BEYOND PREDICTIVE CRITERIA: OPTIMIZATION AND DOWNSTREAM UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimization choices are typically guided by predictive criteria such as accuracy, generalization, and calibration, yet many downstream uncertainty estimators operate on learned representations. We ask how uncertainty readouts respond to representations produced by predictively favorable optimization changes. Using the perturbation radius of Sharpness-Aware Minimization (SAM) as a controlled training path, we find a selective divergence: accuracy and calibration improve, output-based OOD scores remain stable or improve, and kNN improves, whereas covariance-sensitive Mahalanobis scoring and Deep Deterministic Uncertainty (DDU) can deteriorate at large SAM radii. We investigate this degradation using Mahalanobis scoring as a tractable case. Within each checkpoint, we vary the strength of covariance weighting while holding the learned representation, class centroids, and covariance eigensystem fixed. Under large-radius SAM, stronger inverse-covariance weighting can become substantially detrimental to OOD detection. Directional analysis links this behavior to a covariance–contrast mismatch in which later low-variance eigendirections receive increasing emphasis despite providing weaker ID–OOD contrast. Crucially, an unweighted readout over the same representation retains strong OOD discrimination as these directions are added, showing that the degradation does not require a uniform loss of OOD-discriminative structure. This covariance-sensitive degradation under large-radius SAM is not universal across sharpness-aware optimizers or ID training distributions, while the broader separation between predictive improvement and representation-dependent uncertainty also appears under SWA and in a ViT/ImageNet-200 setting. Together, these results show that uncertainty readouts can respond differently across representations produced under different optimization conditions and, in the Mahalanobis case, that detector degradation can arise from detrimental inverse-covariance weighting even when useful OOD-discriminative structure remains accessible.

1 INTRODUCTION

Optimization choices are typically guided by predictive criteria such as classification accuracy, generalization, and calibration. Yet many downstream uncertainty estimators operate not on predictive outputs alone, but on distances, neighborhoods, densities, or covariance structure in the learned representation space (Lee et al., 2018; Mukhoti et al., 2023; Sun et al., 2022; Wang et al., 2022). Changing the optimization procedure can therefore affect both predictive behavior and the learned representation on which downstream uncertainty is evaluated.

This raises a basic question: when an optimization change is favorable according to predictive criteria, how do uncertainty readouts respond to the learned representation produced under that optimization change? Prior work shows that optimization can reshape learned representations (Andriushchenko et al., 2023; Zhao et al., 2026; Tan et al., 2026), feature-based OOD detection depends strongly on representation geometry (Ren et al., 2021; Müller & Hein, 2025; Janiak et al., 2026), and detector performance can interact with training strategy (Krumpl et al., 2026). However, it remains unclear how uncertainty readouts respond to representations produced by predictively favorable optimization changes and, when some readouts deteriorate, what that degradation reflects.

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.

We study this question using Sharpness-Aware Minimization (SAM) (Foret et al., 2021), whose perturbation radius ρ provides a controlled training path while architecture, dataset, and the remaining training protocol are held fixed. Along this path, accuracy improves and calibration error decreases, but downstream uncertainty readouts respond selectively. MSP and Energy remain stable or improve, and representation-based kNN also improves, whereas covariance-sensitive Mahalanobis scoring and Deep Deterministic Uncertainty (DDU) deteriorate at large SAM radii. The opposite responses of kNN and Mahalanobis are particularly informative: both operate on the learned representation, yet move in different directions, showing that a single scalar notion of representation quality is insufficient to explain their behavior.

This divergence raises a more specific explanatory question: does the degradation of a covariance-sensitive readout reflect a loss of useful OOD-discriminative structure, or does training change the effect of the readout applied to structure that remains available? We examine this question using Mahalanobis distance as a tractable case. Within each checkpoint, we vary the strength of inverse-covariance weighting while holding the learned representation, class centroids, covariance eigenvectors, and eigenvalues fixed. Under SGD and smaller SAM radii, covariance weighting has only a modest effect; under large-radius SAM, stronger inverse-covariance weighting becomes substantially detrimental. Because no feature learning occurs during this intervention, the analysis isolates a training-dependent change in the effect of covariance weighting itself.

Directional analysis links this behavior to a covariance–contrast mismatch. Under large-radius SAM, later low-variance eigendirections receive increasing emphasis from inverse-covariance weighting despite providing weaker ID–OOD contrast. Yet an unweighted readout over the same representation retains strong OOD discrimination as these directions are added, showing that the degradation cannot be explained by a uniform loss of OOD-discriminative structure alone.

We further examine the scope and boundaries of this interpretation. DDU exhibits a related covariance-sensitive pattern. SWA and a ViT/ImageNet-200 setting extend the broader separation between predictive improvement and representation-dependent uncertainty, whereas ASAM, GSAM, and a CIFAR-100-ID control show that the large-radius covariance-sensitive failure is neither generic to sharpness-aware optimization nor universal across training settings.

Our contributions are:

- **Readout-selective responses under predictively favorable optimization changes.** We show that accuracy and calibration can improve while downstream uncertainty readouts move in different directions, including opposite responses among readouts operating on the learned representation.
- **Training-dependent changes in covariance-weighting effectiveness.** A within-checkpoint intervention shows that inverse-covariance weighting can become substantially detrimental under large-radius SAM even when strong OOD-discriminative structure remains accessible in the representation.
- **Mechanistic evidence with boundary tests.** Directional analysis links the degradation to a covariance–contrast mismatch, while unweighted readouts and alternative training and dataset controls test competing explanations and delimit its scope.

2 Related Work

2.1 Optimization and Learned Representations

Optimization affects not only predictive performance but also the geometry of learned representations. SAM was introduced to improve generalization by optimizing loss in a neighborhood of the model parameters (Foret et al., 2021), and subsequent studies have shown that sharpness-aware training can alter feature rank and calibration behavior (Andriushchenko et al., 2023; Tan et al., 2026). More broadly, optimizer choice can affect representation geometry such as neural-collapse structure (Zhao et al., 2026). These studies characterize how training reshapes representations, but do not directly examine how such changes affect downstream uncertainty readouts. We study these responses jointly along the same controlled training path.

2.2 Representation Geometry and Covariance-Based OOD Detection

Representation-based OOD detectors exploit different properties of learned features: Mahalanobis scoring uses class means and covariance (Lee et al., 2018), DDU fits class-conditional Gaussian densities (Mukhoti et al., 2023), kNN relies on neighborhood structure (Sun et al., 2022), and ViM combines feature-space residual information with logits (Wang et al., 2022).

Prior work has shown that Mahalanobis OOD performance depends strongly on representation geometry, including feature normalization, covariance estimation, inverse-variance weighting, and directional ID–OOD deviations (Ren et al., 2021; Müller & Hein, 2025; Janiak et al., 2026). Li et al. (2025) further show that modifying covariance use within a fixed representation can improve OOD discrimination. These works characterize how Mahalanobis behavior depends on representation geometry. We instead isolate a training-conditioned change in covariance-weighting effectiveness by varying covariance weighting within each checkpoint along a controlled optimization trajectory.

2.3 Training–Detector Interaction

Krumpl et al. (2026) show that OOD performance depends jointly on training strategy and detector choice, including changes in detector rankings across trained models. This establishes the existence of training–detector interactions, but does not isolate why a particular representation-dependent readout changes. We address this gap by holding each learned representation fixed, varying the Mahalanobis readout, and comparing the resulting weighting effect across training conditions.

3 Divergent Downstream Uncertainty under Predictively Favorable Training

3.1 Experimental Setup

We conduct our primary experiments on CIFAR-10 using WideResNet-28-10, following the DDU-style setting with spectral normalization (Krizhevsky, 2009; Zagoruyko & Komodakis, 2016; Miyato et al., 2018; Mukhoti et al., 2023). Unless otherwise specified, OOD results in the main text refer to CIFAR-100. In Section 3.2, we additionally evaluate Tiny-ImageNet (TIN), SVHN, and MNIST, with CIFAR-100 and TIN treated as near-OOD datasets and SVHN and MNIST as far-OOD datasets, following the near/far grouping used by OpenOOD v1.5 (Zhang et al., 2024).

Models are trained for 350 epochs using either SGD or SAM with

$$\rho \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\},$$

where $\rho = 0$ denotes the SGD baseline. All remaining training settings are kept fixed across the ρ sweep. Results are averaged over three independent runs, and tabulated values are reported as mean $\pm$ standard deviation.

We evaluate predictive behavior using classification accuracy, Expected Calibration Error (ECE), and temperature-scaled ECE (T-ECE) (Guo et al., 2017). For OOD detection, we report AUROC for MSP and Energy (Hendrycks & Gimpel, 2017; Liu et al., 2020), together with Mahalanobis, DDU, kNN, and ViM, covering output-level and representation-dependent readouts. Because our analysis concerns how training-induced representation changes relate to these downstream responses, we additionally report WithinVar, Neural Collapse descriptors $\mathrm{NC1}$ and $\mathrm{NC2}$ (Papyan et al., 2020), and $\mathrm{NC2}^{\mathrm{A}}$. Additional metric definitions, training details, and implementation details are provided in the appendix.

3.2 Divergence across Uncertainty Readouts

Table 1 establishes a predictively favorable training trajectory along the SAM radius sweep. Accuracy improves relative to SGD, while ECE decreases substantially and T-ECE remains below the SGD baseline throughout the sweep.

The downstream uncertainty readouts do not follow a common response. Figure 1 shows that MSP and Energy generally improve as ρ increases. Readouts that use the learned representation

are also heterogeneous: kNN improves consistently across the OOD datasets, ViM shows a mixed response, whereas Mahalanobis and DDU both deteriorate at large SAM radii.

This selective divergence does not by itself identify a mechanism, since different readouts may naturally respond differently when training changes the learned representation. It instead motivates asking what the degradation of a particular representation-dependent readout signifies.

Table 1: Predictive performance across ρ on CIFAR-10 with WideResNet-28-10.

ρ	Accuracy ↑	ECE ↓	T-ECE ↓
0 (SGD)	0.9597±0.0023	0.0241±0.0029	0.0088±0.0027
0.01	0.9622±0.0009	0.0189±0.0010	0.0063±0.0022
0.02	0.9618±0.0020	0.0194±0.0013	0.0069±0.0027
0.05	0.9666±0.0002	0.0146±0.0003	0.0079±0.0004
0.1	0.9659±0.0005	0.0118±0.0016	0.0056±0.0008
0.2	0.9680±0.0018	0.0092±0.0016	0.0053±0.0008
0.5	0.9681±0.0012	0.0097±0.0010	0.0057±0.0020

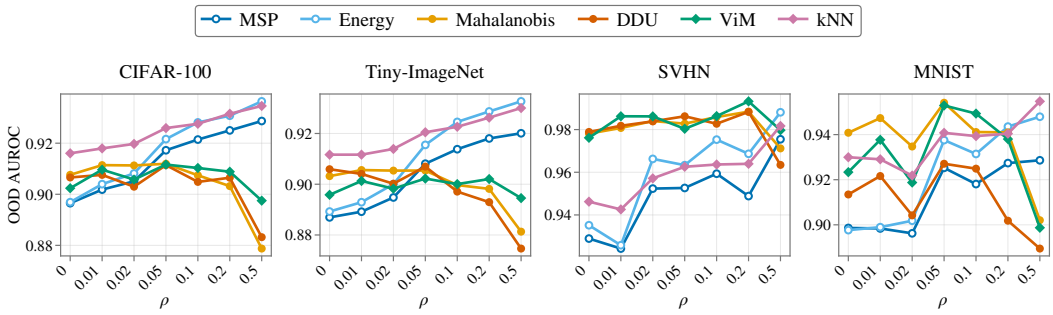

Figure 1: OOD AUROC across the SAM radius ρ for CIFAR-100, TIN, SVHN, and MNIST.

3.3 Representation Changes Do Not Imply a Common Readout Response

The learned representation also changes along the SAM radius sweep. As ρ increases, WithinVar, NC1, and NC2 increase, whereas $\mathrm{NC2^A}$ remains comparatively stable. These measurements indicate substantial changes in representation geometry while the angular organization of centered class means remains largely preserved.

These changes, however, do not correspond to a common trend in representation-dependent OOD performance. Table 2 shows that Mahalanobis AUROC decreases at large ρ, whereas kNN AUROC improves over the same training trajectory.

Table 2: Representation-level statistics and CIFAR-100 OOD readouts across ρ.

ρ	WithinVar	NC1	NC2	$\mathrm{NC2^A}$	Mahalanobis	kNN
0 (SGD)	0.8655±0.0244	0.0221±0.0009	0.0864±0.0028	-0.1110±0.0000	0.9076±0.0035	0.9160±0.0010
0.01	0.9261±0.0629	0.0240±0.0015	0.0972±0.0025	-0.1110±0.0000	0.9114±0.0029	0.9180±0.0017
0.02	1.0805±0.0862	0.0278±0.0016	0.1070±0.0024	-0.1110±0.0000	0.9113±0.0042	0.9197±0.0024
0.05	1.2640±0.1631	0.0342±0.0044	0.1385±0.0084	-0.1110±0.0000	0.9119±0.0030	0.9259±0.0014
0.10	1.7433±0.1295	0.0474±0.0050	0.1993±0.0084	-0.1109±0.0001	0.9074±0.0068	0.9276±0.0021
0.20	2.0108±0.0214	0.0549±0.0014	0.2769±0.0012	-0.1107±0.0000	0.9032±0.0012	0.9315±0.0004
0.50	2.4303±0.0142	0.0639±0.0004	0.3480±0.0025	-0.1106±0.0001	0.8787±0.0023	0.9346±0.0003

The opposite responses of Mahalanobis and kNN are particularly informative because both operate on the learned representation. Thus, the divergence cannot be reduced to a simple output-versus-representation distinction or to a single scalar notion of representation quality. Because Mahalanobis and DDU are the readouts that deteriorate at large ρ and both depend explicitly on learned covariance, we next use Mahalanobis as a tractable case to isolate the effect of covariance weighting.

4 When Does Covariance Weighting Become Detrimental for OOD Detection?

Section 3 shows that representation-dependent uncertainty readouts can respond differently along the same predictively favorable training trajectory. We now examine why covariance-sensitive Mahalanobis scoring deteriorates at large SAM radii. In particular, we ask whether this degradation reflects a loss of useful OOD-discriminative structure or a change in the effect of covariance weighting on structure that remains available. Mahalanobis distance provides a tractable setting for separating these possibilities because its score can be decomposed explicitly into directional residuals and inverse-covariance weights.

4.1 Intervening on Covariance Weighting

Let the tied within-class covariance of a checkpoint have eigendecomposition

$$\Sigma = U\Lambda U^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d),$$

with eigenvalues ordered as

$$\lambda_1 \geq \cdots \geq \lambda_d.$$

Mahalanobis distance weights the squared residual along eigendirection u_i by $1/\lambda_i$ (Lee et al., 2018). Related work has shown that modifying covariance use within a fixed representation, including interpolation toward the identity, can improve OOD discrimination (Li et al., 2025). To vary the strength of covariance weighting, we use the d_{eff} active covariance eigendirections retained by the same pseudoinverse threshold as the baseline Mahalanobis score and define

$$d_\alpha(z, c) = \sum_{i=1}^{d_{\text{eff}}} \frac{\left(u_i^\top (z - \mu_c)\right)^2}{\lambda_i^\alpha}, \qquad \alpha \in [0, 1]. \tag{1}$$

Here, $\alpha = 0$ gives an unweighted class-mean distance over the same active eigensubspace, whereas $\alpha = 1$ recovers the baseline Mahalanobis weighting. Within each checkpoint, the learned representation, class centroids, covariance eigensystem, and active-subspace dimension are fixed while only α is varied. The resulting α-sweep therefore isolates how changing covariance weighting affects OOD discrimination on the same learned representation.

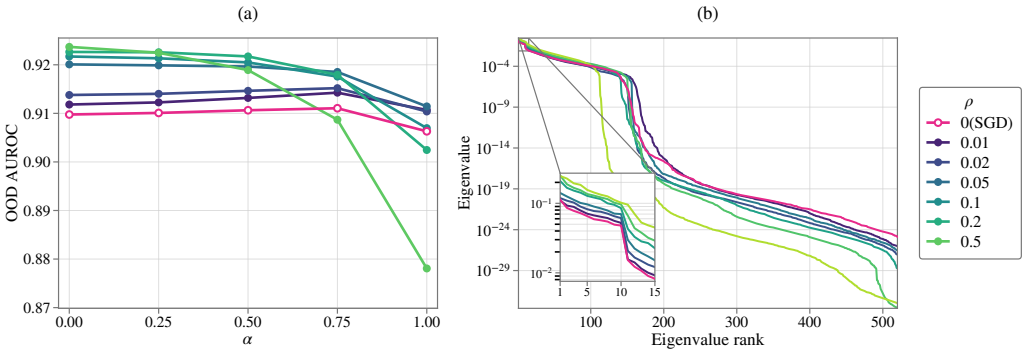

Figure 2: (a) CIFAR-100 OOD AUROC as α varies from the unweighted endpoint ($\alpha = 0$) to Mahalanobis weighting ($\alpha = 1$). (b) Full within-class covariance spectra, sorted independently within each checkpoint; Mahalanobis scoring uses only eigendirections retained by the FP64 pseudoinverse threshold (Appendix C.1).

Figure 2(a) shows that the effect of covariance weighting changes along the SAM radius trajectory. Under SGD and smaller SAM radii, OOD AUROC varies only modestly as α increases. At larger ρ, however, stronger covariance weighting becomes increasingly detrimental, with the largest reduction observed at $\rho = 0.5$. Thus, the large-radius degradation is not explained solely by the fact that SAM produces a different representation: on the representation produced by large-radius SAM, increasing inverse-covariance weighting itself reduces OOD discrimination.

4.2 Covariance Spectrum Alone Does Not Explain the Degradation

We next examine how the within-class covariance spectrum changes with ρ. Figure 2(b) shows that the spectrum is reshaped rather than uniformly scaled as the SAM radius increases: leading eigenvalues increase, whereas later-ranked eigenvalues decrease.

Eigenvalue ranks are defined independently within each checkpoint, so the same rank under SGD and SAM is not interpreted as the same feature direction. More importantly, the covariance spectrum alone does not determine whether emphasizing a direction is beneficial for OOD detection. A smaller eigenvalue gives that direction a larger Mahalanobis weight, but the usefulness of this weighting depends on the ID–OOD contrast carried by the direction. This motivates examining covariance weighting together with directional OOD separation.

4.3 Variance-Aware Directional OOD Separation and Cumulative OOD Discrimination

Following the per-direction Mahalanobis decomposition of Janiak et al. (2026), we examine ID–OOD separation along individual covariance eigendirections. For sample z, let

$$c_M(z) = \arg\min_c d_1(z, c)$$

and define

$$a_i(z) = u_i^\top \left(z - \mu_{c_M(z)}\right). \tag{2}$$

We measure the directional residual contrast and its pooled variability as

$$\Delta_i^{r^2} = \mathbb{E}_{z\sim\mathrm{OOD}}\left[a_i^2(z)\right] - \mathbb{E}_{z\sim\mathrm{ID}}\left[a_i^2(z)\right]. \tag{3}$$

$$V_i = \frac{1}{2}\left[\mathrm{Var}_{z\sim\mathrm{OOD}}\left(a_i^2(z)\right) + \mathrm{Var}_{z\sim\mathrm{ID}}\left(a_i^2(z)\right)\right], \tag{4}$$

and define the variance-normalized directional contrast

$$\widetilde{\Delta}_i^{r^2} = \frac{\Delta_i^{r^2}}{\sqrt{V_i}}. \tag{5}$$

Larger $\widetilde{\Delta}_i^{r^2}$ indicates stronger ID–OOD separation relative to directional variability. Figure 3(a) shows comparable separation under SGD and SAM with $\rho = 0.5$ in the leading eigendirections, but substantially weaker separation under large-radius SAM at later ranks.

Using marginal directional variability and omitting cross-direction covariance, consider two positive-contrast directions i and j satisfying

$$\widetilde{\Delta}_i^{r^2} < \widetilde{\Delta}_j^{r^2}, \qquad \lambda_i < \lambda_j.$$

The effective-weight ratio of direction i to direction j under d_α is

$$t_\alpha = \left(\frac{\lambda_j}{\lambda_i}\right)^\alpha \sqrt{\frac{V_i}{V_j}}.$$

Increasing the relative weight of the weaker-contrast direction reduces the corresponding pairwise variance-normalized separation when

$$t_\alpha > \frac{\widetilde{\Delta}_i^{r^2}}{\widetilde{\Delta}_j^{r^2}}. \tag{6}$$

We refer to this condition as a covariance–contrast mismatch: inverse-covariance weighting increasingly favors a lower-variance direction despite its weaker relative ID–OOD contrast. At $\alpha = 1$, the condition holds for 57.5% of eligible direction pairs under SGD and 85.1% under SAM with $\rho = 0.5$.

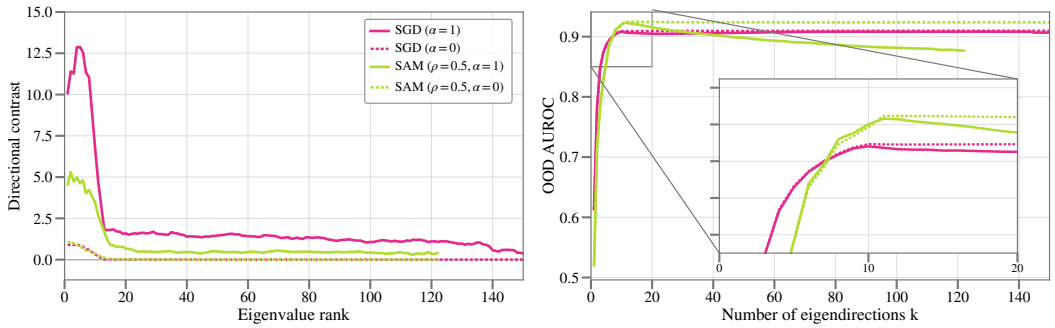

Figure 3: Directional separation and cumulative OOD discrimination with CIFAR-100 as OOD. (a) Variance-normalized directional residual contrast $\widetilde{\Delta}_i^{r^2}$ across eigenvalue rank under SGD and SAM ($\rho = 0.5$). For visualization, the directional curves are smoothed with a centered five-rank moving average. (b) OOD AUROC as eigendirections are progressively accumulated using the unweighted ($\alpha = 0$) and Mahalanobis-weighted ($\alpha = 1$) prefix scores. Solid and dotted lines denote $\alpha = 1$ and $\alpha = 0$, respectively.

