046	1.777	0.655	0.048	1.874	0.619	0.049	1.990	0.561	0.052	2.228
Det. ASAM	0.728	0.053	1.536	0.682	0.056	1.674	0.655	0.057	1.765	0.619	0.059	1.879	0.560	0.061	2.099

A.4 Out-of-Distribution Evaluation Results

A.4.1 ImageNet-C Full 15-Corruption Evaluation

To provide transparency regarding the robustness claims made in Section 4.2.2, we report the complete evaluation matrices for ImageNet-C across all 15 corruption types and 5 severity levels.

Table 7 (ResNet-50) and Table 8 (ViT-B/16) evaluate Top-1 Accuracy, ECE, and Mean Predictive Entropy. A robust uncertainty estimator should exhibit a graceful degradation in accuracy paired with a monotonic scaling in predictive entropy as corruption severity increases. Across high-frequency noise corruptions (Gaussian, Shot, and Impulse Noise), standard Monte Carlo methods frequently experience sudden accuracy degradation at Severity 4 and 5. In contrast, the optimally flat optimization basins natively captured by deterministic SAM (and particularly the scale-aware ASAM formulation) smoothly scale predictive entropy. Without requiring stochastic weight perturbation, these deterministic checkpoints retain higher task performance under extreme visual stress, successfully matching or outperforming the $5\times$ Deep Ensemble.

Table 9: Full OOD Anomaly Detection Evaluation on ImageNet-O.

	ResNet-50				ViT-B/16			
Method	**ECE (↓)**	**NLL (↑)**	**Brier (↓)**	**Entropy (↑)**	**ECE (↓)**	**NLL (↑)**	**Brier (↓)**	**Entropy (↑)**
MCD	0.778	14.052	1.671	1.369	0.765	12.923	1.671	1.451
MCDB	0.826	14.143	1.738	1.104	0.793	**13.162**	1.710	1.286
MCSD	**0.771**	13.749	**1.656**	**1.502**	0.680	12.441	1.549	2.041
SWAG	0.840	12.373	1.762	1.035	0.782	11.577	1.698	1.500
Deep Ensemble	0.822	14.457	1.739	1.038	0.763	10.670	1.666	1.832
Det. SAM	0.827	14.079	1.738	1.179	**0.679**	10.053	**1.539**	**2.642**
Det. ASAM	0.830	**14.657**	1.747	1.056	0.760	10.616	1.662	1.839

Table 10: Full COCO-O Domain Shift Evaluation for YOLOv8x. We evaluate bounding-box regression stability across six visual domain shifts using mAP50-95 (↑), Expected Calibration Error (↓), and spatial Bounding-Box Standard Deviation (↓).

	Weather			Painting			Handmake		
Method	mAP (↑)	ECE (↓)	BBox (↓)	mAP (↑)	ECE (↓)	BBox (↓)	mAP (↑)	ECE (↓)	BBox (↓)
MCD	0.425	0.125	3.48	0.432	0.080	3.57	0.281	0.200	4.57
MCDB	0.427	0.108	2.19	0.413	0.064	2.43	0.265	0.203	3.98
MCSD	0.423	0.106	3.13	0.436	0.062	3.21	0.273	0.204	3.72
SWAG	0.407	0.060	7.04	0.431	0.134	6.88	0.296	0.310	7.17
Deep Ensemble	0.438	0.076	6.15	0.441	0.015	6.19	0.293	0.239	8.67
Det. SAM	0.185	0.259	9.15	0.173	0.226	11.59	0.112	0.052	16.72
Det. ASAM	0.455	0.119	7.61	0.436	0.047	8.10	0.291	0.173	10.58
	Cartoon			**Sketch**			**Tattoo**		
Method	mAP (↑)	ECE (↓)	BBox (↓)	mAP (↑)	ECE (↓)	BBox (↓)	mAP (↑)	ECE (↓)	BBox (↓)
MCD	0.241	0.038	3.85	0.217	0.059	4.20	0.165	0.469	8.15
MCDB	0.229	0.053	2.83	0.195	0.080	4.94	0.155	0.317	21.14
MCSD	0.239	0.046	3.59	0.211	0.073	3.51	0.166	0.436	8.00
SWAG	0.307	0.162	7.15	0.245	0.235	8.37	0.267	0.511	22.91
Deep Ensemble	0.247	0.051	6.41	0.226	0.075	7.11	0.174	0.549	19.34
Det. SAM	0.141	0.181	8.76	0.087	0.133	11.75	0.098	0.124	28.90
Det. ASAM	0.258	0.029	7.85	0.241	0.048	9.20	0.172	0.485	27.10