We next examine whether useful OOD-discriminative structure remains accessible as eigendirections are accumulated. Define

$$d_{\alpha,1:k}(z,c) = \sum_{i=1}^{k} \frac{\left(u_i^\top (z - \mu_c)\right)^2}{\lambda_i^\alpha}, \qquad \alpha \in \{0,1\}, \quad k = 1, \ldots, d_{\text{eff}}, \tag{7}$$

where $\alpha = 0$ and $\alpha = 1$ denote the unweighted and Mahalanobis-weighted prefix scores, respectively.

Figure 3(b) shows comparable OOD discrimination from the leading eigendirections under SGD and large-radius SAM. As later-ranked directions are added, however, the Mahalanobis-weighted prefix under large-radius SAM deteriorates, whereas the unweighted prefix over the same representation and accumulated directions remains stronger. The same qualitative pattern is observed for TIN, SVHN, and MNIST, with weaker later-rank directional separation, higher covariance–contrast mismatch prevalence, and cumulative degradation of the covariance-weighted readout under large-radius SAM; the corresponding results are provided in Appendix G.1.

Motivated by Müller & Hein (2025), we use post-hoc L_2 normalization only as a diagnostic of radial feature scaling. After applying the same normalization to both SGD and SAM representations, the pronounced Mahalanobis degradation at large ρ is no longer observed across the four OOD datasets (Appendix F.1).

Taken together, these analyses show that the large-radius Mahalanobis degradation cannot be explained by a uniform loss of OOD-discriminative structure. Instead, inverse-covariance weighting increasingly emphasizes directions with comparatively weak ID–OOD contrast, while strong OOD discrimination remains accessible under an unweighted readout over the same representation.

5 Does the Covariance-Sensitive Pattern Extend to DDU?

Section 4 used Mahalanobis distance to isolate the effect of inverse-covariance weighting through its tied-covariance quadratic form. DDU is also explicitly covariance-sensitive, but its class-conditional Gaussian densities additionally involve class-specific covariance matrices, log-determinant terms, and aggregation across classes (Mukhoti et al., 2023). Given these structural differences, we ask a narrower question: does DDU exhibit a similar training-conditioned pattern in which OOD discrimination is preserved in leading covariance eigendirections but deteriorates as additional directions are incorporated?

To test this, we decompose each class-conditional DDU density in its covariance eigenbasis and progressively accumulate eigendirections from largest to smallest eigenvalue. For each prefix of k

eigendirections, we recompute the class-conditional density and aggregate across classes to obtain the corresponding spectral-prefix DDU score. Full definitions and numerical details are provided in Appendix D.

Figure 4 shows that the leading eigendirections provide comparable OOD discrimination under SGD and large-radius SAM. As additional directions are incorporated, however, DDU AUROC under large-radius SAM decreases and a gap relative to SGD emerges. Importantly, this gap appears before the numerically extreme covariance tail and persists over a substantial prefix range.

This cumulative behavior is qualitatively consistent with the Mahalanobis analysis: the large-radius degradation is not evident in the leading covariance eigendirections but emerges as additional covariance structure enters the readout. Together, these results show that the training-conditioned covariance-sensitive degradation is not confined to tied-covariance Mahalanobis scoring.

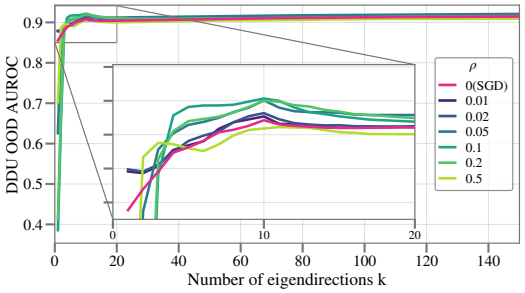

Figure 4: DDU OOD AUROC as class-wise covariance eigendirections are progressively accumulated across ρ.

6 Scope and Boundary Tests

We next examine the boundaries of the covariance-sensitive failure characterized in Section 4 and whether the broader training–readout divergence extends beyond the primary SAM–CIFAR-10 setting.

6.1 Boundary Tests

Under the same primary evaluation protocol, ASAM ($\rho = 0.5$) and GSAM ($\rho_{\max} = 0.5$) maintain competitive predictive performance and calibration but do not reproduce the pronounced Mahalanobis and DDU degradation observed under large-radius vanilla SAM ($\rho = 0.5$) (Kwon et al., 2021; Zhuang et al., 2022) (Appendix E.1.3).

Table 3: Predictive and CIFAR-100 OOD performance across sharpness-aware optimizers.

Training	Accuracy ↑	ECE ↓	T-ECE ↓	MSP ↑	Mahalanobis ↑	DDU ↑
SGD	0.9597±0.0023	0.0241±0.0029	0.0088±0.0027	0.8965±0.0022	0.9076±0.0035	0.9066±0.0024
SAM	0.9681±0.0012	0.0097±0.0010	0.0057±0.0020	0.9287±0.0014	0.8787±0.0023	0.8832±0.0058
ASAM	0.9627±0.0019	0.0197±0.0010	0.0072±0.0015	0.9099±0.0027	0.9124±0.0040	0.9114±0.0024
GSAM	0.9671±0.0009	0.0145±0.0005	0.0060±0.0023	0.9166±0.0023	0.9205±0.0017	0.9070±0.0018

When CIFAR-100 is used as the ID dataset, SAM does not reproduce the same covariance-sensitive degradation observed in the primary CIFAR-10 setting (Appendix E.3); covariance-sensitive readouts instead show mixed, OOD-dependent responses. Together, these controls show that the specific failure characterized in Section 4 is neither a generic consequence of sharpness-aware optimization nor a universal effect of large-radius SAM.

6.2 Broader Training–Readout Divergence

With Stochastic Weight Averaging (SWA), predictive performance and calibration improve relative to SGD, whereas Mahalanobis and DDU do not show a corresponding consistent improvement across OOD datasets (Izmailov et al., 2018) (Appendix E.2). This shows that predictive improvement and covariance-sensitive OOD performance can separate outside the SAM radius path.

A similar broader separation appears in a different architecture and dataset regime. We fine-tune an ImageNet-1K-pretrained ViT-S/16 on ImageNet-200 using AdamW and AdamW+SAM and evaluate OOD detection over iNaturalist (Van Horn et al., 2018), NINCO (Bitterwolf et al., 2023), OpenImage-O, and SSB-Hard (Vaze et al., 2022). Table 4 reports the average OOD performance across the four datasets.

Table 4: Predictive and OOD performance on ImageNet-200 with ViT-S/16. OOD metrics are averaged across the four OOD datasets.

ρ	Accuracy ↑	ECE ↓	MSP ↑	Mahalanobis ↑	DDU ↑
0 (AdamW)	$0.9249_{\pm 0.0009}$	$0.0226_{\pm 0.0001}$	$0.9219_{\pm 0.0022}$	$0.9532_{\pm 0.0004}$	$0.9318_{\pm 0.0015}$
0.05	$0.9294_{\pm 0.0006}$	$0.0147_{\pm 0.0003}$	$0.9227_{\pm 0.0031}$	$0.9519_{\pm 0.0021}$	$0.9285_{\pm 0.0012}$
0.5	$0.9392_{\pm 0.0003}$	$0.0104_{\pm 0.0005}$	$0.9313_{\pm 0.0005}$	$0.9512_{\pm 0.0029}$	$0.9187_{\pm 0.0034}$

As the SAM radius increases, accuracy improves and ECE decreases, while Mahalanobis AUROC remains approximately stable and DDU AUROC deteriorates. This extends the broader training–readout divergence to a different architecture, dataset regime, and base optimizer.

Taken together, these experiments distinguish the specific covariance-sensitive failure characterized in Section 4 from the broader training–readout divergence: the former is not universal across training and data settings, whereas the latter extends beyond the primary SAM–CIFAR-10 setting.

7 Discussion and Conclusion

We study how downstream uncertainty readouts respond to optimization changes that are favorable according to predictive criteria. Along a controlled SAM trajectory, accuracy and calibration improve, but uncertainty readouts respond selectively: output-based scores and kNN remain stable or improve, whereas covariance-sensitive Mahalanobis scoring and DDU can deteriorate at large SAM radii. Thus, predictive improvement does not imply a common response across downstream uncertainty readouts.

For Mahalanobis, our within-checkpoint intervention separates detector degradation from a uniform loss of useful OOD-discriminative structure. Under large-radius SAM, stronger inverse-covariance weighting can become detrimental, and directional analysis links this behavior to a covariance–contrast mismatch in which low-variance directions receive greater weight despite weaker ID–OOD contrast. Yet an unweighted readout over the same representation retains strong OOD discrimination. The observed degradation can therefore arise from how the available OOD-discriminative structure is weighted by the readout rather than from a uniform disappearance of that structure.

This covariance-sensitive behavior is setting-dependent. ASAM, GSAM, and the CIFAR-100-ID setting do not reproduce the same large-radius degradation, whereas SWA and ViT/ImageNet-200 extend the broader separation between predictive improvement and representation-dependent uncertainty beyond the primary SAM–CIFAR-10 setting. The effect of covariance weighting is also OOD-dependent, indicating that the mismatch reflects an interaction among the learned representation, the covariance-based readout, and the downstream OOD distribution.

This analysis raises the further question of how optimization dynamics shape the covariance spectrum and directional residual structure associated with this mismatch. Our normalization control provides one clue: feature magnitude is an important contributor, while its nonuniform effect across OOD datasets and detectors indicates that feature magnitude alone does not account for the full behavior. Taken together, our results show that downstream uncertainty depends not only on the structure produced by optimization but also on how that structure is read out. In the Mahalanobis case, useful OOD-discriminative structure can remain accessible even when inverse-covariance weighting becomes detrimental.

AI USE STATEMENT

Generative AI tools were used to provide feedback on research methodology and experimental design, assist in formulating and checking mathematical analyses, and support the interpretation of experimental results. They were also used to assist with literature search and summarization, identification and formatting of references, review of analysis code, and drafting and editing portions of the manuscript for clarity and structure. All AI-assisted material was reviewed by the authors; manuscript claims and citations were checked against the relevant source literature, and code-related suggestions and reported results were checked against the implementation and experimental outputs. The authors take full responsibility for the final content of the paper, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main experimental setup is described in Section 3.1. Appendix A provides the training configurations, hyperparameters, and dataset settings, while Appendix B specifies the OOD detection, calibration, representation-geometry, and neural-collapse evaluation protocols. Appendix C documents the Mahalanobis spectral analysis, including covariance estimation, numerical thresholding, covariance-weighting interventions, variance-aware directional analysis, and cumulative prefix evaluation. Appendix D provides the corresponding DDU spectral diagnostic and its numerical handling. Additional optimizer, dataset, backbone, feature-normalization, normality, compute, and batch-size controls are described in Appendices E and F. Appendix G reports the cross-OOD validation and extended quantitative results. Unless otherwise stated, the main reported comparisons are averaged over three independent runs, with means and standard deviations provided for the corresponding results.

REFERENCES

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing ImageNet out-of-distribution detection evaluation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2471–2506. PMLR, 2023.

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

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

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

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*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 876–885, 2018.

Denis Janiak, Jakub Binkowski, and Tomasz Kajdanowicz. A geometry-based view of mahalanobis OOD detection. In *International Conference on Machine Learning*, 2026.

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

Gerhard Krumpl, Henning Avenhaus, and Horst Possegger. One model, many behaviors: Training-induced effects on out-of-distribution detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4128–4138, 2026.

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*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914. PMLR, 2021.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 7167–7177, 2018.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised classifiers answer the wrong questions for OOD detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81689–81713. PMLR, 2025. Extended version available as arXiv:2507.01831.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023.

Maximilian Müller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45151–45184. PMLR, 2025.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-OOD detection. *arXiv preprint arXiv:2106.09022*, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840. PMLR, 2022.

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. In *International Conference on Learning Representations*, 2026.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations*, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, pp. 87.1–87.12, 2016. doi: 10.5244/C.30.87.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2(3):1–32, 2024.

Jim Zhao, Tin Sum Cheng, Wojciech Masarczyk, and Aurelien Lucchi. Optimizer choice matters for the emergence of neural collapse. In *International Conference on Learning Representations*, 2026.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C. Dvornek, Sekhar Tatikonda, James S. Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations*, 2022.

A Training Details and Hyperparameter Settings

This appendix provides the training configurations and hyperparameter settings used in our experiments. These implementation details are omitted from the main text for clarity and are reported here to facilitate reproducibility.

A.1 CIFAR-10 Training Setup

For the main CIFAR-10 experiments, we use WideResNet-28-10 with spectral normalization, following the DDU-style setting (Zagoruyko & Komodakis, 2016; Miyato et al., 2018; Mukhoti et al., 2023). Unless otherwise stated, the following settings are shared across the main experiments:

- **Optimizer:** SGD, SAM
- **Base learning rate:** 0.1
- **Momentum:** 0.9
- **Weight decay:** 5×10^{-4}
- **Batch size:** 128
- **Total epochs:** 350

A multistep learning rate scheduler is employed with decay milestones at epochs 150 and 250, where the learning rate is multiplied by 0.1 at each milestone.

A.2 Sharpness-Aware Minimization

For SAM-based optimization, we follow the standard formulation (Foret et al., 2021) and sweep the neighborhood radius

$$\rho \in \{0\ (\mathrm{SGD}), 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.6, 0.7\}.$$

Unless otherwise stated, all optimization hyperparameters other than the SAM radius are kept identical to the SGD baseline, and both SGD and SAM are trained for the same 350 epochs under the same training protocol. This comparison therefore matches the training protocol rather than the gradient-evaluation count; the latter is examined separately in Appendix F.3.

Additional ρ sweep analysis. While the main text focuses on $\rho \leq 0.5$, we additionally evaluate larger values $\rho \in \{0.6, 0.7\}$ for completeness. In these additional runs, the large-ρ covariance-sensitive degradation generally persists, although the detailed detector response is OOD-dependent and not strictly monotonic, while the representation-geometry and neural-collapse metrics continue the trends observed at larger SAM radii.

A.3 ImageNet-200 Fine-Tuning Setup

For the larger-scale extension, we fine-tune an ImageNet-1K-pretrained ViT-S/16 on ImageNet-200 and compare AdamW with SAM using AdamW as the base optimizer (Dosovitskiy et al., 2021; Deng et al., 2009; Loshchilov & Hutter, 2019; Foret et al., 2021).

- **Backbone:** ViT-S/16
- **Pretrained checkpoint:** `vit_small_patch16_224.augreg_in1k` from `timm`
- **Optimizer:** AdamW, AdamW+SAM
- **SAM radius:** $\rho \in \{0.05, 0.5\}$
- **Learning rate:** 2×10^{-4}, constant throughout the 20-epoch fine-tuning run
- **Weight decay:** 0.05
- **Batch size:** 512 (effective batch size 4096 with 8 gradient accumulation steps)
- **Fine-tuning epochs:** 20

All other hyperparameters are kept fixed between AdamW and AdamW+SAM. The 200-way classification head is initialized from the rows of the ImageNet-1K pretrained classifier corresponding to the selected ImageNet-200 classes.

A.4 Datasets

Experiments are conducted using CIFAR-10 as the primary ID training dataset (Krizhevsky, 2009). The OOD dataset settings for the primary CIFAR-10 experiments and the ImageNet-200 extension are summarized in Table 5.

For the ImageNet-200 extension, we construct the ID dataset by randomly selecting 200 eligible ImageNet-1K classes using seed 20260327. For each selected class, 1,000 images from the original ImageNet-1K training split are assigned to the fine-tuning set, and 300 disjoint images from the same split are held out as the ID test set. The original ImageNet-1K validation images for the selected classes are retained as the validation split.

Table 5: ID and OOD dataset settings for the primary CIFAR-10 experiments and the ImageNet-200 extension.

ID dataset	Near-OOD datasets	Far-OOD datasets
CIFAR-10	CIFAR-100, Tiny-ImageNet	MNIST, SVHN
ImageNet-200	SSB-Hard, NINCO	iNaturalist, OpenImage-O

B Supplementary Metrics and Evaluation Protocols

B.1 OOD Detection Methods and Evaluation

We evaluate post-hoc OOD detectors that depend on different components of the trained model: classifier logits, penultimate representations, and feature-space covariance structure. Unless other-

wise stated, all feature-based statistics are fitted using ID training representations and are evaluated on held-out ID and OOD test examples. OOD detection performance in the main analyses is reported using AUROC, while the extended quantitative results in Appendix G additionally report FPR@95, for which lower values indicate better OOD detection performance. OOD scores are oriented so that larger values indicate more ID-like inputs.

Let $f = f_\theta(x) \in \mathbb{R}^d$ denote the penultimate representation of an input x, and let $z_\theta(x) \in \mathbb{R}^K$ denote the classifier logits.

Logit-based scores. We use maximum softmax probability (MSP) and Energy as representative logit-based OOD scores (Hendrycks & Gimpel, 2017; Liu et al., 2020):

$$S_{\mathrm{MSP}}(x) = \max_c p_\theta(c \mid x), \qquad S_{\mathrm{Energy}}(x) = T \log \sum_{c=1}^{K} \exp\left(\frac{z_{\theta,c}(x)}{T}\right). \tag{8}$$

The Energy expression is the negative of the energy convention in Liu et al. (2020), so larger values correspond to greater ID confidence.

Mahalanobis score. Following the class-conditional Gaussian scoring principle of Lee et al. (2018), we estimate the class means μ_c from penultimate ID training representations together with a tied within-class covariance Σ. The Mahalanobis ID confidence is

$$S_{\mathrm{Maha}}(x) = -\min_c (f - \mu_c)^\top \Sigma^\dagger (f - \mu_c). \tag{9}$$

Thus, the score measures proximity to the closest class center under a metric determined by the covariance pseudoinverse, or precision, matrix. Our subsequent spectral analysis examines how this precision weighting behaves across eigendirections of the estimated covariance; the corresponding diagnostic details are given in Appendix C.

DDU. Deep Deterministic Uncertainty (DDU) fits a Gaussian discriminant model to penultimate ID representations after network training (Mukhoti et al., 2023). Unlike the tied covariance used in our Mahalanobis score, DDU estimates a class-conditional covariance Σ_c for each class. We use a uniform class prior, $\pi_c = 1/K$, which is omitted in the implementation because it contributes only a sample-independent additive constant. The corresponding feature-density score is

$$S_{\mathrm{DDU}}(x) = \log \sum_{c=1}^{K} \pi_c \mathcal{N}(f \mid \mu_c, \Sigma_c), \tag{10}$$

where

$$\log \mathcal{N}(f \mid \mu_c, \Sigma_c) = -\frac{1}{2}(f - \mu_c)^\top \Sigma_c^{-1}(f - \mu_c) - \frac{1}{2}\log|\Sigma_c| - \frac{d}{2}\log(2\pi). \tag{11}$$

Following the released DDU implementation (Mukhoti et al., 2023), we add isotropic diagonal jitter to a class-conditional covariance only when required for a valid Gaussian fit. Candidate jitter values are tested in a predefined sequence, and the first value yielding a valid positive-definite covariance is used. Thus, DDU depends on both precision-weighted residuals and the class-conditional covariance-volume term. Appendix D provides the corresponding spectral diagnostic.

kNN. We additionally evaluate the non-parametric nearest-neighbor detector of Sun et al. (2022). Let

$$\tilde{f} = \frac{f}{\max(\|f\|_2, \varepsilon)} \tag{12}$$

denote the normalized penultimate representation, with $\varepsilon = 10^{-12}$, and let $\tilde{f}_{(k)}$ be its k-th nearest neighbor among normalized ID training representations under Euclidean distance. The corresponding ID confidence is

$$S_{k\mathrm{NN}}(x) = -\left\|\tilde{f} - \tilde{f}_{(k)}\right\|_2^2. \tag{13}$$

For the primary CIFAR-10/WideResNet-28-10 experiments, we use $k = 50$. Unlike Mahalanobis and DDU, kNN does not estimate or invert a feature covariance matrix, providing a non-parametric comparison for representation-based OOD detection.

ViM. Virtual-logit Matching (ViM) combines classifier-logit information with the component of the representation lying outside a principal ID feature subspace (Wang et al., 2022). Let $P^{\perp}$ denote the residual subspace estimated from ID training representations and let $\|P^{\perp}(f-o)\|_2$ denote the corresponding residual norm after the feature-space offset o used by ViM. An ID-oriented form of the ViM score, equivalent up to a monotone transformation and sign convention, can be written as

$$S_{\mathrm{ViM}}(x) = \log \sum_{c=1}^{K} \exp\left(z_{\theta,c}(x)\right) - \alpha_{\mathrm{ViM}} \left\| P^{\perp}(f-o) \right\|_2 , \tag{14}$$

where α_{ViM} matches the scale of the feature residual to that of the classifier logits. For the primary CIFAR-10/WideResNet-28-10 experiments, the principal feature dimension used by ViM is 320. ViM is therefore a hybrid detector: it incorporates both logit-space confidence and feature-space information, but does not rely on class-conditional covariance inversion in the manner of Mahalanobis or DDU.

These detectors use different components of the trained model and learned representation. Our mechanistic analysis focuses on Mahalanobis and DDU because both depend explicitly on estimated covariance, while kNN and ViM provide complementary representation-dependent comparisons without the same covariance dependence.

B.2 Calibration Metrics

Expected Calibration Error (ECE). ECE quantifies the mismatch between predictive confidence and empirical accuracy (Naeini et al., 2015). Let $\{B_m\}_{m=1}^{M}$ be a partition of predictions into M confidence bins. ECE is defined as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right| , \tag{15}$$

where

$$\begin{aligned} \mathrm{acc}(B_m) &= \tfrac{1}{|B_m|} \textstyle\sum_{i \in B_m} \mathbb{I}(\hat{y}_i = y_i), \\ \mathrm{conf}(B_m) &= \tfrac{1}{|B_m|} \textstyle\sum_{i \in B_m} \max_k p_{ik}. \end{aligned}$$

We use $M = 15$ confidence bins in all reported ECE calculations.

Temperature-scaled ECE (T-ECE). T-ECE evaluates calibration after temperature scaling (Guo et al., 2017). Given logits z_i, temperature scaling applies

$$p_i^{(T)} = \mathrm{softmax}(z_i / T), \tag{16}$$

where T is selected on the ID validation set by minimizing negative log-likelihood over values from 0.1 to 10.0 in increments of 0.1. T-ECE is then computed on the ID test set as ECE using $p_i^{(T)}$ instead of the original softmax probabilities.

B.3 Representation Geometry Metrics

Within-class Variance (WithinVar). Within-class variance measures feature compactness within each class (Papyan et al., 2020). Let f_n denote the feature representation and μ_c the class-wise mean:

$$\mathrm{Var}_{\mathrm{within}} = \frac{1}{N} \sum_{c=1}^{C} \sum_{n:y_n=c} \|f_n - \mu_c\|_2^2. \tag{17}$$

Inter-class Distance (InterDist). Inter-class distance directly quantifies the separation between class mean representations. Let μ_c denote the mean feature of class c. We define

$$\mathrm{Dist}_{\mathrm{inter}} = \frac{2}{C(C-1)} \sum_{1 \le c < c' \le C} \|\mu_c - \mu_{c'}\|_2. \tag{18}$$

This quantity captures class-level separation in the representation space independently of within-class dispersion.

Anisotropy. We measure how strongly the within-class covariance spectrum is concentrated in its leading direction. For each class c, we compute the population covariance

$$\Sigma_c^{\text{geom}} = \frac{1}{N_c} \sum_{n:y_n=c} (f_n - \mu_c)(f_n - \mu_c)^\top. \tag{19}$$

Let $\bar{\lambda}_{c,1} \geq \cdots \geq \bar{\lambda}_{c,d} \geq 0$ denote its eigenvalues after clamping numerically negative values to zero, and let $s_c = \sum_{i=1}^{d} \bar{\lambda}_{c,i}$. The class-wise anisotropy is

$$A_c = \frac{\bar{\lambda}_{c,1}}{s_c + \varepsilon}, \tag{20}$$

and the reported metric is

$$\text{Anisotropy} = \frac{1}{C} \sum_{c=1}^{C} A_c. \tag{21}$$

Larger values indicate that a greater fraction of within-class covariance is concentrated in the leading eigendirection.

Effective Rank (EffRank). Using the same class-conditional covariance eigenvalues, we define

$$p_{c,i} = \frac{\bar{\lambda}_{c,i}}{s_c + \varepsilon}, \tag{22}$$

and compute the entropy-based effective rank

$$R_{\text{eff},c} = \exp\left(-\sum_{i=1}^{d} p_{c,i} \log\left(p_{c,i} + \varepsilon\right) \right). \tag{23}$$

The reported effective rank is the mean over classes,

$$\text{EffRank} = \frac{1}{C} \sum_{c=1}^{C} R_{\text{eff},c}. \tag{24}$$

Larger values indicate that within-class covariance is distributed over a larger number of eigendirections. In the implementation, $\varepsilon = 10^{-12}$ is used for numerical stability; classes with total covariance mass below this threshold are assigned zero for the corresponding class-wise statistic.

B.4 Neural Collapse Metrics

Following the neural-collapse literature (Papyan et al., 2020), we use several complementary quantities to characterize class-conditional representation geometry.

NC1 **(Variance Ratio).** NC1 measures within-class concentration relative to between-class spread. Here, Σ_W and Σ_B denote the within-class and between-class covariance matrices used in the NC1 computation, respectively.

$$\text{NC1} = \frac{1}{C} \operatorname{tr}\left(\Sigma_W \Sigma_B^{\dagger}\right), \tag{25}$$

where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse. Smaller values indicate stronger within-class contraction relative to between-class variation.

NC2 **(ETF Alignment).** Let $W \in \mathbb{R}^{C \times d}$ denote the linear-classifier weight matrix and

$$G_W = W W^\top. \tag{26}$$

NC2 measures the discrepancy between the normalized classifier-weight Gram matrix and the Gram structure of an ideal simplex equiangular tight frame (ETF).

$$\text{NC2} = \left\| \frac{G_W}{\|G_W\|_F} - \frac{I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top}{\left\|I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top\right\|_F} \right\|_F. \tag{27}$$

$\mathrm{NC2^A}$ **(Angular Structure of Centered Class Means).** Let

$$\bar{\mu} = \frac{1}{C}\sum_{c=1}^{C}\mu_c, \qquad \tilde{\mu}_c = \mu_c - \bar{\mu}.$$

We measure the mean pairwise cosine similarity between centered class means:

$$\mathrm{NC2^A} = \frac{1}{C(C-1)}\sum_{i\neq j}\frac{\langle\tilde{\mu}_i,\tilde{\mu}_j\rangle}{\|\tilde{\mu}_i\|_2\|\tilde{\mu}_j\|_2}. \tag{28}$$

For an ideal simplex ETF, $\mathrm{NC2^A} = -1/(C-1)$.

NC3 **(Classifier Alignment).** Let $H = [\tilde{\mu}_1,\ldots,\tilde{\mu}_C] \in \mathbb{R}^{d\times C}$ denote the matrix of centered class means. We compute

$$\mathrm{NC3} = \left\| \frac{WH}{\|WH\|_F} - \frac{I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top}{\sqrt{C-1}} \right\|_F. \tag{29}$$

NC4 **(Nearest-Class-Center Accuracy).** For each representation f_n, let

$$\hat{y}_n^{\mathrm{NCC}} = \arg\min_c \|f_n - \mu_c\|_2. \tag{30}$$

We compute

$$\mathrm{NC4} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\left(\hat{y}_n^{\mathrm{NCC}} = y_n\right). \tag{31}$$

C Spectral Analysis Protocol for Mahalanobis Detection

The main text analyzes how covariance weighting interacts with OOD-relevant residual structure in representations learned under SGD and SAM. This section provides the corresponding spectral evaluation protocol.

C.1 Tied Within-Class Covariance and Eigendecomposition

Let $z = f(x) \in \mathbb{R}^d$ denote the penultimate representation and let μ_c be the mean representation of ID class c. We estimate a tied within-class covariance matrix from the ID training features:

$$\Sigma_W = \frac{1}{N}\sum_{c=1}^{C}\sum_{n:y_n=c}(z_n-\mu_c)(z_n-\mu_c)^\top. \tag{32}$$

We eigendecompose

$$\Sigma_W = U\Lambda U^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_d), \tag{33}$$

with eigenvalues ordered as

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d.$$

The class means and residual features used in this analysis are formed in FP32. The residuals are then promoted to FP64 for covariance estimation, eigendecomposition, pseudoinverse rank selection, and precision construction, while the final Mahalanobis score is evaluated in FP32, matching the main Mahalanobis evaluation. For numerical stability, we retain the covariance eigendirections selected by the corresponding pseudoinverse threshold

$$\tau = \lambda_1 d\,\epsilon_{\mathrm{fp64}}, \tag{34}$$

where ϵ_{fp64} denotes machine epsilon for FP64. The active covariance eigendirections are

$$\mathcal{I}_{\mathrm{eff}} = \{i : \lambda_i > \tau\}. \tag{35}$$

We denote their number by $d_{\mathrm{eff}} = |\mathcal{I}_{\mathrm{eff}}|$ and index the retained directions in descending-eigenvalue order.

C.2 Generalized Covariance Weighting

To isolate the contribution of inverse-covariance weighting, we apply a power transformation to the active covariance eigenvalues while preserving their total scale. Specifically, we define

$$\tilde{\lambda}_i^{(\alpha)} = \lambda_i^{\alpha} \frac{\sum_{j \in \mathcal{I}_{\text{eff}}} \lambda_j}{\sum_{j \in \mathcal{I}_{\text{eff}}} \lambda_j^{\alpha}}, \qquad i \in \mathcal{I}_{\text{eff}}, \tag{36}$$

and the corresponding class-conditional distance

$$d_\alpha(z, c) = \sum_{i \in \mathcal{I}_{\text{eff}}} \frac{\left[u_i^\top (z - \mu_c)\right]^2}{\tilde{\lambda}_i^{(\alpha)}}, \qquad \alpha \in [0, 1]. \tag{37}$$

The corresponding confidence score is

$$S_\alpha(z) = -\min_c d_\alpha(z, c). \tag{38}$$

At $\alpha = 0$, all active covariance eigendirections receive equal weight, and the score is proportional to a Euclidean class-mean distance within the active eigensubspace. At $\alpha = 1$, the transformed eigenvalues reduce to the original eigenvalues, recovering the usual pseudoinverse Mahalanobis weighting. Intermediate values continuously vary the strength of covariance-derived reweighting without changing the underlying checkpoint representation.

The main text writes the intervention directly using λ_i^α. For a fixed checkpoint and α, the scale-preserving normalization above differs only by a positive factor shared by all samples and classes, and therefore does not change the selected class or OOD AUROC.

This intervention is diagnostic rather than a proposed detector. Its purpose is to separate changes in the learned representation from changes in how that representation is read out by covariance weighting.

C.3 Variance-Aware Directional Separation and Covariance–Contrast Mismatch

Following the per-direction Mahalanobis decomposition used by Janiak et al. (2026), for a sample z we let

$$c_M(z) = \arg\min_c \sum_{i \in \mathcal{I}_{\text{eff}}} \frac{\left[u_i^\top (z - \mu_c)\right]^2}{\lambda_i} \tag{39}$$

denote the class selected by the full Mahalanobis score within the active eigensubspace. We define the residual coordinate along eigendirection i as

$$a_i(z) = u_i^\top \left(z - \mu_{c_M(z)}\right). \tag{40}$$

The directional residual contrast is

$$\Delta_i^{r^2} = \mathbb{E}_{z \sim P_{\text{OOD}}}\left[a_i^2(z)\right] - \mathbb{E}_{z \sim P_{\text{ID}}}\left[a_i^2(z)\right]. \tag{41}$$

We measure its pooled marginal variability as

$$V_i = \frac{1}{2}\left(\text{Var}_{z \sim P_{\text{OOD}}}\left[a_i^2(z)\right] + \text{Var}_{z \sim P_{\text{ID}}}\left[a_i^2(z)\right]\right). \tag{42}$$

The empirical variances are computed using the population normalization. We then define the variance-normalized directional contrast

$$\widetilde{\Delta}_i^{r^2} = \frac{\Delta_i^{r^2}}{\sqrt{V_i}}. \tag{43}$$

Larger $\widetilde{\Delta}_i^{r^2}$ indicates stronger ID–OOD separation relative to the marginal variability of that direction.

To derive the covariance–contrast mismatch condition in Eq. 6 of the main text, consider two positive-contrast directions i and j satisfying

$$\widetilde{\Delta}_i^{r^2} < \widetilde{\Delta}_j^{r^2}, \qquad \lambda_i < \lambda_j.$$

Under covariance weighting strength α, define

$$\beta_i(\alpha) = \lambda_i^{-\alpha}\sqrt{V_i}, \qquad t_\alpha = \frac{\beta_i(\alpha)}{\beta_j(\alpha)} = \left(\frac{\lambda_j}{\lambda_i}\right)^{\alpha}\sqrt{\frac{V_i}{V_j}}. \tag{44}$$

For this pairwise diagnostic, we use the marginal pooled variances of the two directions and omit cross-direction covariance. The corresponding variance-normalized separation can then be written as

$$J_{ij}(t_\alpha) = \frac{\widetilde{\Delta}_j^{r^2} + t_\alpha \widetilde{\Delta}_i^{r^2}}{\sqrt{1+t_\alpha^2}}. \tag{45}$$

Differentiating with respect to the relative effective weight gives

$$\frac{\partial J_{ij}}{\partial t_\alpha} = \frac{\widetilde{\Delta}_i^{r^2} - t_\alpha \widetilde{\Delta}_j^{r^2}}{(1+t_\alpha^2)^{3/2}}. \tag{46}$$

Thus, increasing the relative effective weight on direction i reduces this pairwise variance-normalized separation when

$$t_\alpha > \frac{\widetilde{\Delta}_i^{r^2}}{\widetilde{\Delta}_j^{r^2}}, \tag{47}$$

which is the covariance–contrast mismatch condition reported in Eq. 6 of the main text. The reported mismatch prevalence is the fraction of direction pairs satisfying the stated positive-contrast and ordering conditions for which this inequality holds. Because this diagnostic omits cross-direction covariance, it characterizes the alignment between covariance-derived weighting and marginal directional contrast rather than the exact pooled variance of the full multi-direction score.

C.4 Cumulative Prefix Scores

We additionally evaluate how OOD discrimination accumulates as eigendirections are progressively included. Using the active eigendirections ordered by descending eigenvalue, for $k = 1, \ldots, d_{\text{eff}}$ we define

$$d_{\alpha,1:k}(z,c) = \sum_{i=1}^{k} \frac{\left[u_i^\top (z-\mu_c)\right]^2}{\lambda_i^\alpha}, \qquad \alpha \in \{0,1\}. \tag{48}$$

The corresponding prefix confidence score is

$$S_{\alpha,1:k}(z) = -\min_c d_{\alpha,1:k}(z,c). \tag{49}$$

The winning class is selected separately for every prefix and weighting condition:

$$c_{\alpha,1:k}(z) = \arg\min_c d_{\alpha,1:k}(z,c). \tag{50}$$

Thus, each prefix evaluates the corresponding truncated readout directly rather than fixing the class selected by the full-dimensional detector.

At $\alpha = 0$, the included eigendirections receive equal weight, whereas $\alpha = 1$ applies the inverse-covariance Mahalanobis weighting. Prefix AUROC therefore measures how OOD discrimination changes as progressively later-ranked covariance eigendirections are incorporated under the two weighting schemes.

C.5 Comparison Across Checkpoints

The covariance eigendirections of independently trained checkpoints are not assumed to admit a one-to-one correspondence. Accordingly, our analysis compares the *ranked covariance structure* within each checkpoint rather than matching individual eigendirections across SGD and SAM models.

The covariance eigensystem, pseudoinverse threshold, active-subspace dimension, directional residual statistics, mismatch diagnostics, and prefix scores are computed independently within each checkpoint. This avoids attributing semantic identity to eigendirections obtained from different independently trained models.

Across runs, directional statistics and prefix AUROCs are averaged at matching ranks or prefix sizes within the common active support of each training condition, while mismatch prevalence is computed independently within each run and then averaged across runs.

D DDU Spectral Diagnostic Details

DDU uses class-conditional Gaussian densities and therefore depends explicitly on class-specific covariance structure (Mukhoti et al., 2023). This section provides the spectral prefix protocol used to examine how OOD discrimination changes as class-wise covariance eigendirections are progressively incorporated.

D.1 Class-Conditional Gaussian Model

For each class c, let μ_c denote the empirical mean of the ID training representations and let

$$\Sigma_c = \frac{1}{N_c - 1} \sum_{n:y_n=c} (z_n - \mu_c)(z_n - \mu_c)^\top \tag{51}$$

denote the corresponding empirical class covariance. We eigendecompose

$$\Sigma_c = U_c \Lambda_c U_c^\top, \tag{52}$$

where

$$\Lambda_c = \mathrm{diag}(\lambda_{c,1}, \ldots, \lambda_{c,d}), \qquad \lambda_{c,1} \geq \lambda_{c,2} \geq \cdots \geq \lambda_{c,d}. \tag{53}$$

For a feature z, the class-conditional Gaussian log density is

$$\ell_c(z) = -\frac{1}{2} \left[(z - \mu_c)^\top \Sigma_c^{-1} (z - \mu_c) + \log |\Sigma_c| + d \log(2\pi) \right]. \tag{54}$$

Let π_c denote the class prior. DDU confidence is obtained by aggregating the class-conditional densities as

$$S_{\mathrm{DDU}}(z) = \log \sum_{c=1}^{C} \pi_c \exp\left(\ell_c(z)\right). \tag{55}$$

In the balanced CIFAR-10 setting used for the spectral analysis, the class prior is uniform, so omitting π_c changes the score only by a sample-independent additive constant.

D.2 Spectral Prefix Diagnostic

For the spectral diagnostic, we retain all class-wise covariance eigendirections and progressively include them in descending-eigenvalue order. For the first k eigendirections, the corresponding class-conditional prefix log density is

$$\ell_{c,1:k}(z) = -\frac{1}{2} \left[\sum_{i=1}^{k} \frac{\left[u_{c,i}^\top (z - \mu_c)\right]^2}{\tilde{\lambda}_{c,i}} + \sum_{i=1}^{k} \log \tilde{\lambda}_{c,i} + k \log(2\pi) \right], \tag{56}$$

where

$$\tilde{\lambda}_{c,i} = \max(\lambda_{c,i}, \epsilon_{\mathrm{mach}}) \tag{57}$$

and ϵ_{mach} denotes FP64 machine epsilon. This clamp is used only to ensure numerical stability in the spectral diagnostic.

Both the inverse-quadratic and log-determinant terms are retained because DDU is a Gaussian density score rather than a score based only on Mahalanobis distance. Under the uniform class prior of the balanced CIFAR-10 setting, the corresponding prefix confidence score can be written up to a sample-independent additive constant as

$$S_{1:k}(z) = \log \sum_{c=1}^{C} \exp\left(\ell_{c,1:k}(z)\right). \tag{58}$$

The spectral prefix experiment is used as a diagnostic and is not treated as a replacement detector. The main DDU implementation and the spectral diagnostic use different numerical stabilization procedures: the former uses diagonal jitter when required, whereas the latter clamps the eigenvalues used in the truncated eigenspectrum. Consequently, the final prefix value at $k = d$ need not exactly reproduce the full-dimensional DDU score reported in the main evaluation. The prefix diagnostic is instead used to examine how OOD discrimination changes as progressively more class-wise covariance eigendirections are incorporated across training conditions.

E Extended Controls

E.1 ASAM and GSAM

This appendix provides additional details for Adaptive Sharpness-Aware Minimization (ASAM) and Surrogate Gap Guided Sharpness-Aware Minimization (GSAM), the two sharpness-aware optimizer variants compared with vanilla SAM in the main text.

E.1.1 Adaptive Sharpness-Aware Minimization (ASAM)

ASAM (Kwon et al., 2021) addresses a scale-dependency issue in vanilla SAM. In SAM, sharpness is measured within a fixed-radius perturbation ball around the current weights. However, this estimate can change under parameter rescaling even when the network function and training loss remain unchanged:

$$\max_{\|\epsilon\|_2 \le \rho} \mathcal{L}_S(w+\epsilon) \neq \max_{\|\epsilon\|_2 \le \rho} \mathcal{L}_S(Aw+\epsilon).$$

To reduce this sensitivity, ASAM replaces the fixed spherical perturbation region with a weight-adaptive ellipsoidal region defined via

$$T_w = \mathrm{diag}(|w_1| + \eta, \ldots, |w_k| + \eta), \qquad \eta > 0, \tag{59}$$

and computes the perturbation as

$$\hat{\epsilon}_t = \rho \frac{T_{w_t}^2 \nabla \mathcal{L}_S(w_t)}{\|T_{w_t} \nabla \mathcal{L}_S(w_t)\|_2}. \tag{60}$$

Compared with SAM, the perturbation adapts to the scale of the current parameters through T_{w_t}. In our experiments, we use $\eta = 0.01$.

E.1.2 Surrogate Gap Guided Sharpness-Aware Minimization (GSAM)

GSAM (Zhuang et al., 2022) observes that minimizing the perturbed loss $f_p(w) \triangleq \max_{\|\delta\| \le \rho} f(w+\delta)$ alone does not guarantee a flat solution, since sharp and flat minima can share similar f_p values. To address this, GSAM introduces the *surrogate gap*

$$h(w) \triangleq f_p(w) - f(w), \tag{61}$$

which serves as a tractable proxy for local sharpness. GSAM jointly minimizes f_p and h without conflating their gradients. To this end, $\nabla f(w)$ is decomposed into components parallel and orthogonal to $\nabla f_p(w)$:

$$\nabla f_\perp(w) = \nabla f(w) - \frac{\nabla f(w) \cdot \nabla f_p(w)}{\|\nabla f_p(w)\|^2} \nabla f_p(w). \tag{62}$$

Table 6: Effect of optimizer choice on representation geometry and OOD detection for WideResNet-28-10 trained on CIFAR-10, with CIFAR-100 used as OOD data. Values are mean ± standard deviation over three seeds. For GSAM, ρ denotes $\rho_{\max}$, with $\alpha = 0.02$ and $\rho_{\min} = 0.01$.