A.4.2 IMAGENET-O OOD ANOMALY EVALUATION

ImageNet-O is a strict anomaly detection benchmark consisting of images from classes that do not exist in the ImageNet-1k training set. The ideal behavior for a robust model on this dataset is, that whilst predictive accuracy falls, uncertainty metrics (such as predictive entropy and NLL) spike, indicating that the network successfully recognizes the semantic shift. As task accuracy and AUARC collapse to 0.0 across all evaluated methods on this dataset, they are omitted to be shown. Instead, Table 9 details a comprehensive suite of calibration metrics: ECE, NLL, Brier Score, and Mean Predictive Entropy.

Deterministic flat minima natively demonstrate strong anomaly rejection, particularly within the Vision Transformer (ViT-B/16) architecture. A single deterministic SAM forward pass on ViT-B/16 achieves the highest predictive entropy (2.642), lowest ECE (0.679) and Brier Score (1.539), thereby also outperforming the multi-modal Deep Ensemble (1.832 Entropy). While Monte Carlo Stochastic Depth (MCSD) produces the highest raw entropy on ResNet-50 (1.502), this can be seen as a structural artifact caused by dropping entire convolutional layers, thus artificially distorting feature activation graphs to force logit spread. Deterministic SAM avoids such destructive intervention. By utilizing the smooth local geometry of the optimum, it naturally yields scaled entropy spikes on OOD anomalies. This confirms that the anomaly rejection capabilities historically sought through local sampling are in SAM actually intrinsic geometric properties of the flat minimum itself.

A.4.3 COCO-O Domain Shift & Spatial Bounding-Box Uncertainty

To thoroughly evaluate predictive robustness and bounding-box spatial stability under visual domain shift, we benchmark YOLOv8x across the full COCO-O dataset. This dataset translates the standard COCO validation set into six distinct visual domains (Weather, Painting, Handmake, Cartoon, Sketch, and Tattoo). Unlike standard classification, evaluating dense object detection under domain shift requires measuring both predictive confidence (ECE) as well as the stability of the regression head (spatial Bounding-Box Standard Deviation).

As shown in Table 10, the uniform optimal transport boundary (standard SAM) collapses across all domains as it does not account for multi-scale feature topography, resulting in enlarged bounding-box variance. In contrast, the scale-aware deterministic ASAM formulation successfully stabilizes the regression predictions. Utilizing only a single base checkpoint, ASAM can match or exceed the Deep Ensemble across severe stylistic shifts like Sketch and Cartoon, while providing structurally superior ECE. However, it is notable that SWAG explicitly yields superior mAP on highly stylized domains such as Cartoon and Tattoo.

Spatial Stability Tradeoff. Crucially, our evaluation also reveals a distinct limitation of weight-space uncertainty estimators. When measuring pure spatial bounding-box stability, stochastic activation sampling methods (such as MCD and MCSD) consistently outperform weight-perturbation and weight-shifting methods (Deep Ensembles, SWAG, and the modified optimization boundaries of det. ASAM) by larger margins. As methods like MCD only drop intermediate feature activations, they leave the base regression weights physically intact, yielding highly stable, tight coordinate predictions (e.g., 3.48 BBox StdDev for MCD on Weather). Conversely, altering the global weight graph (whether via Deep Ensemble random initializations, SWAG covariance sampling, or ASAM's geometry) inherently induces higher spatial coordinate jitter in sensitive detection heads (e.g., 7.85 for Det. ASAM on Weather), even when the method retains high overall detection rates (mAP). This demonstrates that while ASAM can optimally bound detection confidence and classification rates under domain shift, activation sampling remains the strictly superior paradigm if absolute spatial coordinate consistency is the primary deployment constraint.

Table 11: Scale Sweep Evaluation for ResNet-50 on ImageNet-C. Scale denotes the multiplier of the base ρ radius. Metrics are averaged across all 5 severities.