Optimizer	ρ	WithinVar	InterDist	$\mathcal{NC}_1$	DDU AUROC	Maha AUROC
SGD	0	$0.8655_{\pm 0.0244}$	$7.3106_{\pm 0.3599}$	$0.0221_{\pm 0.0009}$	$0.9066_{\pm 0.0024}$	$0.9076_{\pm 0.0035}$
SAM	0.01	$0.9261_{\pm 0.0629}$	$7.4538_{\pm 0.0153}$	$0.0240_{\pm 0.0015}$	$0.9076_{\pm 0.0021}$	$0.9114_{\pm 0.0029}$
	0.02	$1.0805_{\pm 0.0862}$	$7.4644_{\pm 0.0632}$	$0.0278_{\pm 0.0016}$	$0.9030_{\pm 0.0106}$	$0.9113_{\pm 0.0042}$
	0.05	$1.2640_{\pm 0.1631}$	$7.1803_{\pm 0.3163}$	$0.0342_{\pm 0.0044}$	$0.9114_{\pm 0.0042}$	$0.9119_{\pm 0.0030}$
	0.1	$1.7433_{\pm 0.1295}$	$7.1992_{\pm 0.0944}$	$0.0474_{\pm 0.0050}$	$0.9048_{\pm 0.0009}$	$0.9074_{\pm 0.0068}$
	0.2	$2.0108_{\pm 0.0214}$	$6.9563_{\pm 0.0558}$	$0.0549_{\pm 0.0014}$	$0.9065_{\pm 0.0048}$	$0.9032_{\pm 0.0012}$
	0.5	$2.4303_{\pm 0.0142}$	$6.8398_{\pm 0.0089}$	$0.0639_{\pm 0.0004}$	$0.8832_{\pm 0.0058}$	$0.8787_{\pm 0.0023}$
ASAM	0.01	$0.9319_{\pm 0.1233}$	$7.5076_{\pm 0.0421}$	$0.0238_{\pm 0.0032}$	$0.9038_{\pm 0.0022}$	$0.9064_{\pm 0.0034}$
	0.02	$0.9392_{\pm 0.1514}$	$7.5100_{\pm 0.0797}$	$0.0237_{\pm 0.0037}$	$0.9033_{\pm 0.0065}$	$0.9083_{\pm 0.0068}$
	0.05	$0.8869_{\pm 0.0645}$	$7.4745_{\pm 0.0567}$	$0.0223_{\pm 0.0016}$	$0.9047_{\pm 0.0034}$	$0.9090_{\pm 0.0017}$
	0.1	$0.9613_{\pm 0.0868}$	$7.4780_{\pm 0.0510}$	$0.0243_{\pm 0.0021}$	$0.9069_{\pm 0.0021}$	$0.9106_{\pm 0.0010}$
	0.2	$0.9397_{\pm 0.0787}$	$7.4405_{\pm 0.0737}$	$0.0234_{\pm 0.0021}$	$0.9078_{\pm 0.0051}$	$0.9104_{\pm 0.0039}$
	0.5	$1.0127_{\pm 0.0658}$	$7.4404_{\pm 0.0466}$	$0.0248_{\pm 0.0019}$	$0.9114_{\pm 0.0024}$	$0.9124_{\pm 0.0040}$
GSAM	0.01	$0.9991_{\pm 0.0580}$	$7.4534_{\pm 0.0332}$	$0.0260_{\pm 0.0013}$	$0.9039_{\pm 0.0009}$	$0.9093_{\pm 0.0028}$
	0.02	$1.0057_{\pm 0.0426}$	$7.4571_{\pm 0.0303}$	$0.0261_{\pm 0.0011}$	$0.9056_{\pm 0.0022}$	$0.9082_{\pm 0.0013}$
	0.05	$0.8966_{\pm 0.1063}$	$7.3898_{\pm 0.0222}$	$0.0232_{\pm 0.0032}$	$0.9087_{\pm 0.0026}$	$0.9139_{\pm 0.0010}$
	0.1	$0.9448_{\pm 0.0496}$	$7.3154_{\pm 0.0349}$	$0.0250_{\pm 0.0016}$	$0.9084_{\pm 0.0017}$	$0.9129_{\pm 0.0020}$
	0.2	$0.8771_{\pm 0.0058}$	$7.2486_{\pm 0.0380}$	$0.0228_{\pm 0.0004}$	$0.9133_{\pm 0.0029}$	$0.9182_{\pm 0.0016}$
	0.5	$1.0925_{\pm 0.0332}$	$7.1839_{\pm 0.0365}$	$0.0282_{\pm 0.0016}$	$0.9070_{\pm 0.0018}$	$0.9205_{\pm 0.0017}$

The weight update then uses the surrogate gradient

$$\nabla f^{\mathrm{GSAM}} \triangleq \nabla f_p(w) - \alpha \nabla f_\perp(w), \tag{63}$$

where $\alpha \geq 0$ scales the orthogonal ascent step. Because $\nabla f_\perp \perp \nabla f_p$ by construction, this step decreases h by increasing $f(w)$ while leaving $f_p(w)$ unchanged to first order.

In our experiments, we use $\alpha = 0.02$ and $\rho_{\min} = 0.01$. The perturbation radius is scheduled proportionally to the learning rate between $\rho_{\max}$ and $\rho_{\min}$, and the reported GSAM value of ρ denotes $\rho_{\max}$.

E.1.3 Extended ρ Sweep

Table 6 reports additional results for ASAM and GSAM across a wider range of perturbation radii.

This extended sweep complements the representative comparison in the main text by providing a more detailed view of how these optimizer variants are associated with representation geometry and covariance-sensitive detector behavior.

Across the sweep, ASAM shows relatively mild changes in representation geometry. WithinVar and $\mathrm{NC1}$ remain close to the SGD baseline across the tested values of ρ, while DDU and Mahalanobis AUROC remain comparatively stable. Thus, ASAM does not reproduce the large increase in within-class dispersion or the pronounced covariance-sensitive degradation observed under larger-radius vanilla SAM in the main setting.

GSAM shows a similarly stable profile across the tested perturbation radii. Although its geometry measures vary with ρ, WithinVar and $\mathrm{NC1}$ remain substantially closer to the SGD range than under large-radius vanilla SAM, while DDU and Mahalanobis AUROC remain comparatively stable.

These results complement the representative comparison in the main text: ASAM and GSAM do not reproduce the large-radius degradation observed along the vanilla SAM trajectory. The observed degradation should therefore not be interpreted as a generic consequence of sharpness-aware optimization.

E.2 Stochastic Weight Averaging (SWA)

SWA (Izmailov et al., 2018) constructs a single model by averaging weights sampled along an SGD optimization trajectory. Rather than modifying the training objective through adversarial weight perturbations, as in SAM, SWA maintains an equally weighted average of the sampled parameter vectors. Given weights $w_1, \ldots, w_T$ collected during training, the SWA solution is

$$w_{\mathrm{SWA}} = \frac{1}{T} \sum_{t=1}^{T} w_t. \tag{64}$$

Equivalently, the average can be updated incrementally as

$$w_{\mathrm{SWA}}^{(t)} = \frac{t-1}{t} w_{\mathrm{SWA}}^{(t-1)} + \frac{1}{t} w_t. \tag{65}$$

In its original formulation, these weights are sampled from an SGD trajectory using a constant or cyclical learning-rate schedule, allowing the trajectory to explore different points within the same high-performing region of the loss landscape. The resulting averaged solution has been associated with wider regions of low loss than conventional SGD solutions.

We include SWA as a complementary optimization control because, unlike SAM, ASAM, and GSAM, it does not explicitly optimize a perturbation-based worst-case objective. It therefore provides a qualitatively different training change for testing whether predictive improvement and covariance-sensitive OOD performance can separate outside the vanilla SAM radius trajectory.

For the 350-epoch SWA runs, weight averaging begins at approximately 75% of training (zero-based epoch 262, corresponding to the 263rd training epoch) and is updated once per epoch thereafter. During the averaging phase, the learning rate is fixed to 10^{-2}, while the remaining SGD hyperparameters match the baseline in Appendix A.1. After training, the averaged weights are copied into the model and the BatchNorm running statistics are recomputed once over the training set.

Table 7: ID predictive performance with Stochastic Weight Averaging (SWA) on CIFAR-10. Values are mean ± standard deviation over three seeds.

Training	Accuracy ↑	ECE ↓	T-ECE ↓
SGD	$0.9597_{\pm 0.0023}$	$0.0241_{\pm 0.0029}$	$0.0088_{\pm 0.0027}$
SWA	$0.9639_{\pm 0.0004}$	$0.0206_{\pm 0.0002}$	$0.0061_{\pm 0.0004}$

Table 8: Per-dataset OOD AUROC with Stochastic Weight Averaging (SWA) on CIFAR-10. Values are mean ± standard deviation over three seeds.

Method	Training	OOD AUROC ↑			
		CIFAR-100	Tiny-ImageNet	SVHN	MNIST
MSP	SGD	$0.8965_{\pm 0.0022}$	$0.8869_{\pm 0.0015}$	$0.9289_{\pm 0.0399}$	$0.8986_{\pm 0.0365}$
	SWA	$0.9114_{\pm 0.0025}$	$0.9015_{\pm 0.0029}$	$0.9663_{\pm 0.0101}$	$0.9282_{\pm 0.0125}$
Mahalanobis	SGD	$0.9076_{\pm 0.0035}$	$0.9032_{\pm 0.0018}$	$0.9781_{\pm 0.0028}$	$0.9408_{\pm 0.0224}$
	SWA	$0.9030_{\pm 0.0044}$	$0.8973_{\pm 0.0050}$	$0.9863_{\pm 0.0004}$	$0.9384_{\pm 0.0072}$
DDU	SGD	$0.9066_{\pm 0.0024}$	$0.9059_{\pm 0.0038}$	$0.9789_{\pm 0.0018}$	$0.9135_{\pm 0.0275}$
	SWA	$0.9006_{\pm 0.0003}$	$0.8976_{\pm 0.0018}$	$0.9780_{\pm 0.0039}$	$0.9094_{\pm 0.0098}$

As reported in the main text, SWA improves accuracy and calibration relative to SGD, while Mahalanobis and DDU do not show a corresponding consistent improvement across OOD datasets. We therefore use SWA as evidence that the separation between predictive improvement and covariance-sensitive OOD performance is not specific to vanilla SAM, rather than as evidence that the covariance-weighting mechanism isolated in Section 4 is reproduced under SWA.

E.3 Dataset and Backbone Controls

The primary mechanistic analysis focuses on CIFAR-10 with WideResNet-28-10, where increasing the SAM radius produces a clear detector-dependent divergence that can be studied in detail. We additionally evaluate other datasets and training regimes to examine the scope of this observation.

CIFAR-100 as ID. When CIFAR-100 is used as the ID training dataset, SAM does not reproduce the same covariance-sensitive degradation that motivates our CIFAR-10 analysis. In this setting, covariance-sensitive readouts instead show mixed, OOD-dependent responses under SAM. We therefore do not interpret the CIFAR-10 phenomenon as a universal consequence of SAM, nor do we claim that the specific mechanism characterized in the main text generalizes to every dataset.

Table 9: ID predictive performance when CIFAR-100 is used as the ID training dataset with WideResNet-28-10. Values are mean ± standard deviation over three seeds.

ρ	Accuracy ↑	ECE ↓	T-ECE ↓
0.00 (SGD)	$0.8099_{\pm 0.0024}$	$0.0559_{\pm 0.0027}$	$0.0385_{\pm 0.0049}$
0.01	$0.8097_{\pm 0.0014}$	$0.0553_{\pm 0.0020}$	$0.0362_{\pm 0.0038}$
0.02	$0.8151_{\pm 0.0040}$	$0.0511_{\pm 0.0024}$	$0.0370_{\pm 0.0025}$
0.05	$0.8180_{\pm 0.0015}$	$0.0445_{\pm 0.0055}$	$0.0316_{\pm 0.0040}$
0.10	$0.8251_{\pm 0.0021}$	$0.0388_{\pm 0.0047}$	$0.0274_{\pm 0.0037}$
0.20	$0.8271_{\pm 0.0012}$	$0.0351_{\pm 0.0056}$	$0.0221_{\pm 0.0036}$
0.50	$0.8310_{\pm 0.0012}$	$0.0303_{\pm 0.0027}$	$0.0158_{\pm 0.0014}$

Table 10: Per-dataset OOD AUROC when CIFAR-100 is used as the ID training dataset with WideResNet-28-10. Values are mean ± standard deviation over three seeds.

Method	ρ	OOD AUROC ↑			
		CIFAR-10	Tiny-ImageNet	SVHN	MNIST
MSP	0.00 (SGD)	$0.8116_{\pm 0.0011}$	$0.8188_{\pm 0.0006}$	$0.8362_{\pm 0.0205}$	$0.6589_{\pm 0.0436}$
	0.01	$0.8129_{\pm 0.0006}$	$0.8197_{\pm 0.0019}$	$0.8353_{\pm 0.0049}$	$0.6593_{\pm 0.0501}$
	0.02	$0.8179_{\pm 0.0023}$	$0.8222_{\pm 0.0034}$	$0.8329_{\pm 0.0187}$	$0.6583_{\pm 0.0245}$
	0.05	$0.8199_{\pm 0.0032}$	$0.8248_{\pm 0.0006}$	$0.8443_{\pm 0.0056}$	$0.6789_{\pm 0.0326}$
	0.10	$0.8239_{\pm 0.0031}$	$0.8276_{\pm 0.0008}$	$0.8338_{\pm 0.0133}$	$0.6220_{\pm 0.0149}$
	0.20	$0.8225_{\pm 0.0041}$	$0.8286_{\pm 0.0028}$	$0.8138_{\pm 0.0099}$	$0.6559_{\pm 0.0128}$
	0.50	$0.8200_{\pm 0.0027}$	$0.8264_{\pm 0.0015}$	$0.8364_{\pm 0.0172}$	$0.6271_{\pm 0.0281}$
Mahalanobis	0.00 (SGD)	$0.7831_{\pm 0.0021}$	$0.8220_{\pm 0.0018}$	$0.8872_{\pm 0.0105}$	$0.6591_{\pm 0.0199}$
	0.01	$0.7786_{\pm 0.0050}$	$0.8221_{\pm 0.0033}$	$0.9109_{\pm 0.0130}$	$0.6850_{\pm 0.0355}$
	0.02	$0.7814_{\pm 0.0043}$	$0.8237_{\pm 0.0038}$	$0.8922_{\pm 0.0192}$	$0.6689_{\pm 0.0488}$
	0.05	$0.7767_{\pm 0.0049}$	$0.8258_{\pm 0.0007}$	$0.9116_{\pm 0.0224}$	$0.6981_{\pm 0.0374}$
	0.10	$0.7718_{\pm 0.0062}$	$0.8224_{\pm 0.0046}$	$0.9011_{\pm 0.0255}$	$0.6826_{\pm 0.0126}$
	0.20	$0.7613_{\pm 0.0100}$	$0.8133_{\pm 0.0050}$	$0.9299_{\pm 0.0162}$	$0.6819_{\pm 0.0335}$
	0.50	$0.7654_{\pm 0.0050}$	$0.8127_{\pm 0.0053}$	$0.8945_{\pm 0.0277}$	$0.7255_{\pm 0.0366}$
DDU	0.00 (SGD)	$0.5422_{\pm 0.0409}$	$0.5902_{\pm 0.0213}$	$0.6379_{\pm 0.1038}$	$0.4977_{\pm 0.0669}$
	0.01	$0.5296_{\pm 0.0412}$	$0.5856_{\pm 0.0291}$	$0.6910_{\pm 0.0555}$	$0.5471_{\pm 0.0456}$
	0.02	$0.5114_{\pm 0.0472}$	$0.5610_{\pm 0.0299}$	$0.5597_{\pm 0.1721}$	$0.6034_{\pm 0.0324}$
	0.05	$0.5021_{\pm 0.0258}$	$0.5499_{\pm 0.0386}$	$0.6153_{\pm 0.0673}$	$0.5238_{\pm 0.0784}$
	0.10	$0.5558_{\pm 0.0395}$	$0.5860_{\pm 0.0036}$	$0.6480_{\pm 0.0062}$	$0.5735_{\pm 0.0215}$
	0.20	$0.5322_{\pm 0.0699}$	$0.5592_{\pm 0.0569}$	$0.6872_{\pm 0.0472}$	$0.5807_{\pm 0.0478}$
	0.50	$0.5614_{\pm 0.0548}$	$0.5915_{\pm 0.0221}$	$0.6204_{\pm 0.1300}$	$0.6139_{\pm 0.0682}$

DDU jitter-grid check. In the CIFAR-100-ID setting, the feature dimension is larger than the number of training samples available per class, so the class-conditional empirical covariance is rank-deficient before regularization. DDU already uses the first valid diagonal jitter as described in Appendix B. For this setting, we additionally evaluate the same checkpoints over an extended

jitter grid and report the best AUROC obtained among those jitter values in Table 11. This table is a diagnostic upper envelope rather than the selection rule used for the DDU rows above.

Table 11: Best DDU OOD AUROC over the extended covariance-jitter grid when CIFAR-100 is used as the ID dataset. For each checkpoint and OOD dataset, the highest AUROC over the extended jitter grid is selected before averaging over three seeds.

ρ	OOD AUROC ↑			
	CIFAR-10	Tiny-ImageNet	SVHN	MNIST
0.00 (SGD)	$0.7703_{\pm 0.0024}$	$0.8183_{\pm 0.0030}$	$0.8756_{\pm 0.0121}$	$0.6831_{\pm 0.0455}$
0.01	$0.7651_{\pm 0.0060}$	$0.8187_{\pm 0.0040}$	$0.8918_{\pm 0.0172}$	$0.7039_{\pm 0.0604}$
0.02	$0.7716_{\pm 0.0045}$	$0.8221_{\pm 0.0036}$	$0.8767_{\pm 0.0486}$	$0.6996_{\pm 0.0233}$
0.05	$0.7754_{\pm 0.0018}$	$0.8271_{\pm 0.0025}$	$0.8733_{\pm 0.0272}$	$0.7332_{\pm 0.0192}$
0.10	$0.7799_{\pm 0.0039}$	$0.8282_{\pm 0.0039}$	$0.8718_{\pm 0.0111}$	$0.7269_{\pm 0.0908}$
0.20	$0.7809_{\pm 0.0054}$	$0.8296_{\pm 0.0024}$	$0.8782_{\pm 0.0439}$	$0.7051_{\pm 0.0555}$
0.50	$0.7763_{\pm 0.0013}$	$0.8239_{\pm 0.0064}$	$0.8506_{\pm 0.0411}$	$0.7520_{\pm 0.0415}$

ImageNet-200 fine-tuning. We also evaluate AdamW and AdamW+SAM when fine-tuning a Vision Transformer on ImageNet-200. As reported in the main text, accuracy, ECE, and MSP improve as the SAM radius increases, whereas Mahalanobis remains approximately stable and DDU decreases. This setting therefore exhibits a broader separation between predictive improvement and representation-dependent uncertainty outside the primary CIFAR-scale CNN setting.

Table 12: ID predictive performance for ImageNet-200 fine-tuning with ViT-S/16. Values are mean $\pm$ standard deviation over three seeds.

ρ	Accuracy ↑	ECE ↓	T-ECE ↓
0.00 (AdamW)	$0.9249_{\pm 0.0009}$	$0.0226_{\pm 0.0001}$	$0.0084_{\pm 0.0003}$
0.05	$0.9294_{\pm 0.0006}$	$0.0147_{\pm 0.0003}$	$0.0084_{\pm 0.0004}$
0.50	$0.9392_{\pm 0.0003}$	$0.0104_{\pm 0.0005}$	$0.0104_{\pm 0.0005}$

Table 13: Per-dataset OOD AUROC for ImageNet-200 fine-tuning with ViT-S/16. Values are mean $\pm$ standard deviation over three seeds.

Method	ρ	OOD AUROC ↑			
		iNaturalist	NINCO	OpenImage-O	SSB-Hard
MSP	0.00 (AdamW)	$0.9488_{\pm 0.0025}$	$0.9158_{\pm 0.0023}$	$0.9337_{\pm 0.0042}$	$0.8891_{\pm 0.0006}$
	0.05	$0.9487_{\pm 0.0049}$	$0.9175_{\pm 0.0009}$	$0.9345_{\pm 0.0054}$	$0.8901_{\pm 0.0011}$
	0.50	$0.9607_{\pm 0.0018}$	$0.9267_{\pm 0.0015}$	$0.9448_{\pm 0.0012}$	$0.8930_{\pm 0.0009}$
Mahalanobis	0.00 (AdamW)	$0.9782_{\pm 0.0010}$	$0.9505_{\pm 0.0009}$	$0.9621_{\pm 0.0016}$	$0.9222_{\pm 0.0010}$
	0.05	$0.9762_{\pm 0.0039}$	$0.9492_{\pm 0.0008}$	$0.9613_{\pm 0.0038}$	$0.9207_{\pm 0.0005}$
	0.50	$0.9787_{\pm 0.0035}$	$0.9498_{\pm 0.0025}$	$0.9595_{\pm 0.0044}$	$0.9167_{\pm 0.0019}$
DDU	0.00 (AdamW)	$0.9677_{\pm 0.0027}$	$0.9182_{\pm 0.0009}$	$0.9429_{\pm 0.0016}$	$0.8981_{\pm 0.0025}$
	0.05	$0.9630_{\pm 0.0028}$	$0.9150_{\pm 0.0022}$	$0.9407_{\pm 0.0008}$	$0.8952_{\pm 0.0004}$
	0.50	$0.9566_{\pm 0.0054}$	$0.9076_{\pm 0.0026}$	$0.9280_{\pm 0.0052}$	$0.8827_{\pm 0.0007}$

We use the ImageNet-200 experiment to test whether heterogeneous downstream responses to optimization extend beyond the primary setting, rather than as evidence that the covariance-weighting mechanism isolated on CIFAR-10 is reproduced unchanged.

F Additional Diagnostic and Robustness Checks

F.1 L_2 Feature Normalization Control

Müller & Hein (2025) show that post-hoc L_2 feature normalization improves conventional Mahalanobis OOD detection across most of the models they evaluate, although some model settings remain exceptions. Our purpose here is different: we use normalization not as an improved detector, but as a diagnostic for whether the training-conditioned degradation observed along the SAM trajectory is sensitive to feature magnitude.

To test whether the observed covariance-sensitive degradation can be explained primarily by differences in feature magnitude, we additionally apply post-hoc L_2 normalization to the penultimate representations before computing the corresponding OOD scores. For a representation $f(x)$, we use

$$\hat{f}(x) = \frac{f(x)}{\max(\|f(x)\|_2, \varepsilon)}, \tag{66}$$

where $\varepsilon = 10^{-12}$.

We normalize both the ID training features used to fit the detector and the ID/OOD evaluation features, and then refit the corresponding Mahalanobis or DDU model in the normalized feature space. We use this experiment only as a diagnostic control: normalization removes per-sample radial scale variation while preserving each feature's direction from the origin, allowing us to assess the sensitivity of detector behavior to feature magnitude.

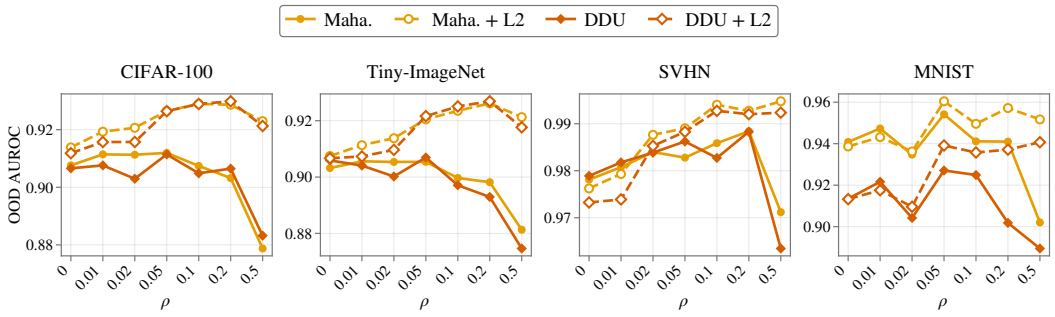

Figure 5: L_2 feature-normalization control for Mahalanobis and DDU OOD readouts in the CIFAR-10 / WideResNet-28-10 setting. Each panel follows the same OOD ordering and visual style as Figure 1, and compares the original detector with its post-hoc L_2-normalized counterpart.

Figure 5 shows that L_2 normalization substantially changes the large-radius detector response. At $\rho = 0.5$, the normalized Mahalanobis and DDU readouts no longer exhibit the corresponding degradation relative to their normalized SGD baselines across the four OOD datasets. This indicates that feature magnitude is an important contributor to the observed covariance-sensitive response. However, the magnitude of the recovery is not uniform across OOD datasets or detectors, indicating that detector behavior also depends on other aspects of the representation and the corresponding covariance-based readout. We use L_2 normalization only as a diagnostic control rather than as a remedy or an alternative detection method.

F.2 Feature Normality Diagnostics

Because Mahalanobis and Gaussian-density-based detectors use second-order distributional structure, we additionally examine whether the degradation observed under SAM coincides with a poorer held-out Gaussian fit. This diagnostic does not constitute a test of full multivariate Gaussianity; rather, it measures how well the fitted Gaussian reference describes independent ID features using the same active-subspace convention as the precision-geometry analysis.

For each model and SAM radius ρ, we fit the Gaussian reference on the seed-defined CIFAR-10 training split and evaluate it on a disjoint held-out ID split. We report two geometries: a tied within-

class covariance estimate and a macro-average over classwise covariance estimates. The diagnostics include the active dimension, Shapiro–Wilk statistic along retained directions, mean absolute excess kurtosis, radial Wasserstein distance per active dimension, and radial Kolmogorov–Smirnov distance. All covariance eigenspectra and inverse operations use FP64 linear algebra.

Table 14: Held-out Gaussian-reference fit diagnostics for WideResNet-28-10 trained on CIFAR-10. The Gaussian reference is fit on the seed-defined CIFAR-10 training split and evaluated on a disjoint held-out ID split. Shared covariance uses a tied within-class estimate, while class-wise covariance macro-averages per-class fits. Values are mean ± standard deviation over three seeds, using FP64 linear algebra.

ρ	Covariance type	Active dim.	Directional Shapiro $W \uparrow$	Abs. excess kurt. $\downarrow$	Radial W_1 / dim. $\downarrow$	Radial KS dist. $\downarrow$
0.00	Shared	$161.3_{\pm 15.8}$	$0.9783_{\pm 0.0014}$	$2.0857_{\pm 0.0689}$	$0.4990_{\pm 0.0194}$	$0.3160_{\pm 0.0115}$
	Class-wise	$159.7_{\pm 15.9}$	$0.9381_{\pm 0.0142}$	$5.2774_{\pm 1.4610}$	$0.9995_{\pm 0.1891}$	$0.3694_{\pm 0.0054}$
0.01	Shared	$162.0_{\pm 25.2}$	$0.9785_{\pm 0.0032}$	$2.0479_{\pm 0.1875}$	$0.4923_{\pm 0.0462}$	$0.3168_{\pm 0.0069}$
	Class-wise	$159.3_{\pm 24.3}$	$0.9465_{\pm 0.0084}$	$4.3732_{\pm 0.5461}$	$0.9173_{\pm 0.0922}$	$0.3748_{\pm 0.0067}$
0.02	Shared	$140.3_{\pm 21.0}$	$0.9792_{\pm 0.0021}$	$1.9874_{\pm 0.1793}$	$0.4941_{\pm 0.0072}$	$0.3167_{\pm 0.0099}$
	Class-wise	$138.2_{\pm 22.0}$	$0.9492_{\pm 0.0008}$	$4.1016_{\pm 0.1885}$	$0.9057_{\pm 0.0156}$	$0.3678_{\pm 0.0081}$
0.05	Shared	$153.0_{\pm 20.3}$	$0.9858_{\pm 0.0021}$	$1.4490_{\pm 0.1617}$	$0.4430_{\pm 0.0297}$	$0.3337_{\pm 0.0113}$
	Class-wise	$150.1_{\pm 19.5}$	$0.9556_{\pm 0.0162}$	$3.6037_{\pm 1.4109}$	$0.8831_{\pm 0.2134}$	$0.3691_{\pm 0.0092}$
0.10	Shared	$147.7_{\pm 15.6}$	$0.9894_{\pm 0.0025}$	$1.1984_{\pm 0.2378}$	$0.3783_{\pm 0.0516}$	$0.3237_{\pm 0.0091}$
	Class-wise	$146.0_{\pm 14.9}$	$0.9685_{\pm 0.0080}$	$2.5236_{\pm 0.6843}$	$0.7326_{\pm 0.1059}$	$0.3745_{\pm 0.0047}$
0.20	Shared	$150.7_{\pm 18.2}$	$0.9919_{\pm 0.0012}$	$0.9817_{\pm 0.1175}$	$0.3127_{\pm 0.0208}$	$0.3020_{\pm 0.0051}$
	Class-wise	$150.0_{\pm 17.6}$	$0.9686_{\pm 0.0046}$	$2.4699_{\pm 0.2852}$	$0.7485_{\pm 0.0611}$	$0.3712_{\pm 0.0127}$
0.50	Shared	$117.7_{\pm 5.1}$	$0.9890_{\pm 0.0013}$	$1.2146_{\pm 0.1471}$	$0.2527_{\pm 0.0048}$	$0.2496_{\pm 0.0086}$
	Class-wise	$116.3_{\pm 4.2}$	$0.9680_{\pm 0.0002}$	$2.5598_{\pm 0.0286}$	$0.6750_{\pm 0.0121}$	$0.3536_{\pm 0.0054}$

Table 14 shows that the large-radius SAM settings do not produce a uniformly worse held-out Gaussian reference fit. In the shared-covariance geometry, radial distances and Shapiro statistics generally move toward closer agreement with the Gaussian reference as ρ increases. Classwise fits show a weaker and more variable pattern, but they likewise do not support the interpretation that the covariance-sensitive OOD degradation is caused simply by increasingly non-Gaussian ID features.

F.3 Extended-Training SGD Control

SAM requires an additional gradient evaluation per parameter update, so equal training epochs do not correspond to equal optimization compute. To control for this difference, we additionally train SGD for 700 epochs. Since a SAM update requires approximately two gradient evaluations, this experiment approximately matches the gradient-evaluation count of 350-epoch SAM.

For the 700-epoch SGD control, the learning-rate decay milestones are proportionally extended from epochs 150/250 to 300/500, while all other hyperparameters are kept identical to the SGD baseline.

Table 15: Extended-training SGD control on WideResNet-28-10 trained on CIFAR-10, with CIFAR-100 as OOD. SGD trained for 700 epochs approximately matches the gradient-evaluation budget of 350-epoch SAM. OOD detection performance is reported in AUROC.

Optimizer	Epoch	Acc.	NC1	WithinVar	InterDist	MSP	DDU	Maha.
SGD	350	$0.9597_{\pm 0.0023}$	$0.0221_{\pm 0.0009}$	$0.8655_{\pm 0.0244}$	$7.3106_{\pm 0.3599}$	$0.8965_{\pm 0.0022}$	$0.9066_{\pm 0.0024}$	$0.9076_{\pm 0.0035}$
SGD	700	$0.9607_{\pm 0.0007}$	$0.0117_{\pm 0.0006}$	$0.4425_{\pm 0.0218}$	$7.3751_{\pm 0.0070}$	$0.8962_{\pm 0.0013}$	$0.9109_{\pm 0.0017}$	$0.9135_{\pm 0.0009}$
SAM ($\rho = 0.5$)	350	$0.9681_{\pm 0.0012}$	$0.0639_{\pm 0.0004}$	$2.4303_{\pm 0.0142}$	$6.8398_{\pm 0.0089}$	$0.9287_{\pm 0.0014}$	$0.8832_{\pm 0.0058}$	$0.8787_{\pm 0.0023}$

Table 15 reports the results on WideResNet-28-10 trained on CIFAR-10, with CIFAR-100 used as OOD data. Extended SGD does not reproduce the representation changes or detector behavior observed under SAM. Longer SGD training further reduces NC1 and within-class variance while maintaining or improving Mahalanobis performance. In contrast, SAM exhibits increased within-class variation together with degraded covariance-based OOD detection, despite improved predictive performance and calibration.

These results indicate that the additional gradient evaluations required by SAM are not sufficient, by themselves, to explain the observed representation–detector mismatch.

Table 16: Effect of SAM batch size on OOD detection performance for WideResNet-28-10 trained on CIFAR-10, with CIFAR-100 used as OOD data. Values are mean ± standard deviation over three seeds.

Model	Batch Size	Metric	ρ		
			0	0.05	0.50
WideResNet-28-10	64	DDU	$0.8788_{\pm 0.0081}$	$0.8796_{\pm 0.0035}$	$0.8349_{\pm 0.0063}$
		Maha	$0.8812_{\pm 0.0037}$	$0.8730_{\pm 0.0044}$	$0.7583_{\pm 0.0112}$
	128	DDU	$0.9066_{\pm 0.0024}$	$0.9114_{\pm 0.0042}$	$0.8832_{\pm 0.0058}$
		Maha	$0.9076_{\pm 0.0035}$	$0.9119_{\pm 0.0030}$	$0.8787_{\pm 0.0023}$
	256	DDU	$0.9077_{\pm 0.0023}$	$0.9079_{\pm 0.0045}$	$0.8944_{\pm 0.0094}$
		Maha	$0.8921_{\pm 0.0026}$	$0.8938_{\pm 0.0018}$	$0.8960_{\pm 0.0031}$

Table 17: Effect of SAM batch size on representation geometry for WideResNet-28-10 trained on CIFAR-10. Values are mean ± standard deviation over three seeds.

Model	Batch Size	Metric	ρ		
			0	0.05	0.50
WideResNet-28-10	64	WithinVar	$1.7279_{\pm 0.0569}$	$2.1676_{\pm 0.0119}$	$3.7911_{\pm 0.0715}$
		InterDist	$7.1517_{\pm 0.0128}$	$6.9931_{\pm 0.0485}$	$5.7908_{\pm 0.0186}$
		$\mathcal{NC}_1$	$0.0488_{\pm 0.0017}$	$0.0634_{\pm 0.0012}$	$0.1293_{\pm 0.0015}$
	128	WithinVar	$0.8655_{\pm 0.0244}$	$1.2640_{\pm 0.1631}$	$2.4303_{\pm 0.0142}$
		InterDist	$7.3106_{\pm 0.3599}$	$7.1803_{\pm 0.3163}$	$6.8398_{\pm 0.0089}$
		$\mathcal{NC}_1$	$0.0221_{\pm 0.0009}$	$0.0342_{\pm 0.0044}$	$0.0639_{\pm 0.0004}$
	256	WithinVar	$0.4801_{\pm 0.0537}$	$0.9091_{\pm 0.0590}$	$1.5964_{\pm 0.0136}$
		InterDist	$7.6930_{\pm 0.0834}$	$7.7524_{\pm 0.0177}$	$7.4726_{\pm 0.0665}$
		$\mathcal{NC}_1$	$0.0120_{\pm 0.0014}$	$0.0232_{\pm 0.0019}$	$0.0364_{\pm 0.0013}$

F.4 Batch-Size Sensitivity under SAM

We additionally examine whether the geometric changes and OOD detection trends observed under SAM depend critically on the default training batch size. We do not use this experiment to claim that SAM is independent of batch size; instead, we test whether the main qualitative tendencies remain visible when the training batch size is changed.

F.4.1 Experimental Protocol

The main experiments use a batch size of 128. We additionally evaluate batch sizes of 64 and 256 using WideResNet-28-10 and three representative optimization settings: SGD ($\rho = 0$), SAM with $\rho = 0.05$, and SAM with $\rho = 0.5$.

The number of training epochs and nominal optimization hyperparameters are kept fixed. Consequently, changing the batch size also changes the total number of parameter updates. This experiment should therefore be interpreted as a robustness check across batch-size configurations rather than as an update-matched or compute-matched comparison.

All results are averaged over three random seeds.

F.4.2 Observations

Tables 16 and 17 show that the qualitative trends observed in the main experiments are visible across the tested batch sizes, but not every detector degradation is reproduced in every batch-size setting.

For OOD detection, batch sizes 64 and 128 reproduce the large-radius degradation for both DDU and Mahalanobis. At batch size 256, however, Mahalanobis degradation is not reproduced: its AUROC

changes from 0.8921 under SGD to 0.8960 at $\rho = 0.5$, while DDU still decreases from 0.9077 to 0.8944. Thus, batch size can affect whether the Mahalanobis response appears as a degradation, even when the broader large-radius DDU trend remains visible.

The representation geometry exhibits a similar overall tendency. As ρ increases, WithinVar and $\mathrm{NC1}$ generally increase across the tested batch sizes, indicating greater within-class dispersion relative to between-class structure. InterDist also changes with ρ, although it is not strictly monotonic in every configuration.

The relevant observation is therefore qualitative rather than universal or strictly monotonic: the representation-geometry trends are not restricted to the default batch size of 128, while the covariance-sensitive detector response remains batch-size dependent. We consequently interpret this experiment as evidence that the default batch-size choice is not the sole explanation for the representation changes, rather than as evidence of batch-size-invariant detector degradation.

G Extended Quantitative Results

G.1 Cross-OOD Validation of the Covariance-Weighting Analysis

We repeat the covariance-weighting intervention from Figure 2(a) across CIFAR-100, Tiny-ImageNet, SVHN, and MNIST as OOD datasets, using the same CIFAR-10 WRN-28-10 checkpoints. Figure 6 shows the OOD AUROC response as the covariance exponent α varies from the unweighted endpoint to Mahalanobis weighting.

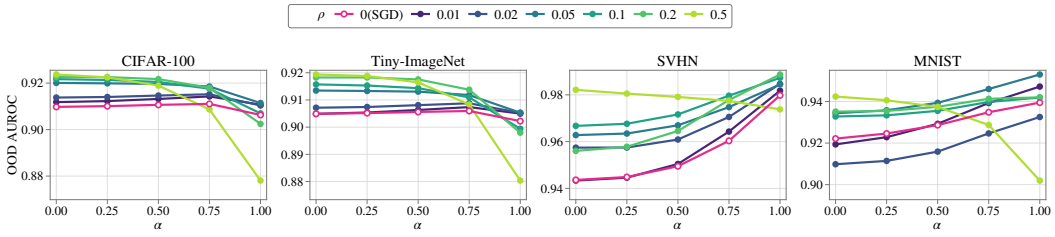

Figure 6: Cross-OOD replication of the covariance-weighting intervention for CIFAR-10 WRN-28-10. Each panel reports OOD AUROC as α varies from the unweighted endpoint ($\alpha = 0$) to Mahalanobis weighting ($\alpha = 1$).

Figure 6 shows that the effect of covariance weighting is OOD-dependent but the large-radius SAM setting is consistently sensitive to increasing covariance weighting.

Figures 7–9 show the corresponding directional-contrast, Mahalanobis-prefix, and DDU-prefix analyses across the four OOD datasets. The Mahalanobis prefixes reproduce the divergence between covariance-weighted and unweighted accumulation, while the DDU prefixes show that related cumulative covariance-sensitive behavior also appears in the class-wise Gaussian readout.

Figure 10 reports the corresponding covariance–contrast mismatch prevalence across α. The prevalence increases as covariance weighting strengthens, and SAM with large ρ remains above SGD across the OOD datasets, matching the direction of the CIFAR-100 result in the main text.

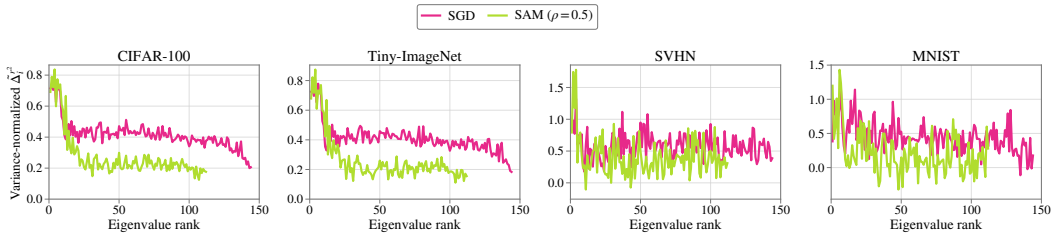

Figure 7: Cross-OOD replication of the directional contrast panel for CIFAR-10 WRN-28-10.

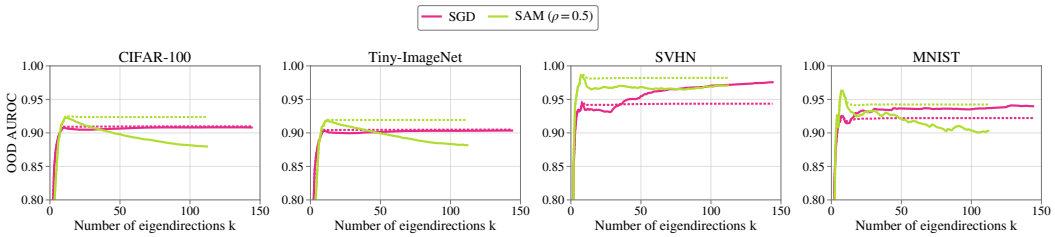

Figure 8: Cross-OOD replication of the prefix AUROC panel for CIFAR-10 WRN-28-10. Solid and dotted lines denote Mahalanobis-weighted ($\alpha = 1$) and unweighted ($\alpha = 0$) prefixes, respectively.

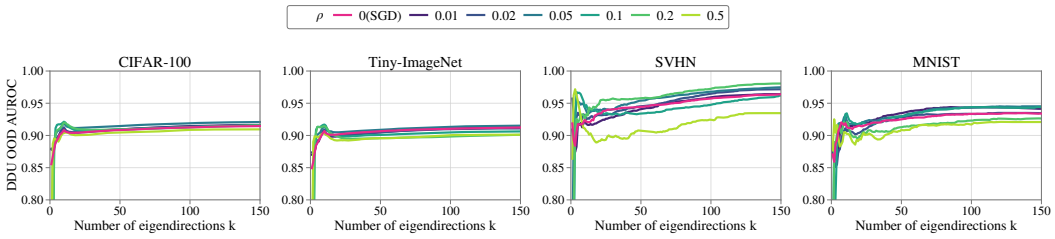

Figure 9: Cross-OOD replication of the DDU prefix AUROC analysis for CIFAR-10 WRN-28-10.

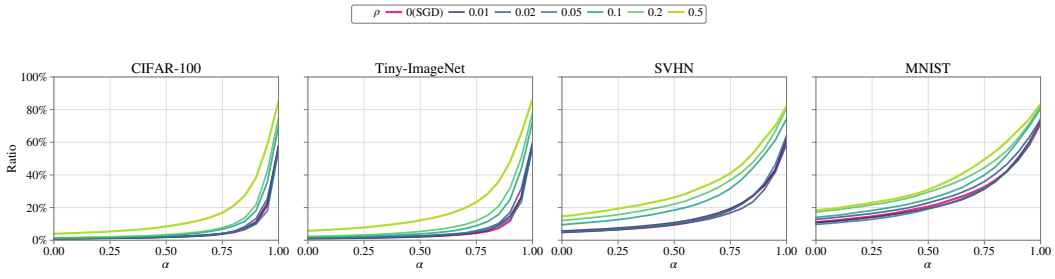

Figure 10: α sweep of the covariance–contrast mismatch prevalence across OOD datasets for CIFAR-10 WRN-28-10. Each panel reports the percentage of direction pairs satisfying the positive-contrast and ordering conditions for which Eq. 6 holds as the covariance exponent α varies.

G.2 Extended Spectral-Normalization and Backbone Results

We additionally report the corresponding CIFAR-10 results across WideResNet-28-10, ResNet-18, and VGG-16, with and without spectral normalization, to examine sensitivity to backbone and spectral normalization.

In the following tables, SN indicates whether spectral normalization is used: O denotes enabled and X denotes disabled. Unless otherwise stated, entries in this subsection are means over three independent runs. Entries shown as — indicate unavailable geometric statistics because the corresponding covariance-spectrum computation did not yield a finite reported value for that checkpoint setting.

Table 18: Accuracy and calibration metrics on CIFAR-10 under backbone and spectral-normalization settings. Values are mean ± standard deviation over three seeds.