Method	Scale 10			Scale 50			Scale 100			Scale 200		
	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)
Zoom Blur												
SAM	0.403	0.039	4.136	0.402	0.034	4.257	0.387	0.043	4.669	0.164	0.075	5.453
ASAM	0.401	0.066	3.747	0.386	0.035	4.387	0.319	0.053	5.521	0.114	0.068	6.733
Gaussian Noise												
SAM	0.397	0.033	4.538	0.396	0.035	4.658	0.380	0.056	5.038	0.181	0.106	5.581
ASAM	0.412	0.045	4.022	0.405	0.038	4.426	0.358	0.081	5.310	0.157	0.088	6.577
Snow												
SAM	0.346	0.065	4.413	0.341	0.056	4.537	0.317	0.041	4.951	0.125	0.099	5.441
ASAM	0.348	0.095	4.064	0.338	0.060	4.620	0.287	0.058	5.626	0.120	0.079	6.718
Pixelate												
SAM	0.478	0.024	3.538	0.473	0.020	3.668	0.451	0.035	4.112	0.241	0.083	5.335
ASAM	0.527	0.065	2.632	0.521	0.023	3.204	0.467	0.070	4.450	0.223	0.068	6.331
Brightness												
SAM	0.694	0.029	2.120	0.691	0.037	2.222	0.675	0.072	2.612	0.458	0.144	4.405
ASAM	0.689	0.020	1.856	0.684	0.029	2.278	0.640	0.111	3.467	0.387	0.134	5.719
Frost												
SAM	0.421	0.039	4.078	0.418	0.032	4.226	0.400	0.042	4.712	0.175	0.087	5.553
ASAM	0.421	0.067	3.664	0.417	0.032	4.317	0.365	0.070	5.498	0.152	0.069	6.829
Impulse Noise												
SAM	0.356	0.032	4.934	0.354	0.036	5.071	0.332	0.060	5.517	0.118	0.093	5.826
ASAM	0.371	0.044	4.370	0.363	0.042	4.796	0.309	0.076	5.693	0.122	0.081	6.728
Glass Blur												
SAM	0.276	0.077	5.011	0.277	0.070	5.118	0.269	0.065	5.461	0.103	0.110	5.704

Method	Scale 10			Scale 50			Scale 100			Scale 200		
	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)
ASAM	0.299	0.094	4.508	0.299	0.064	5.030	0.267	0.065	5.966	0.098	0.075	6.958
Fog												
SAM	0.498	0.032	3.446	0.496	0.035	3.615	0.479	0.069	4.212	0.211	0.085	5.557
ASAM	0.490	0.054	3.081	0.463	0.045	4.150	0.340	0.098	5.962	0.066	0.072	7.255
Defocus Blur												
SAM	0.375	0.033	5.021	0.373	0.040	5.116	0.353	0.062	5.448	0.126	0.089	5.825
ASAM	0.373	0.027	4.649	0.353	0.031	5.173	0.277	0.060	6.128	0.085	0.073	6.994
Contrast												
SAM	0.489	0.042	4.194	0.485	0.054	4.334	0.461	0.089	4.846	0.193	0.101	5.722
ASAM	0.433	0.030	4.091	0.389	0.052	5.079	0.266	0.105	6.445	0.055	0.078	7.355
Jpeg Compression												
SAM	0.559	0.021	3.042	0.555	0.026	3.163	0.534	0.057	3.611	0.295	0.088	5.092
ASAM	0.575	0.035	2.522	0.571	0.013	2.957	0.527	0.074	4.050	0.289	0.085	6.113
Elastic Transform												
SAM	0.449	0.068	3.466	0.447	0.062	3.592	0.432	0.070	4.041	0.219	0.100	5.155
ASAM	0.466	0.085	3.015	0.467	0.048	3.642	0.425	0.087	4.914	0.196	0.073	6.541
Motion Blur												
SAM	0.406	0.043	4.295	0.403	0.046	4.419	0.383	0.062	4.821	0.165	0.104	5.516
ASAM	0.397	0.061	3.900	0.387	0.047	4.501	0.321	0.078	5.615	0.113	0.090	6.762
Shot Noise												
SAM	0.383	0.036	4.623	0.381	0.038	4.749	0.364	0.057	5.148	0.164	0.102	5.595
ASAM	0.397	0.049	4.118	0.388	0.042	4.545	0.339	0.083	5.470	0.140	0.086	6.641

Table 12: Scale Sweep Evaluation for ViT-B/16 on ImageNet-C. Scale denotes the multiplier of the base ρ radius. The Transformer architecture structurally degrades at much lower noise thresholds compared to convolutional networks.

Method	Scale 10			Scale 50			Scale 100			Scale 200		
	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)
Zoom Blur												
SAM	0.464	0.027	4.465	0.355	0.256	8.597	0.085	0.078	9.813	0.001	0.001	9.900
ASAM	0.474	0.092	3.405	0.475	0.078	5.654	0.468	0.192	7.069	0.389	0.290	8.938
Gaussian Noise												
SAM	0.564	0.109	4.600	0.429	0.321	8.565	0.075	0.069	9.832	0.001	0.001	9.903
ASAM	0.575	0.042	2.940	0.574	0.115	5.043	0.564	0.241	6.649	0.481	0.342	8.584
Snow												
SAM	0.509	0.052	4.489	0.330	0.255	8.900	0.030	0.025	9.844	0.001	0.002	9.898
ASAM	0.512	0.084	3.139	0.516	0.090	5.400	0.510	0.207	6.847	0.434	0.311	8.742
Pixelate												
SAM	0.621	0.036	3.224	0.531	0.367	7.992	0.154	0.146	9.808	0.002	0.001	9.900
ASAM	0.628	0.078	2.119	0.629	0.101	4.447	0.624	0.241	6.146	0.560	0.396	8.394
Brightness												
SAM	0.738	0.063	2.503	0.620	0.432	7.763	0.170	0.162	9.796	0.001	0.002	9.896
ASAM	0.739	0.050	1.448	0.739	0.158	4.111	0.735	0.305	5.834	0.679	0.479	8.126
Frost												
SAM	0.590	0.086	4.115	0.440	0.335	8.597	0.075	0.069	9.834	0.001	0.002	9.895
ASAM	0.589	0.040	2.910	0.594	0.146	5.264	0.592	0.265	6.670	0.529	0.385	8.567
Impulse Noise												
SAM	0.553	0.116	4.770	0.418	0.319	8.659	0.065	0.059	9.836	0.001	0.002	9.903
ASAM	0.561	0.040	3.090	0.562	0.110	5.073	0.552	0.232	6.650	0.460	0.334	8.699
Glass Blur												
SAM	0.437	0.034	4.886	0.353	0.254	8.586	0.079	0.073	9.815	0.001	0.001	9.900
ASAM	0.444	0.098	3.596	0.447	0.062	5.669	0.442	0.169	7.055	0.373	0.268	8.865

Method	Scale 10			Scale 50			Scale 100			Scale 200		
	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)	Acc (↑)	ECE (↓)	Ent (↑)
Fog												
SAM	0.631	0.114	4.091	0.517	0.386	8.398	0.168	0.160	9.821	0.001	0.001	9.902
ASAM	0.613	0.073	3.533	0.618	0.205	5.606	0.613	0.321	6.995	0.520	0.409	8.868
Defocus Blur												
SAM	0.480	0.080	5.031	0.357	0.266	8.638	0.067	0.061	9.808	0.002	0.001	9.900
ASAM	0.496	0.050	3.575	0.493	0.112	5.762	0.477	0.219	7.200	0.374	0.284	8.980
Contrast												
SAM	0.596	0.126	4.602	0.475	0.342	8.355	0.198	0.188	9.789	0.003	0.001	9.895
ASAM	0.585	0.082	3.785	0.582	0.199	5.943	0.561	0.295	7.284	0.449	0.349	8.970
Jpeg Compression												
SAM	0.639	0.049	3.187	0.561	0.392	7.914	0.168	0.159	9.796	0.002	0.001	9.901
ASAM	0.650	0.070	1.978	0.650	0.110	4.330	0.646	0.258	6.105	0.595	0.419	8.282
Elastic Transform												
SAM	0.606	0.046	3.340	0.503	0.351	8.089	0.146	0.138	9.793	0.002	0.001	9.900
ASAM	0.599	0.079	2.396	0.600	0.100	4.701	0.593	0.235	6.384	0.525	0.374	8.515
Motion Blur												
SAM	0.531	0.064	4.355	0.413	0.302	8.445	0.094	0.087	9.811	0.002	0.001	9.900
ASAM	0.532	0.067	3.075	0.533	0.103	5.319	0.526	0.221	6.790	0.440	0.325	8.785
Shot Noise												
SAM	0.548	0.109	4.758	0.402	0.302	8.648	0.060	0.055	9.836	0.002	0.001	9.902
ASAM	0.557	0.048	3.037	0.557	0.108	5.115	0.544	0.225	6.654	0.453	0.321	8.639