Metric	ρ	WideResNet-28-10		ResNet-18		VGG-16	
		SNO	SNX	SNO	SNX	SNO	SNX
Accuracy	0.00	$0.9597_{\pm 0.0023}$	$0.9588_{\pm 0.0012}$	$0.9516_{\pm 0.0018}$	$0.9506_{\pm 0.0008}$	$0.9353_{\pm 0.0017}$	$0.9351_{\pm 0.0009}$
	0.01	$0.9622_{\pm 0.0009}$	$0.9596_{\pm 0.0003}$	$0.9541_{\pm 0.0022}$	$0.9546_{\pm 0.0021}$	$0.9398_{\pm 0.0026}$	$0.9398_{\pm 0.0020}$
	0.02	$0.9618_{\pm 0.0020}$	$0.9630_{\pm 0.0006}$	$0.9557_{\pm 0.0008}$	$0.9561_{\pm 0.0013}$	$0.9419_{\pm 0.0014}$	$0.9424_{\pm 0.0002}$
	0.05	$0.9666_{\pm 0.0002}$	$0.9653_{\pm 0.0005}$	$0.9579_{\pm 0.0007}$	$0.9567_{\pm 0.0013}$	$0.9426_{\pm 0.0015}$	$0.9438_{\pm 0.0022}$
	0.10	$0.9659_{\pm 0.0005}$	$0.9665_{\pm 0.0005}$	$0.9583_{\pm 0.0014}$	$0.9584_{\pm 0.0004}$	$0.9461_{\pm 0.0009}$	$0.9447_{\pm 0.0007}$
	0.20	$0.9680_{\pm 0.0018}$	$0.9673_{\pm 0.0020}$	$0.9595_{\pm 0.0010}$	$0.9569_{\pm 0.0014}$	$0.9439_{\pm 0.0013}$	$0.9441_{\pm 0.0016}$
	0.50	$0.9681_{\pm 0.0012}$	$0.9672_{\pm 0.0015}$	$0.9541_{\pm 0.0011}$	$0.9542_{\pm 0.0009}$	$0.9347_{\pm 0.0016}$	$0.9367_{\pm 0.0031}$
	0.60	$0.9667_{\pm 0.0008}$	$0.9652_{\pm 0.0010}$	$0.9523_{\pm 0.0023}$	$0.9522_{\pm 0.0017}$	$0.9321_{\pm 0.0014}$	$0.9323_{\pm 0.0022}$
	0.70	$0.9652_{\pm 0.0008}$	$0.9643_{\pm 0.0012}$	$0.9460_{\pm 0.0033}$	$0.9478_{\pm 0.0017}$	$0.9266_{\pm 0.0018}$	$0.9272_{\pm 0.0020}$
ECE	0.00	$0.0241_{\pm 0.0029}$	$0.0232_{\pm 0.0003}$	$0.0243_{\pm 0.0006}$	$0.0253_{\pm 0.0010}$	$0.0481_{\pm 0.0016}$	$0.0489_{\pm 0.0006}$
	0.01	$0.0189_{\pm 0.0010}$	$0.0218_{\pm 0.0008}$	$0.0207_{\pm 0.0024}$	$0.0205_{\pm 0.0007}$	$0.0419_{\pm 0.0019}$	$0.0422_{\pm 0.0021}$
	0.02	$0.0194_{\pm 0.0013}$	$0.0186_{\pm 0.0009}$	$0.0186_{\pm 0.0015}$	$0.0176_{\pm 0.0012}$	$0.0353_{\pm 0.0008}$	$0.0342_{\pm 0.0003}$
	0.05	$0.0146_{\pm 0.0003}$	$0.0143_{\pm 0.0009}$	$0.0151_{\pm 0.0004}$	$0.0157_{\pm 0.0008}$	$0.0291_{\pm 0.0025}$	$0.0293_{\pm 0.0021}$
	0.10	$0.0118_{\pm 0.0016}$	$0.0116_{\pm 0.0012}$	$0.0136_{\pm 0.0012}$	$0.0136_{\pm 0.0016}$	$0.0242_{\pm 0.0023}$	$0.0255_{\pm 0.0013}$
	0.20	$0.0092_{\pm 0.0016}$	$0.0097_{\pm 0.0012}$	$0.0117_{\pm 0.0001}$	$0.0141_{\pm 0.0009}$	$0.0230_{\pm 0.0008}$	$0.0224_{\pm 0.0014}$
	0.50	$0.0097_{\pm 0.0010}$	$0.0097_{\pm 0.0013}$	$0.0119_{\pm 0.0019}$	$0.0143_{\pm 0.0015}$	$0.0202_{\pm 0.0034}$	$0.0194_{\pm 0.0043}$
	0.60	$0.0089_{\pm 0.0018}$	$0.0098_{\pm 0.0003}$	$0.0081_{\pm 0.0031}$	$0.0149_{\pm 0.0054}$	$0.0151_{\pm 0.0064}$	$0.0167_{\pm 0.0021}$
	0.70	$0.0075_{\pm 0.0011}$	$0.0084_{\pm 0.0012}$	$0.0048_{\pm 0.0013}$	$0.0053_{\pm 0.0009}$	$0.0155_{\pm 0.0035}$	$0.0159_{\pm 0.0023}$
T-ECE	0.00	$0.0088_{\pm 0.0027}$	$0.0081_{\pm 0.0026}$	$0.0089_{\pm 0.0020}$	$0.0105_{\pm 0.0009}$	$0.0176_{\pm 0.0014}$	$0.0187_{\pm 0.0027}$
	0.01	$0.0063_{\pm 0.0022}$	$0.0100_{\pm 0.0031}$	$0.0066_{\pm 0.0021}$	$0.0087_{\pm 0.0025}$	$0.0186_{\pm 0.0035}$	$0.0158_{\pm 0.0015}$
	0.02	$0.0069_{\pm 0.0027}$	$0.0076_{\pm 0.0005}$	$0.0087_{\pm 0.0032}$	$0.0090_{\pm 0.0024}$	$0.0154_{\pm 0.0020}$	$0.0133_{\pm 0.0013}$
	0.05	$0.0079_{\pm 0.0004}$	$0.0068_{\pm 0.0023}$	$0.0065_{\pm 0.0004}$	$0.0066_{\pm 0.0009}$	$0.0131_{\pm 0.0020}$	$0.0114_{\pm 0.0014}$
	0.10	$0.0056_{\pm 0.0008}$	$0.0075_{\pm 0.0016}$	$0.0061_{\pm 0.0015}$	$0.0061_{\pm 0.0017}$	$0.0101_{\pm 0.0017}$	$0.0112_{\pm 0.0028}$
	0.20	$0.0053_{\pm 0.0008}$	$0.0045_{\pm 0.0022}$	$0.0065_{\pm 0.0010}$	$0.0088_{\pm 0.0017}$	$0.0113_{\pm 0.0045}$	$0.0113_{\pm 0.0037}$
	0.50	$0.0057_{\pm 0.0020}$	$0.0062_{\pm 0.0004}$	$0.0080_{\pm 0.0021}$	$0.0076_{\pm 0.0029}$	$0.0107_{\pm 0.0011}$	$0.0091_{\pm 0.0006}$
	0.60	$0.0044_{\pm 0.0005}$	$0.0045_{\pm 0.0013}$	$0.0062_{\pm 0.0009}$	$0.0064_{\pm 0.0022}$	$0.0092_{\pm 0.0032}$	$0.0085_{\pm 0.0015}$
	0.70	$0.0037_{\pm 0.0012}$	$0.0055_{\pm 0.0011}$	$0.0059_{\pm 0.0018}$	$0.0059_{\pm 0.0007}$	$0.0120_{\pm 0.0035}$	$0.0081_{\pm 0.0011}$

Table 19: Performance on CIFAR-10 (ID) with CIFAR-100 as the OOD dataset under spectral normalization settings. Values are mean ± standard deviation over three seeds.

Method	ρ	Metric	WideResNet-28-10		ResNet-18		VGG-16	
			SNO	SNX	SNO	SNX	SNO	SNX
MSP	0.00	AUROC	$0.8965_{\pm 0.0022}$	$0.8934_{\pm 0.0017}$	$0.8833_{\pm 0.0007}$	$0.8869_{\pm 0.0006}$	$0.8249_{\pm 0.0044}$	$0.8238_{\pm 0.0037}$
		FPR@95	$0.5091_{\pm 0.0072}$	$0.5410_{\pm 0.0132}$	$0.5729_{\pm 0.0102}$	$0.5696_{\pm 0.0031}$	$0.6473_{\pm 0.0137}$	$0.6464_{\pm 0.0090}$
	0.01	AUROC	$0.9018_{\pm 0.0028}$	$0.9028_{\pm 0.0017}$	$0.8921_{\pm 0.0040}$	$0.8934_{\pm 0.0015}$	$0.8524_{\pm 0.0072}$	$0.8539_{\pm 0.0016}$
		FPR@95	$0.5209_{\pm 0.0199}$	$0.5107_{\pm 0.0133}$	$0.5608_{\pm 0.0164}$	$0.5520_{\pm 0.0029}$	$0.6488_{\pm 0.0189}$	$0.6405_{\pm 0.0079}$
	0.02	AUROC	$0.9051_{\pm 0.0012}$	$0.9046_{\pm 0.0027}$	$0.8984_{\pm 0.0020}$	$0.8972_{\pm 0.0026}$	$0.8749_{\pm 0.0025}$	$0.8771_{\pm 0.0047}$
		FPR@95	$0.5046_{\pm 0.0133}$	$0.5076_{\pm 0.0059}$	$0.5385_{\pm 0.0013}$	$0.5578_{\pm 0.0087}$	$0.6466_{\pm 0.0082}$	$0.6497_{\pm 0.0111}$
	0.05	AUROC	$0.9172_{\pm 0.0023}$	$0.9161_{\pm 0.0007}$	$0.9072_{\pm 0.0009}$	$0.9065_{\pm 0.0016}$	$0.8856_{\pm 0.0021}$	$0.8832_{\pm 0.0045}$
		FPR@95	$0.4752_{\pm 0.0204}$	$0.4832_{\pm 0.0025}$	$0.5299_{\pm 0.0043}$	$0.5323_{\pm 0.0124}$	$0.6423_{\pm 0.0019}$	$0.6397_{\pm 0.0159}$
	0.10	AUROC	$0.9214_{\pm 0.0020}$	$0.9225_{\pm 0.0005}$	$0.9084_{\pm 0.0020}$	$0.9079_{\pm 0.0033}$	$0.8905_{\pm 0.0035}$	$0.8916_{\pm 0.0008}$
		FPR@95	$0.4710_{\pm 0.0063}$	$0.4699_{\pm 0.0026}$	$0.5319_{\pm 0.0090}$	$0.5330_{\pm 0.0131}$	$0.6176_{\pm 0.0175}$	$0.6148_{\pm 0.0073}$
	0.20	AUROC	$0.9250_{\pm 0.0019}$	$0.9254_{\pm 0.0023}$	$0.9111_{\pm 0.0036}$	$0.9092_{\pm 0.0024}$	$0.8904_{\pm 0.0019}$	$0.8902_{\pm 0.0033}$
		FPR@95	$0.4497_{\pm 0.0153}$	$0.4420_{\pm 0.0223}$	$0.5125_{\pm 0.0270}$	$0.5346_{\pm 0.0183}$	$0.5970_{\pm 0.0105}$	$0.5997_{\pm 0.0174}$
	0.50	AUROC	$0.9287_{\pm 0.0014}$	$0.9271_{\pm 0.0010}$	$0.9069_{\pm 0.0034}$	$0.9077_{\pm 0.0010}$	$0.8853_{\pm 0.0006}$	$0.8922_{\pm 0.0006}$
		FPR@95	$0.4298_{\pm 0.0165}$	$0.4378_{\pm 0.0036}$	$0.5324_{\pm 0.0159}$	$0.5340_{\pm 0.0128}$	$0.6203_{\pm 0.0143}$	$0.5906_{\pm 0.0084}$
	0.60	AUROC	$0.9274_{\pm 0.0007}$	$0.9245_{\pm 0.0022}$	$0.9053_{\pm 0.0014}$	$0.9051_{\pm 0.0039}$	$0.8763_{\pm 0.0117}$	$0.8808_{\pm 0.0040}$
		FPR@95	$0.4451_{\pm 0.0057}$	$0.4574_{\pm 0.0103}$	$0.5384_{\pm 0.0028}$	$0.5384_{\pm 0.0175}$	$0.6197_{\pm 0.0015}$	$0.6141_{\pm 0.0141}$
	0.70	AUROC	$0.9262_{\pm 0.0016}$	$0.9241_{\pm 0.0022}$	$0.9024_{\pm 0.0048}$	$0.9017_{\pm 0.0022}$	$0.8715_{\pm 0.0063}$	$0.8827_{\pm 0.0038}$
		FPR@95	$0.4577_{\pm 0.0062}$	$0.4767_{\pm 0.0136}$	$0.5527_{\pm 0.0178}$	$0.5541_{\pm 0.0041}$	$0.6273_{\pm 0.0100}$	$0.6059_{\pm 0.0142}$
Energy	0.00	AUROC	$0.8969_{\pm 0.0042}$	$0.8938_{\pm 0.0027}$	$0.8798_{\pm 0.0014}$	$0.8852_{\pm 0.0017}$	$0.8170_{\pm 0.0056}$	$0.8151_{\pm 0.0046}$
		FPR@95	$0.4357_{\pm 0.0150}$	$0.4457_{\pm 0.0057}$	$0.4894_{\pm 0.0081}$	$0.4895_{\pm 0.0108}$	$0.5653_{\pm 0.0032}$	$0.5709_{\pm 0.0110}$
	0.01	AUROC	$0.9040_{\pm 0.0027}$	$0.9044_{\pm 0.0014}$	$0.8912_{\pm 0.0034}$	$0.8914_{\pm 0.0031}$	$0.8503_{\pm 0.0098}$	$0.8520_{\pm 0.0035}$
		FPR@95	$0.4286_{\pm 0.0077}$	$0.4274_{\pm 0.0055}$	$0.4771_{\pm 0.0049}$	$0.4720_{\pm 0.0064}$	$0.5528_{\pm 0.0121}$	$0.5437_{\pm 0.0087}$
	0.02	AUROC	$0.9082_{\pm 0.0005}$	$0.9071_{\pm 0.0036}$	$0.8995_{\pm 0.0019}$	$0.8985_{\pm 0.0033}$	$0.8821_{\pm 0.0035}$	$0.8864_{\pm 0.0086}$
		FPR@95	$0.4100_{\pm 0.0088}$	$0.4115_{\pm 0.0068}$	$0.4483_{\pm 0.0028}$	$0.4520_{\pm 0.0175}$	$0.5365_{\pm 0.0072}$	$0.5250_{\pm 0.0179}$
	0.05	AUROC	$0.9216_{\pm 0.0034}$	$0.9220_{\pm 0.0014}$	$0.9095_{\pm 0.0018}$	$0.9114_{\pm 0.0017}$	$0.9006_{\pm 0.0046}$	$0.8975_{\pm 0.0054}$
		FPR@95	$0.3758_{\pm 0.0122}$	$0.3713_{\pm 0.0031}$	$0.4260_{\pm 0.0153}$	$0.4237_{\pm 0.0034}$	$0.5085_{\pm 0.0043}$	$0.5108_{\pm 0.0135}$
	0.10	AUROC	$0.9281_{\pm 0.0019}$	$0.9301_{\pm 0.0004}$	$0.9123_{\pm 0.0019}$	$0.9140_{\pm 0.0030}$	$0.9034_{\pm 0.0047}$	$0.9040_{\pm 0.0015}$
		FPR@95	$0.3623_{\pm 0.0100}$	$0.3559_{\pm 0.0032}$	$0.4097_{\pm 0.0148}$	$0.4137_{\pm 0.0051}$	$0.4910_{\pm 0.0172}$	$0.4892_{\pm 0.0082}$
	0.20	AUROC	$0.9308_{\pm 0.0020}$	$0.9316_{\pm 0.0016}$	$0.9163_{\pm 0.0026}$	$0.9200_{\pm 0.0016}$	$0.9020_{\pm 0.0027}$	$0.9016_{\pm 0.0034}$
		FPR@95	$0.3454_{\pm 0.0081}$	$0.3404_{\pm 0.0142}$	$0.4024_{\pm 0.0093}$	$0.4098_{\pm 0.0099}$	$0.4977_{\pm 0.0155}$	$0.4962_{\pm 0.0147}$
	0.50	AUROC	$0.9364_{\pm 0.0015}$	$0.9357_{\pm 0.0011}$	$0.9176_{\pm 0.0021}$	$0.9196_{\pm 0.0022}$	$0.9021_{\pm 0.0002}$	$0.9063_{\pm 0.0006}$
		FPR@95	$0.3156_{\pm 0.0133}$	$0.3243_{\pm 0.0028}$	$0.4189_{\pm 0.0109}$	$0.4157_{\pm 0.0045}$	$0.4993_{\pm 0.0137}$	$0.4847_{\pm 0.0106}$
	0.60	AUROC	$0.9374_{\pm 0.0005}$	$0.9346_{\pm 0.0032}$	$0.9174_{\pm 0.0025}$	$0.9171_{\pm 0.0049}$	$0.8918_{\pm 0.0118}$	$0.8983_{\pm 0.0028}$
		FPR@95	$0.3215_{\pm 0.0079}$	$0.3336_{\pm 0.0116}$	$0.4224_{\pm 0.0097}$	$0.4239_{\pm 0.0251}$	$0.5299_{\pm 0.0214}$	$0.5143_{\pm 0.0078}$
	0.70	AUROC	$0.9371_{\pm 0.0019}$	$0.9355_{\pm 0.0017}$	$0.9159_{\pm 0.0056}$	$0.9152_{\pm 0.0022}$	$0.8889_{\pm 0.0090}$	$0.8986_{\pm 0.0035}$
		FPR@95	$0.3302_{\pm 0.0130}$	$0.3352_{\pm 0.0099}$	$0.4295_{\pm 0.0155}$	$0.4303_{\pm 0.0047}$	$0.5248_{\pm 0.0146}$	$0.5111_{\pm 0.0076}$
Mahalanobis	0.00	AUROC	$0.9076_{\pm 0.0035}$	$0.9071_{\pm 0.0024}$	$0.8726_{\pm 0.0024}$	$0.8736_{\pm 0.0036}$	$0.8790_{\pm 0.0028}$	$0.8745_{\pm 0.0017}$
		FPR@95	$0.4767_{\pm 0.0253}$	$0.4849_{\pm 0.0196}$	$0.6266_{\pm 0.0050}$	$0.6409_{\pm 0.0140}$	$0.6508_{\pm 0.0083}$	$0.6685_{\pm 0.0155}$
	0.01	AUROC	$0.9114_{\pm 0.0029}$	$0.9108_{\pm 0.0007}$	$0.8735_{\pm 0.0038}$	$0.8700_{\pm 0.0042}$	$0.8780_{\pm 0.0016}$	$0.8780_{\pm 0.0002}$
		FPR@95	$0.4619_{\pm 0.0118}$	$0.4579_{\pm 0.0083}$	$0.6316_{\pm 0.0112}$	$0.6494_{\pm 0.0184}$	$0.6528_{\pm 0.0069}$	$0.6527_{\pm 0.0075}$
	0.02	AUROC	$0.9113_{\pm 0.0042}$	$0.9120_{\pm 0.0037}$	$0.8642_{\pm 0.0040}$	$0.8645_{\pm 0.0114}$	$0.8778_{\pm 0.0008}$	$0.8784_{\pm 0.0063}$
		FPR@95	$0.4724_{\pm 0.0180}$	$0.4573_{\pm 0.0143}$	$0.6674_{\pm 0.0080}$	$0.6663_{\pm 0.0379}$	$0.6864_{\pm 0.0042}$	$0.6744_{\pm 0.0454}$
	0.05	AUROC	$0.9119_{\pm 0.0030}$	$0.9094_{\pm 0.0030}$	$0.8493_{\pm 0.0021}$	$0.8493_{\pm 0.0158}$	$0.8755_{\pm 0.0029}$	$0.8741_{\pm 0.0050}$
		FPR@95	$0.4769_{\pm 0.0043}$	$0.4897_{\pm 0.0025}$	$0.7021_{\pm 0.0076}$	$0.7024_{\pm 0.0456}$	$0.6716_{\pm 0.0116}$	$0.6798_{\pm 0.0254}$
	0.10	AUROC	$0.9074_{\pm 0.0068}$	$0.9073_{\pm 0.0026}$	$0.8338_{\pm 0.0068}$	$0.8352_{\pm 0.0209}$	$0.8660_{\pm 0.0105}$	$0.8768_{\pm 0.0043}$
		FPR@95	$0.4879_{\pm 0.0277}$	$0.5042_{\pm 0.0067}$	$0.7224_{\pm 0.0103}$	$0.7110_{\pm 0.0481}$	$0.6652_{\pm 0.0258}$	$0.6265_{\pm 0.0304}$
	0.20	AUROC	$0.9032_{\pm 0.0012}$	$0.9052_{\pm 0.0015}$	$0.8407_{\pm 0.0091}$	$0.8624_{\pm 0.0077}$	$0.8743_{\pm 0.0027}$	$0.8757_{\pm 0.0039}$
		FPR@95	$0.4903_{\pm 0.0063}$	$0.4720_{\pm 0.0028}$	$0.6989_{\pm 0.0194}$	$0.6362_{\pm 0.0353}$	$0.6379_{\pm 0.0094}$	$0.6349_{\pm 0.0304}$
	0.50	AUROC	$0.8787_{\pm 0.0023}$	$0.8720_{\pm 0.0033}$	$0.7925_{\pm 0.0008}$	$0.8023_{\pm 0.0061}$	$0.8514_{\pm 0.0060}$	$0.8491_{\pm 0.0128}$
		FPR@95	$0.5359_{\pm 0.0206}$	$0.5651_{\pm 0.0185}$	$0.7810_{\pm 0.0156}$	$0.7769_{\pm 0.0300}$	$0.7395_{\pm 0.0418}$	$0.7356_{\pm 0.0271}$
	0.60	AUROC	$0.8645_{\pm 0.0104}$	$0.8589_{\pm 0.0035}$	$0.7658_{\pm 0.0176}$	$0.7846_{\pm 0.0365}$	$0.8360_{\pm 0.0187}$	$0.8381_{\pm 0.0024}$
		FPR@95	$0.5849_{\pm 0.0400}$	$0.6070_{\pm 0.0061}$	$0.8146_{\pm 0.0238}$	$0.7886_{\pm 0.0184}$	$0.7479_{\pm 0.0338}$	$0.7298_{\pm 0.0124}$
	0.70	AUROC	$0.8360_{\pm 0.0066}$	$0.8423_{\pm 0.0082}$	$0.6997_{\pm 0.0031}$	$0.7294_{\pm 0.0145}$	$0.8426_{\pm 0.0103}$	$0.8345_{\pm 0.0119}$
		FPR@95	$0.6584_{\pm 0.0140}$	$0.6547_{\pm 0.0342}$	$0.8638_{\pm 0.0199}$	$0.8454_{\pm 0.0188}$	$0.7458_{\pm 0.0091}$	$0.7520_{\pm 0.0396}$
DDU	0.00	AUROC	$0.9066_{\pm 0.0024}$	$0.9010_{\pm 0.0039}$	$0.8987_{\pm 0.0021}$	$0.9014_{\pm 0.0028}$	$0.8610_{\pm 0.0108}$	$0.8625_{\pm 0.0029}$
		FPR@95	$0.4569_{\pm 0.0110}$	$0.4963_{\pm 0.0195}$	$0.5303_{\pm 0.0035}$	$0.5187_{\pm 0.0140}$	$0.6736_{\pm 0.0225}$	$0.6883_{\pm 0.0114}$
	0.01	AUROC	$0.9076_{\pm 0.0021}$	$0.9061_{\pm 0.0051}$	$0.9044_{\pm 0.0034}$	$0.9056_{\pm 0.0015}$	$0.8661_{\pm 0.0014}$	$0.8677_{\pm 0.0043}$
		FPR@95	$0.4608_{\pm 0.0069}$	$0.4677_{\pm 0.0184}$	$0.5252_{\pm 0.0168}$	$0.5194_{\pm 0.0175}$	$0.6734_{\pm 0.0178}$	$0.6849_{\pm 0.0017}$
	0.02	AUROC	$0.9030_{\pm 0.0106}$	$0.9060_{\pm 0.0061}$	$0.9042_{\pm 0.0008}$	$0.9037_{\pm 0.0025}$	$0.8707_{\pm 0.0068}$	$0.8715_{\pm 0.0076}$
		FPR@95	$0.4781_{\pm 0.0285}$	$0.4708_{\pm 0.0257}$	$0.5198_{\pm 0.0137}$	$0.5379_{\pm 0.0143}$	$0.7138_{\pm 0.0074}$	$0.7155_{\pm 0.0366}$
	0.05	AUROC	$0.9114_{\pm 0.0042}$	$0.9026_{\pm 0.0008}$	$0.9074_{\pm 0.0010}$	$0.9059_{\pm 0.0006}$	$0.8700_{\pm 0.0059}$	$0.8678_{\pm 0.0057}$
		FPR@95	$0.4474_{\pm 0.0216}$	$0.4906_{\pm 0.0122}$	$0.5177_{\pm 0.0108}$	$0.5244_{\pm 0.0175}$	$0.6805_{\pm 0.0217}$	$0.7000_{\pm 0.0179}$
	0.10	AUROC	$0.9048_{\pm 0.0009}$	$0.8963_{\pm 0.0075}$	$0.9046_{\pm 0.0026}$	$0.9048_{\pm 0.0005}$	$0.8608_{\pm 0.0096}$	$0.8764_{\pm 0.0088}$
		FPR@95	$0.4961_{\pm 0.0101}$	$0.5212_{\pm 0.0286}$	$0.5405_{\pm 0.0154}$	$0.5298_{\pm 0.0237}$	$0.6923_{\pm 0.0532}$	$0.6434_{\pm 0.0264}$
	0.20	AUROC	$0.9065_{\pm 0.0048}$	$0.9029_{\pm 0.0040}$	$0.9067_{\pm 0.0011}$	$0.9059_{\pm 0.0041}$	$0.8779_{\pm 0.0044}$	$0.8729_{\pm 0.0061}$
		FPR@95	$0.4935_{\pm 0.0116}$	$0.5077_{\pm 0.0225}$	$0.5152_{\pm 0.0219}$	$0.5059_{\pm 0.0205}$	$0.6115_{\pm 0.0275}$	$0.6608_{\pm 0.0548}$
	0.50	AUROC	$0.8832_{\pm 0.0058}$	$0.8850_{\pm 0.0004}$	$0.8938_{\pm 0.0038}$	$0.8921_{\pm 0.0038}$	$0.8599_{\pm 0.0008}$	$0.8620_{\pm 0.0054}$
		FPR@95	$0.5746_{\pm 0.0273}$	$0.5743_{\pm 0.0032}$	$0.5801_{\pm 0.0280}$	$0.5889_{\pm 0.0175}$	$0.7082_{\pm 0.0132}$	$0.6902_{\pm 0.0145}$
	0.60	AUROC	$0.8679_{\pm 0.0071}$	$0.8700_{\pm 0.0072}$	$0.8871_{\pm 0.0049}$	$0.8902_{\pm 0.0064}$	$0.8534_{\pm 0.0074}$	$0.8548_{\pm 0.0002}$
		FPR@95	$0.6242_{\pm 0.0273}$	$0.6251_{\pm 0.0231}$	$0.5929_{\pm 0.0222}$	$0.5811_{\pm 0.0166}$	$0.6986_{\pm 0.0117}$	$0.6850_{\pm 0.0031}$
	0.70	AUROC	$0.8467_{\pm 0.0128}$	$0.8490_{\pm 0.0068}$	$0.8687_{\pm 0.0017}$	$0.8702_{\pm 0.0183}$	$0.8549_{\pm 0.0072}$	$0.8542_{\pm 0.0069}$
		FPR@95	$0.6722_{\pm 0.0403}$	$0.6778_{\pm 0.0196}$	$0.6251_{\pm 0.0051}$	$0.6196_{\pm 0.0361}$	$0.7144_{\pm 0.0008}$	$0.6910_{\pm 0.0303}$