Table 13: Scale Sweep Evaluation for YOLOv8x across COCO-O domain shifts. Scale denotes the multiplier of the base ρ radius.

Method	Scale 10			Scale 50			Scale 100			Scale 200			Scale 500		
	mAP ↑	ECE ↓	Ent ↑	mAP ↑	ECE ↓	Ent ↑	mAP ↑	ECE ↓	Ent ↑	mAP ↑	ECE ↓	Ent ↑	mAP ↑	ECE ↓	Ent ↑
COCO Weather															
SAM	0.185	0.263	0.013	0.185	0.262	0.013	0.185	0.262	0.013	0.185	0.261	0.013	0.187	0.261	0.013
ASAM	0.455	0.118	0.011	0.451	0.116	0.011	0.452	0.114	0.011	0.452	0.112	0.011	0.451	0.094	0.011
COCO Sketch															
SAM	0.086	0.132	0.013	0.086	0.132	0.013	0.086	0.132	0.013	0.086	0.132	0.013	0.087	0.125	0.013
ASAM	0.241	0.048	0.011	0.242	0.052	0.011	0.243	0.051	0.011	0.242	0.043	0.011	0.242	0.040	0.011
COCO Painting															
SAM	0.173	0.229	0.013	0.173	0.228	0.013	0.173	0.229	0.013	0.173	0.226	0.013	0.173	0.228	0.013
ASAM	0.436	0.047	0.011	0.442	0.047	0.011	0.441	0.041	0.011	0.447	0.039	0.011	0.447	0.034	0.011
COCO Handmake															
SAM	0.113	0.061	0.014	0.113	0.064	0.014	0.113	0.064	0.014	0.113	0.063	0.014	0.113	0.062	0.014
ASAM	0.291	0.174	0.011	0.291	0.174	0.011	0.291	0.178	0.011	0.292	0.181	0.011	0.292	0.190	0.011
COCO Cartoon															
SAM	0.140	0.176	0.013	0.140	0.176	0.013	0.140	0.176	0.013	0.140	0.175	0.013	0.140	0.175	0.013
ASAM	0.258	0.027	0.011	0.258	0.024	0.011	0.258	0.024	0.011	0.259	0.029	0.011	0.261	0.030	0.011
COCO Tattoo															
SAM	0.099	0.109	0.013	0.099	0.109	0.013	0.099	0.109	0.013	0.099	0.109	0.013	0.099	0.111	0.013
ASAM	0.172	0.484	0.012	0.172	0.485	0.013	0.171	0.484	0.013	0.172	0.491	0.013	0.172	0.501	0.013

A.5 Empirical Validation of the Variance Dilemma (Full Scale Sweeps)

Tables 11 (ResNet-50), 12 (ViT-B/16), and 13 (YOLOv8x) detail the complete diagnostic scale sweeps across in-distribution and domain-shift benchmarks. By artificially inflating the isotropic noise multiplier (k) well beyond the theoretically safe Wasserstein radius, we empirically trace the phase transition between Dimensionality Dilution and Functional Disruption. This illustrates why

practical samplers, which must operate at macroscopic noise scales to yield diversity, inevitably encounter structural degradation.

At bounded macroscopic multipliers (e.g., $10\times$ to $50\times$), the stochastic noise is entirely absorbed by the high-dimensional null-space. Task accuracy and ECE remain virtually static compared to the deterministic baseline, confirming that mathematically bounded sampling is functionally invisible. When forced to uncertified extremes to induce measurable logit variance (e.g., $100\times$ to $200\times$ for classifiers, or $> 500\times$ for the YOLOv8x backbone), the variance breaches the flat minimum. This degrades the internal feature manifolds, causing accuracy to fall and calibration to severely degrade.