Table 20: Performance on CIFAR-10 (ID) with Tiny-ImageNet as the OOD dataset under spectral normalization settings. Values are mean ± standard deviation over three seeds.

Method	ρ	Metric	WideResNet-28-10		ResNet-18		VGG-16	
			SNO	SNX	SNO	SNX	SNO	SNX
MSP	0.00	AUROC	$0.8869_{\pm 0.0015}$	$0.8848_{\pm 0.0007}$	$0.8800_{\pm 0.0014}$	$0.8812_{\pm 0.0005}$	$0.8290_{\pm 0.0030}$	$0.8295_{\pm 0.0063}$
		FPR@95	$0.5030_{\pm 0.0040}$	$0.5262_{\pm 0.0139}$	$0.5591_{\pm 0.0119}$	$0.5530_{\pm 0.0099}$	$0.6304_{\pm 0.0177}$	$0.6192_{\pm 0.0137}$
	0.01	AUROC	$0.8892_{\pm 0.0009}$	$0.8929_{\pm 0.0012}$	$0.8859_{\pm 0.0034}$	$0.8875_{\pm 0.0013}$	$0.8534_{\pm 0.0067}$	$0.8539_{\pm 0.0017}$
		FPR@95	$0.5162_{\pm 0.0163}$	$0.5031_{\pm 0.0087}$	$0.5484_{\pm 0.0103}$	$0.5390_{\pm 0.0094}$	$0.6296_{\pm 0.0145}$	$0.6221_{\pm 0.0087}$
	0.02	AUROC	$0.8948_{\pm 0.0022}$	$0.8947_{\pm 0.0018}$	$0.8928_{\pm 0.0019}$	$0.8907_{\pm 0.0021}$	$0.8711_{\pm 0.0008}$	$0.8730_{\pm 0.0074}$
		FPR@95	$0.4979_{\pm 0.0138}$	$0.4938_{\pm 0.0047}$	$0.5257_{\pm 0.0067}$	$0.5445_{\pm 0.0024}$	$0.6336_{\pm 0.0054}$	$0.6429_{\pm 0.0097}$
	0.05	AUROC	$0.9081_{\pm 0.0011}$	$0.9064_{\pm 0.0008}$	$0.8999_{\pm 0.0007}$	$0.9002_{\pm 0.0009}$	$0.8818_{\pm 0.0021}$	$0.8789_{\pm 0.0042}$
		FPR@95	$0.4620_{\pm 0.0197}$	$0.4751_{\pm 0.0084}$	$0.5174_{\pm 0.0072}$	$0.5198_{\pm 0.0146}$	$0.6342_{\pm 0.0104}$	$0.6245_{\pm 0.0078}$
	0.10	AUROC	$0.9138_{\pm 0.0016}$	$0.9145_{\pm 0.0004}$	$0.9038_{\pm 0.0010}$	$0.9049_{\pm 0.0016}$	$0.8892_{\pm 0.0034}$	$0.8903_{\pm 0.0022}$
		FPR@95	$0.4523_{\pm 0.0047}$	$0.4534_{\pm 0.0055}$	$0.5129_{\pm 0.0017}$	$0.5147_{\pm 0.0084}$	$0.5918_{\pm 0.0138}$	$0.5894_{\pm 0.0112}$
	0.20	AUROC	$0.9180_{\pm 0.0020}$	$0.9179_{\pm 0.0012}$	$0.9076_{\pm 0.0045}$	$0.9034_{\pm 0.0021}$	$0.8916_{\pm 0.0015}$	$0.8898_{\pm 0.0008}$
		FPR@95	$0.4279_{\pm 0.0130}$	$0.4233_{\pm 0.0172}$	$0.4981_{\pm 0.0351}$	$0.5228_{\pm 0.0206}$	$0.5705_{\pm 0.0084}$	$0.5742_{\pm 0.0178}$
	0.50	AUROC	$0.9201_{\pm 0.0024}$	$0.9187_{\pm 0.0014}$	$0.9021_{\pm 0.0040}$	$0.9054_{\pm 0.0024}$	$0.8841_{\pm 0.0011}$	$0.8902_{\pm 0.0027}$
		FPR@95	$0.4210_{\pm 0.0196}$	$0.4276_{\pm 0.0097}$	$0.5240_{\pm 0.0222}$	$0.5114_{\pm 0.0082}$	$0.6012_{\pm 0.0214}$	$0.5684_{\pm 0.0105}$
	0.60	AUROC	$0.9195_{\pm 0.0013}$	$0.9165_{\pm 0.0030}$	$0.9007_{\pm 0.0028}$	$0.9015_{\pm 0.0059}$	$0.8733_{\pm 0.0163}$	$0.8777_{\pm 0.0029}$
		FPR@95	$0.4325_{\pm 0.0125}$	$0.4491_{\pm 0.0194}$	$0.5333_{\pm 0.0024}$	$0.5335_{\pm 0.0212}$	$0.5959_{\pm 0.0085}$	$0.5929_{\pm 0.0191}$
	0.70	AUROC	$0.9192_{\pm 0.0009}$	$0.9169_{\pm 0.0011}$	$0.8968_{\pm 0.0051}$	$0.8979_{\pm 0.0016}$	$0.8698_{\pm 0.0076}$	$0.8811_{\pm 0.0032}$
		FPR@95	$0.4450_{\pm 0.0050}$	$0.4614_{\pm 0.0139}$	$0.5509_{\pm 0.0199}$	$0.5507_{\pm 0.0076}$	$0.6046_{\pm 0.0128}$	$0.5833_{\pm 0.0132}$
Energy	0.00	AUROC	$0.8892_{\pm 0.0033}$	$0.8881_{\pm 0.0013}$	$0.8787_{\pm 0.0019}$	$0.8815_{\pm 0.0025}$	$0.8277_{\pm 0.0040}$	$0.8268_{\pm 0.0061}$
		FPR@95	$0.4137_{\pm 0.0072}$	$0.4202_{\pm 0.0019}$	$0.4714_{\pm 0.0077}$	$0.4687_{\pm 0.0124}$	$0.5322_{\pm 0.0103}$	$0.5287_{\pm 0.0177}$
	0.01	AUROC	$0.8929_{\pm 0.0014}$	$0.8966_{\pm 0.0016}$	$0.8868_{\pm 0.0026}$	$0.8870_{\pm 0.0025}$	$0.8562_{\pm 0.0089}$	$0.8567_{\pm 0.0040}$
		FPR@95	$0.4155_{\pm 0.0032}$	$0.4066_{\pm 0.0063}$	$0.4572_{\pm 0.0071}$	$0.4524_{\pm 0.0067}$	$0.5138_{\pm 0.0116}$	$0.5105_{\pm 0.0116}$
	0.02	AUROC	$0.9000_{\pm 0.0027}$	$0.8994_{\pm 0.0018}$	$0.8952_{\pm 0.0024}$	$0.8935_{\pm 0.0024}$	$0.8819_{\pm 0.0008}$	$0.8856_{\pm 0.0105}$
		FPR@95	$0.3941_{\pm 0.0100}$	$0.3904_{\pm 0.0034}$	$0.4256_{\pm 0.0078}$	$0.4300_{\pm 0.0132}$	$0.5033_{\pm 0.0075}$	$0.4971_{\pm 0.0140}$
	0.05	AUROC	$0.9155_{\pm 0.0014}$	$0.9147_{\pm 0.0004}$	$0.9059_{\pm 0.0015}$	$0.9086_{\pm 0.0004}$	$0.9010_{\pm 0.0059}$	$0.8964_{\pm 0.0048}$
		FPR@95	$0.3473_{\pm 0.0016}$	$0.3480_{\pm 0.0067}$	$0.4002_{\pm 0.0117}$	$0.3931_{\pm 0.0095}$	$0.4767_{\pm 0.0113}$	$0.4792_{\pm 0.0086}$
	0.10	AUROC	$0.9246_{\pm 0.0016}$	$0.9259_{\pm 0.0004}$	$0.9126_{\pm 0.0017}$	$0.9160_{\pm 0.0040}$	$0.9063_{\pm 0.0038}$	$0.9062_{\pm 0.0033}$
		FPR@95	$0.3262_{\pm 0.0097}$	$0.3213_{\pm 0.0036}$	$0.3723_{\pm 0.0098}$	$0.3740_{\pm 0.0121}$	$0.4457_{\pm 0.0159}$	$0.4514_{\pm 0.0074}$
	0.20	AUROC	$0.9287_{\pm 0.0027}$	$0.9287_{\pm 0.0011}$	$0.9193_{\pm 0.0034}$	$0.9189_{\pm 0.0025}$	$0.9047_{\pm 0.0027}$	$0.9028_{\pm 0.0004}$
		FPR@95	$0.3046_{\pm 0.0031}$	$0.3059_{\pm 0.0067}$	$0.3591_{\pm 0.0224}$	$0.3730_{\pm 0.0159}$	$0.4558_{\pm 0.0078}$	$0.4563_{\pm 0.0089}$
	0.50	AUROC	$0.9327_{\pm 0.0020}$	$0.9322_{\pm 0.0010}$	$0.9187_{\pm 0.0031}$	$0.9223_{\pm 0.0049}$	$0.9038_{\pm 0.0016}$	$0.9074_{\pm 0.0029}$
		FPR@95	$0.2901_{\pm 0.0101}$	$0.2950_{\pm 0.0073}$	$0.3790_{\pm 0.0204}$	$0.3669_{\pm 0.0198}$	$0.4554_{\pm 0.0166}$	$0.4405_{\pm 0.0219}$
	0.60	AUROC	$0.9351_{\pm 0.0007}$	$0.9317_{\pm 0.0045}$	$0.9197_{\pm 0.0048}$	$0.9189_{\pm 0.0077}$	$0.8928_{\pm 0.0147}$	$0.8990_{\pm 0.0026}$
		FPR@95	$0.2907_{\pm 0.0014}$	$0.3039_{\pm 0.0195}$	$0.3800_{\pm 0.0164}$	$0.3816_{\pm 0.0328}$	$0.4860_{\pm 0.0260}$	$0.4720_{\pm 0.0158}$
	0.70	AUROC	$0.9360_{\pm 0.0014}$	$0.9342_{\pm 0.0006}$	$0.9189_{\pm 0.0063}$	$0.9191_{\pm 0.0019}$	$0.8922_{\pm 0.0101}$	$0.9014_{\pm 0.0026}$
		FPR@95	$0.2926_{\pm 0.0061}$	$0.2934_{\pm 0.0052}$	$0.3860_{\pm 0.0212}$	$0.3835_{\pm 0.0021}$	$0.4824_{\pm 0.0214}$	$0.4608_{\pm 0.0059}$
Mahalanobis	0.00	AUROC	$0.9032_{\pm 0.0018}$	$0.9010_{\pm 0.0017}$	$0.8730_{\pm 0.0014}$	$0.8724_{\pm 0.0050}$	$0.8777_{\pm 0.0017}$	$0.8750_{\pm 0.0024}$
		FPR@95	$0.4718_{\pm 0.0229}$	$0.4768_{\pm 0.0109}$	$0.6234_{\pm 0.0063}$	$0.6415_{\pm 0.0167}$	$0.6493_{\pm 0.0110}$	$0.6580_{\pm 0.0187}$
	0.01	AUROC	$0.9056_{\pm 0.0012}$	$0.9044_{\pm 0.0010}$	$0.8723_{\pm 0.0034}$	$0.8693_{\pm 0.0017}$	$0.8783_{\pm 0.0035}$	$0.8763_{\pm 0.0013}$
		FPR@95	$0.4579_{\pm 0.0076}$	$0.4551_{\pm 0.0122}$	$0.6328_{\pm 0.0178}$	$0.6453_{\pm 0.0093}$	$0.6445_{\pm 0.0186}$	$0.6554_{\pm 0.0074}$
	0.02	AUROC	$0.9054_{\pm 0.0018}$	$0.9065_{\pm 0.0027}$	$0.8619_{\pm 0.0065}$	$0.8625_{\pm 0.0093}$	$0.8748_{\pm 0.0012}$	$0.8752_{\pm 0.0065}$
		FPR@95	$0.4683_{\pm 0.0076}$	$0.4508_{\pm 0.0093}$	$0.6812_{\pm 0.0192}$	$0.6723_{\pm 0.0322}$	$0.6927_{\pm 0.0082}$	$0.6885_{\pm 0.0458}$
	0.05	AUROC	$0.9054_{\pm 0.0029}$	$0.9023_{\pm 0.0024}$	$0.8421_{\pm 0.0038}$	$0.8404_{\pm 0.0170}$	$0.8717_{\pm 0.0027}$	$0.8704_{\pm 0.0052}$
		FPR@95	$0.4769_{\pm 0.0126}$	$0.4971_{\pm 0.0041}$	$0.7268_{\pm 0.0020}$	$0.7273_{\pm 0.0466}$	$0.6809_{\pm 0.0063}$	$0.6828_{\pm 0.0212}$
	0.10	AUROC	$0.8997_{\pm 0.0067}$	$0.8991_{\pm 0.0045}$	$0.8220_{\pm 0.0054}$	$0.8245_{\pm 0.0191}$	$0.8633_{\pm 0.0128}$	$0.8730_{\pm 0.0048}$
		FPR@95	$0.5025_{\pm 0.0325}$	$0.5163_{\pm 0.0090}$	$0.7603_{\pm 0.0031}$	$0.7479_{\pm 0.0432}$	$0.6710_{\pm 0.0352}$	$0.6253_{\pm 0.0327}$
	0.20	AUROC	$0.8982_{\pm 0.0032}$	$0.8997_{\pm 0.0034}$	$0.8272_{\pm 0.0122}$	$0.8533_{\pm 0.0076}$	$0.8706_{\pm 0.0026}$	$0.8738_{\pm 0.0023}$
		FPR@95	$0.4897_{\pm 0.0109}$	$0.4750_{\pm 0.0185}$	$0.7370_{\pm 0.0370}$	$0.6546_{\pm 0.0389}$	$0.6383_{\pm 0.0073}$	$0.6395_{\pm 0.0455}$
	0.50	AUROC	$0.8813_{\pm 0.0029}$	$0.8734_{\pm 0.0046}$	$0.7799_{\pm 0.0038}$	$0.8053_{\pm 0.0018}$	$0.8421_{\pm 0.0074}$	$0.8420_{\pm 0.0122}$
		FPR@95	$0.5334_{\pm 0.0151}$	$0.5831_{\pm 0.0184}$	$0.8118_{\pm 0.0172}$	$0.7767_{\pm 0.0157}$	$0.7521_{\pm 0.0553}$	$0.7516_{\pm 0.0283}$
	0.60	AUROC	$0.8631_{\pm 0.0138}$	$0.8598_{\pm 0.0058}$	$0.7605_{\pm 0.0223}$	$0.7850_{\pm 0.0440}$	$0.8314_{\pm 0.0204}$	$0.8408_{\pm 0.0042}$
		FPR@95	$0.6063_{\pm 0.0574}$	$0.6186_{\pm 0.0164}$	$0.8272_{\pm 0.0280}$	$0.7943_{\pm 0.0367}$	$0.7477_{\pm 0.0318}$	$0.7017_{\pm 0.0250}$
	0.70	AUROC	$0.8335_{\pm 0.0092}$	$0.8370_{\pm 0.0113}$	$0.6915_{\pm 0.0121}$	$0.7317_{\pm 0.0125}$	$0.8349_{\pm 0.0080}$	$0.8311_{\pm 0.0122}$
		FPR@95	$0.6943_{\pm 0.0278}$	$0.6867_{\pm 0.0459}$	$0.8798_{\pm 0.0118}$	$0.8532_{\pm 0.0098}$	$0.7611_{\pm 0.0331}$	$0.7547_{\pm 0.0491}$
DDU	0.00	AUROC	$0.9059_{\pm 0.0038}$	$0.8973_{\pm 0.0027}$	$0.8982_{\pm 0.0011}$	$0.9001_{\pm 0.0032}$	$0.8631_{\pm 0.0078}$	$0.8662_{\pm 0.0029}$
		FPR@95	$0.4330_{\pm 0.0225}$	$0.4933_{\pm 0.0164}$	$0.5103_{\pm 0.0035}$	$0.4997_{\pm 0.0170}$	$0.6686_{\pm 0.0293}$	$0.6776_{\pm 0.0163}$
	0.01	AUROC	$0.9041_{\pm 0.0031}$	$0.9023_{\pm 0.0058}$	$0.9032_{\pm 0.0023}$	$0.9035_{\pm 0.0006}$	$0.8690_{\pm 0.0015}$	$0.8691_{\pm 0.0057}$
		FPR@95	$0.4539_{\pm 0.0132}$	$0.4635_{\pm 0.0231}$	$0.5020_{\pm 0.0193}$	$0.4977_{\pm 0.0173}$	$0.6707_{\pm 0.0307}$	$0.6869_{\pm 0.0067}$
	0.02	AUROC	$0.9002_{\pm 0.0070}$	$0.9023_{\pm 0.0050}$	$0.9021_{\pm 0.0013}$	$0.9009_{\pm 0.0027}$	$0.8692_{\pm 0.0043}$	$0.8693_{\pm 0.0074}$
		FPR@95	$0.4754_{\pm 0.0261}$	$0.4631_{\pm 0.0310}$	$0.5125_{\pm 0.0111}$	$0.5276_{\pm 0.0150}$	$0.7219_{\pm 0.0101}$	$0.7238_{\pm 0.0295}$
	0.05	AUROC	$0.9069_{\pm 0.0067}$	$0.8952_{\pm 0.0011}$	$0.9024_{\pm 0.0021}$	$0.8992_{\pm 0.0014}$	$0.8692_{\pm 0.0058}$	$0.8638_{\pm 0.0052}$
		FPR@95	$0.4434_{\pm 0.0355}$	$0.5102_{\pm 0.0098}$	$0.5177_{\pm 0.0080}$	$0.5323_{\pm 0.0175}$	$0.6848_{\pm 0.0251}$	$0.7107_{\pm 0.0201}$
	0.10	AUROC	$0.8971_{\pm 0.0025}$	$0.8859_{\pm 0.0101}$	$0.8969_{\pm 0.0025}$	$0.8975_{\pm 0.0022}$	$0.8561_{\pm 0.0129}$	$0.8742_{\pm 0.0098}$
		FPR@95	$0.5233_{\pm 0.0191}$	$0.5580_{\pm 0.0415}$	$0.5686_{\pm 0.0191}$	$0.5501_{\pm 0.0169}$	$0.7153_{\pm 0.0636}$	$0.6358_{\pm 0.0213}$
	0.20	AUROC	$0.8929_{\pm 0.0060}$	$0.8893_{\pm 0.0038}$	$0.8960_{\pm 0.0011}$	$0.8992_{\pm 0.0032}$	$0.8768_{\pm 0.0034}$	$0.8722_{\pm 0.0060}$
		FPR@95	$0.5485_{\pm 0.0200}$	$0.5640_{\pm 0.0180}$	$0.5419_{\pm 0.0317}$	$0.5059_{\pm 0.0145}$	$0.5988_{\pm 0.0306}$	$0.6498_{\pm 0.0578}$
	0.50	AUROC	$0.8746_{\pm 0.0079}$	$0.8744_{\pm 0.0020}$	$0.8845_{\pm 0.0027}$	$0.8899_{\pm 0.0072}$	$0.8534_{\pm 0.0027}$	$0.8581_{\pm 0.0052}$
		FPR@95	$0.6213_{\pm 0.0349}$	$0.6267_{\pm 0.0115}$	$0.6041_{\pm 0.0289}$	$0.5823_{\pm 0.0387}$	$0.7097_{\pm 0.0333}$	$0.6906_{\pm 0.0209}$
	0.60	AUROC	$0.8565_{\pm 0.0066}$	$0.8606_{\pm 0.0067}$	$0.8812_{\pm 0.0073}$	$0.8855_{\pm 0.0114}$	$0.8494_{\pm 0.0073}$	$0.8568_{\pm 0.0013}$
		FPR@95	$0.6771_{\pm 0.0303}$	$0.6769_{\pm 0.0241}$	$0.6134_{\pm 0.0349}$	$0.5858_{\pm 0.0456}$	$0.6805_{\pm 0.0131}$	$0.6585_{\pm 0.0093}$
	0.70	AUROC	$0.8309_{\pm 0.0153}$	$0.8341_{\pm 0.0086}$	$0.8662_{\pm 0.0029}$	$0.8709_{\pm 0.0148}$	$0.8513_{\pm 0.0084}$	$0.8516_{\pm 0.0074}$
		FPR@95	$0.7385_{\pm 0.0438}$	$0.7427_{\pm 0.0253}$	$0.6403_{\pm 0.0133}$	$0.6183_{\pm 0.0355}$	$0.7119_{\pm 0.0076}$	$0.6814_{\pm 0.0461}$

Table 21: Performance on CIFAR-10 (ID) with SVHN as the OOD dataset under spectral normalization settings. Values are mean ± standard deviation over three seeds.

Method	ρ	Metric	WideResNet-28-10		ResNet-18		VGG-16	
			SNO	SNX	SNO	SNX	SNO	SNX
MSP	0.00	AUROC	$0.9289_{\pm 0.0399}$	$0.9350_{\pm 0.0115}$	$0.9405_{\pm 0.0088}$	$0.9362_{\pm 0.0109}$	$0.8562_{\pm 0.0202}$	$0.8946_{\pm 0.0099}$
		FPR@95	$0.4255_{\pm 0.1616}$	$0.4272_{\pm 0.0368}$	$0.4215_{\pm 0.0530}$	$0.4546_{\pm 0.0768}$	$0.6536_{\pm 0.0595}$	$0.5722_{\pm 0.0497}$
	0.01	AUROC	$0.9242_{\pm 0.0230}$	$0.9341_{\pm 0.0140}$	$0.9374_{\pm 0.0027}$	$0.9400_{\pm 0.0102}$	$0.8905_{\pm 0.0327}$	$0.8957_{\pm 0.0139}$
		FPR@95	$0.5083_{\pm 0.1424}$	$0.4209_{\pm 0.0552}$	$0.4648_{\pm 0.0016}$	$0.4205_{\pm 0.0774}$	$0.6420_{\pm 0.0846}$	$0.6233_{\pm 0.0798}$
	0.02	AUROC	$0.9523_{\pm 0.0080}$	$0.9463_{\pm 0.0060}$	$0.9436_{\pm 0.0171}$	$0.9393_{\pm 0.0053}$	$0.9014_{\pm 0.0332}$	$0.9148_{\pm 0.0117}$
		FPR@95	$0.3366_{\pm 0.0633}$	$0.3940_{\pm 0.0425}$	$0.3968_{\pm 0.1346}$	$0.4768_{\pm 0.0583}$	$0.6264_{\pm 0.1271}$	$0.6247_{\pm 0.0419}$
	0.05	AUROC	$0.9526_{\pm 0.0236}$	$0.9533_{\pm 0.0023}$	$0.9509_{\pm 0.0139}$	$0.9545_{\pm 0.0110}$	$0.9137_{\pm 0.0154}$	$0.9019_{\pm 0.0055}$
		FPR@95	$0.3130_{\pm 0.1346}$	$0.3570_{\pm 0.0355}$	$0.3688_{\pm 0.1316}$	$0.3444_{\pm 0.1069}$	$0.6313_{\pm 0.0791}$	$0.6739_{\pm 0.0124}$
	0.10	AUROC	$0.9593_{\pm 0.0206}$	$0.9603_{\pm 0.0087}$	$0.9592_{\pm 0.0079}$	$0.9569_{\pm 0.0056}$	$0.9246_{\pm 0.0190}$	$0.9218_{\pm 0.0079}$
		FPR@95	$0.2927_{\pm 0.1896}$	$0.2960_{\pm 0.0820}$	$0.3057_{\pm 0.0692}$	$0.3258_{\pm 0.0523}$	$0.5758_{\pm 0.1357}$	$0.6140_{\pm 0.0593}$
	0.20	AUROC	$0.9488_{\pm 0.0144}$	$0.9692_{\pm 0.0015}$	$0.9589_{\pm 0.0092}$	$0.9627_{\pm 0.0029}$	$0.9403_{\pm 0.0075}$	$0.9273_{\pm 0.0068}$
		FPR@95	$0.3776_{\pm 0.1172}$	$0.1920_{\pm 0.0177}$	$0.3043_{\pm 0.1042}$	$0.2636_{\pm 0.0306}$	$0.5088_{\pm 0.0550}$	$0.5822_{\pm 0.0588}$
	0.50	AUROC	$0.9755_{\pm 0.0013}$	$0.9782_{\pm 0.0022}$	$0.9732_{\pm 0.0048}$	$0.9677_{\pm 0.0037}$	$0.9532_{\pm 0.0036}$	$0.9595_{\pm 0.0087}$
		FPR@95	$0.1381_{\pm 0.0248}$	$0.1101_{\pm 0.0186}$	$0.1490_{\pm 0.0528}$	$0.2156_{\pm 0.0292}$	$0.4081_{\pm 0.0915}$	$0.3035_{\pm 0.0976}$
	0.60	AUROC	$0.9782_{\pm 0.0044}$	$0.9796_{\pm 0.0072}$	$0.9775_{\pm 0.0023}$	$0.9748_{\pm 0.0041}$	$0.9492_{\pm 0.0024}$	$0.9541_{\pm 0.0039}$
		FPR@95	$0.1039_{\pm 0.0340}$	$0.1057_{\pm 0.0572}$	$0.1206_{\pm 0.0230}$	$0.1331_{\pm 0.0260}$	$0.4548_{\pm 0.0498}$	$0.4042_{\pm 0.0654}$
	0.70	AUROC	$0.9784_{\pm 0.0009}$	$0.9808_{\pm 0.0035}$	$0.9739_{\pm 0.0037}$	$0.9756_{\pm 0.0009}$	$0.9569_{\pm 0.0081}$	$0.9682_{\pm 0.0026}$
		FPR@95	$0.1091_{\pm 0.0138}$	$0.0934_{\pm 0.0199}$	$0.1559_{\pm 0.0520}$	$0.1479_{\pm 0.0085}$	$0.3633_{\pm 0.1332}$	$0.2212_{\pm 0.0399}$
Energy	0.00	AUROC	$0.9352_{\pm 0.0479}$	$0.9397_{\pm 0.0144}$	$0.9509_{\pm 0.0093}$	$0.9485_{\pm 0.0173}$	$0.8083_{\pm 0.0253}$	$0.8791_{\pm 0.0195}$
		FPR@95	$0.3294_{\pm 0.2060}$	$0.3033_{\pm 0.0667}$	$0.2733_{\pm 0.0504}$	$0.2929_{\pm 0.1211}$	$0.6354_{\pm 0.0831}$	$0.5119_{\pm 0.0095}$
	0.01	AUROC	$0.9258_{\pm 0.0330}$	$0.9459_{\pm 0.0101}$	$0.9489_{\pm 0.0065}$	$0.9505_{\pm 0.0121}$	$0.8463_{\pm 0.0737}$	$0.8893_{\pm 0.0082}$
		FPR@95	$0.4201_{\pm 0.1970}$	$0.2846_{\pm 0.0221}$	$0.3126_{\pm 0.0375}$	$0.2910_{\pm 0.0844}$	$0.6305_{\pm 0.1811}$	$0.5126_{\pm 0.0783}$
	0.02	AUROC	$0.9663_{\pm 0.0050}$	$0.9589_{\pm 0.0077}$	$0.9565_{\pm 0.0185}$	$0.9548_{\pm 0.0067}$	$0.9055_{\pm 0.0402}$	$0.9066_{\pm 0.0421}$
		FPR@95	$0.1858_{\pm 0.0316}$	$0.2338_{\pm 0.0457}$	$0.2447_{\pm 0.1227}$	$0.2811_{\pm 0.0471}$	$0.5033_{\pm 0.1912}$	$0.5179_{\pm 0.1392}$
	0.05	AUROC	$0.9633_{\pm 0.0252}$	$0.9699_{\pm 0.0032}$	$0.9651_{\pm 0.0117}$	$0.9723_{\pm 0.0113}$	$0.9256_{\pm 0.0083}$	$0.9035_{\pm 0.0073}$
		FPR@95	$0.1762_{\pm 0.1230}$	$0.1713_{\pm 0.0273}$	$0.2061_{\pm 0.0914}$	$0.1532_{\pm 0.0730}$	$0.5112_{\pm 0.0716}$	$0.5870_{\pm 0.0469}$
	0.10	AUROC	$0.9753_{\pm 0.0175}$	$0.9770_{\pm 0.0053}$	$0.9737_{\pm 0.0088}$	$0.9761_{\pm 0.0036}$	$0.9352_{\pm 0.0219}$	$0.9194_{\pm 0.0152}$
		FPR@95	$0.1453_{\pm 0.1519}$	$0.1163_{\pm 0.0477}$	$0.1404_{\pm 0.0628}$	$0.1270_{\pm 0.0146}$	$0.4413_{\pm 0.1792}$	$0.5556_{\pm 0.0861}$
	0.20	AUROC	$0.9686_{\pm 0.0097}$	$0.9802_{\pm 0.0027}$	$0.9769_{\pm 0.0069}$	$0.9783_{\pm 0.0037}$	$0.9557_{\pm 0.0119}$	$0.9465_{\pm 0.0097}$
		FPR@95	$0.1818_{\pm 0.0823}$	$0.0821_{\pm 0.0176}$	$0.1187_{\pm 0.0544}$	$0.1107_{\pm 0.0139}$	$0.3421_{\pm 0.1327}$	$0.3846_{\pm 0.0692}$
	0.50	AUROC	$0.9881_{\pm 0.0012}$	$0.9880_{\pm 0.0009}$	$0.9828_{\pm 0.0029}$	$0.9788_{\pm 0.0024}$	$0.9682_{\pm 0.0048}$	$0.9719_{\pm 0.0061}$
		FPR@95	$0.0340_{\pm 0.0072}$	$0.0338_{\pm 0.0099}$	$0.0786_{\pm 0.0223}$	$0.0974_{\pm 0.0189}$	$0.1862_{\pm 0.0454}$	$0.1413_{\pm 0.0182}$
	0.60	AUROC	$0.9884_{\pm 0.0020}$	$0.9902_{\pm 0.0047}$	$0.9843_{\pm 0.0018}$	$0.9856_{\pm 0.0020}$	$0.9636_{\pm 0.0023}$	$0.9717_{\pm 0.0039}$
		FPR@95	$0.0312_{\pm 0.0087}$	$0.0272_{\pm 0.0158}$	$0.0629_{\pm 0.0077}$	$0.0524_{\pm 0.0109}$	$0.2551_{\pm 0.0352}$	$0.1733_{\pm 0.0438}$
	0.70	AUROC	$0.9882_{\pm 0.0010}$	$0.9895_{\pm 0.0034}$	$0.9803_{\pm 0.0041}$	$0.9821_{\pm 0.0016}$	$0.9697_{\pm 0.0073}$	$0.9787_{\pm 0.0031}$
		FPR@95	$0.0283_{\pm 0.0035}$	$0.0247_{\pm 0.0088}$	$0.0822_{\pm 0.0319}$	$0.0890_{\pm 0.0063}$	$0.1743_{\pm 0.0692}$	$0.1241_{\pm 0.0254}$
Mahalanobis	0.00	AUROC	$0.9781_{\pm 0.0028}$	$0.9810_{\pm 0.0034}$	$0.9426_{\pm 0.0110}$	$0.9378_{\pm 0.0236}$	$0.8967_{\pm 0.0068}$	$0.9040_{\pm 0.0125}$
		FPR@95	$0.1267_{\pm 0.0224}$	$0.1087_{\pm 0.0351}$	$0.4299_{\pm 0.1125}$	$0.4754_{\pm 0.2045}$	$0.6707_{\pm 0.0516}$	$0.7053_{\pm 0.1032}$
	0.01	AUROC	$0.9808_{\pm 0.0095}$	$0.9826_{\pm 0.0039}$	$0.9521_{\pm 0.0046}$	$0.9456_{\pm 0.0056}$	$0.9055_{\pm 0.0279}$	$0.8984_{\pm 0.0060}$
		FPR@95	$0.1120_{\pm 0.0754}$	$0.0903_{\pm 0.0320}$	$0.3428_{\pm 0.0473}$	$0.4060_{\pm 0.0501}$	$0.6622_{\pm 0.1370}$	$0.7730_{\pm 0.0259}$
	0.02	AUROC	$0.9840_{\pm 0.0021}$	$0.9804_{\pm 0.0057}$	$0.9467_{\pm 0.0119}$	$0.9389_{\pm 0.0033}$	$0.9000_{\pm 0.0305}$	$0.9199_{\pm 0.0097}$
		FPR@95	$0.0869_{\pm 0.0150}$	$0.1068_{\pm 0.0392}$	$0.3833_{\pm 0.1190}$	$0.4620_{\pm 0.0280}$	$0.7149_{\pm 0.0788}$	$0.6090_{\pm 0.0511}$
	0.05	AUROC	$0.9828_{\pm 0.0070}$	$0.9854_{\pm 0.0015}$	$0.9392_{\pm 0.0111}$	$0.9490_{\pm 0.0138}$	$0.9004_{\pm 0.0277}$	$0.9008_{\pm 0.0114}$
		FPR@95	$0.0919_{\pm 0.0474}$	$0.0701_{\pm 0.0083}$	$0.4457_{\pm 0.1263}$	$0.3451_{\pm 0.1200}$	$0.6782_{\pm 0.1278}$	$0.6958_{\pm 0.0569}$
	0.10	AUROC	$0.9859_{\pm 0.0044}$	$0.9812_{\pm 0.0032}$	$0.9490_{\pm 0.0102}$	$0.9466_{\pm 0.0128}$	$0.8654_{\pm 0.0506}$	$0.8937_{\pm 0.0104}$
		FPR@95	$0.0652_{\pm 0.0224}$	$0.1002_{\pm 0.0266}$	$0.3291_{\pm 0.0752}$	$0.3389_{\pm 0.0797}$	$0.7922_{\pm 0.1161}$	$0.6904_{\pm 0.1252}$
	0.20	AUROC	$0.9884_{\pm 0.0016}$	$0.9897_{\pm 0.0017}$	$0.9419_{\pm 0.0156}$	$0.9648_{\pm 0.0105}$	$0.8566_{\pm 0.0101}$	$0.8716_{\pm 0.0158}$
		FPR@95	$0.0538_{\pm 0.0116}$	$0.0444_{\pm 0.0111}$	$0.3883_{\pm 0.1253}$	$0.2047_{\pm 0.0769}$	$0.8519_{\pm 0.0460}$	$0.7726_{\pm 0.1024}$
	0.50	AUROC	$0.9712_{\pm 0.0097}$	$0.9683_{\pm 0.0055}$	$0.9299_{\pm 0.0341}$	$0.9195_{\pm 0.0380}$	$0.8830_{\pm 0.0366}$	$0.8670_{\pm 0.0462}$
		FPR@95	$0.1664_{\pm 0.0763}$	$0.1728_{\pm 0.0315}$	$0.3689_{\pm 0.1819}$	$0.4504_{\pm 0.2215}$	$0.6872_{\pm 0.1995}$	$0.7448_{\pm 0.1973}$
	0.60	AUROC	$0.9734_{\pm 0.0051}$	$0.9712_{\pm 0.0016}$	$0.9334_{\pm 0.0141}$	$0.9283_{\pm 0.0611}$	$0.8441_{\pm 0.0860}$	$0.8799_{\pm 0.0330}$
		FPR@95	$0.1343_{\pm 0.0277}$	$0.1512_{\pm 0.0076}$	$0.3419_{\pm 0.0683}$	$0.3462_{\pm 0.2701}$	$0.7556_{\pm 0.1628}$	$0.6236_{\pm 0.1649}$
	0.70	AUROC	$0.9675_{\pm 0.0021}$	$0.9729_{\pm 0.0085}$	$0.8627_{\pm 0.0705}$	$0.9185_{\pm 0.0104}$	$0.8654_{\pm 0.0186}$	$0.9025_{\pm 0.0554}$
		FPR@95	$0.1652_{\pm 0.0198}$	$0.1476_{\pm 0.0533}$	$0.5748_{\pm 0.2534}$	$0.3850_{\pm 0.0296}$	$0.8312_{\pm 0.0992}$	$0.5256_{\pm 0.4017}$
DDU	0.00	AUROC	$0.9789_{\pm 0.0018}$	$0.9807_{\pm 0.0049}$	$0.9464_{\pm 0.0042}$	$0.9445_{\pm 0.0093}$	$0.8698_{\pm 0.0063}$	$0.9002_{\pm 0.0136}$
		FPR@95	$0.1101_{\pm 0.0093}$	$0.1084_{\pm 0.0333}$	$0.3653_{\pm 0.0215}$	$0.4057_{\pm 0.0685}$	$0.7463_{\pm 0.0414}$	$0.7213_{\pm 0.0925}$
	0.01	AUROC	$0.9818_{\pm 0.0065}$	$0.9811_{\pm 0.0034}$	$0.9528_{\pm 0.0110}$	$0.9565_{\pm 0.0030}$	$0.9003_{\pm 0.0414}$	$0.8991_{\pm 0.0125}$
		FPR@95	$0.0972_{\pm 0.0432}$	$0.1039_{\pm 0.0292}$	$0.3523_{\pm 0.1118}$	$0.3000_{\pm 0.0352}$	$0.6281_{\pm 0.1382}$	$0.7918_{\pm 0.0842}$
	0.02	AUROC	$0.9838_{\pm 0.0032}$	$0.9781_{\pm 0.0047}$	$0.9593_{\pm 0.0042}$	$0.9487_{\pm 0.0092}$	$0.9005_{\pm 0.0387}$	$0.9219_{\pm 0.0102}$
		FPR@95	$0.0816_{\pm 0.0166}$	$0.1234_{\pm 0.0289}$	$0.2803_{\pm 0.0381}$	$0.3989_{\pm 0.1050}$	$0.6797_{\pm 0.0677}$	$0.5953_{\pm 0.0795}$
	0.05	AUROC	$0.9862_{\pm 0.0011}$	$0.9836_{\pm 0.0038}$	$0.9570_{\pm 0.0204}$	$0.9621_{\pm 0.0141}$	$0.9139_{\pm 0.0234}$	$0.9046_{\pm 0.0040}$
		FPR@95	$0.0708_{\pm 0.0036}$	$0.0845_{\pm 0.0257}$	$0.3181_{\pm 0.1992}$	$0.2601_{\pm 0.1378}$	$0.5826_{\pm 0.1784}$	$0.7413_{\pm 0.0564}$
	0.10	AUROC	$0.9827_{\pm 0.0060}$	$0.9800_{\pm 0.0019}$	$0.9503_{\pm 0.0123}$	$0.9580_{\pm 0.0113}$	$0.8834_{\pm 0.0321}$	$0.9033_{\pm 0.0141}$
		FPR@95	$0.0924_{\pm 0.0378}$	$0.1084_{\pm 0.0126}$	$0.3668_{\pm 0.1101}$	$0.3057_{\pm 0.0993}$	$0.7058_{\pm 0.1259}$	$0.7386_{\pm 0.1700}$
	0.20	AUROC	$0.9883_{\pm 0.0045}$	$0.9901_{\pm 0.0034}$	$0.9438_{\pm 0.0166}$	$0.9469_{\pm 0.0048}$	$0.8680_{\pm 0.0184}$	$0.8634_{\pm 0.0113}$
		FPR@95	$0.0510_{\pm 0.0223}$	$0.0473_{\pm 0.0171}$	$0.4281_{\pm 0.1519}$	$0.3838_{\pm 0.0372}$	$0.7874_{\pm 0.0306}$	$0.8478_{\pm 0.0622}$
	0.50	AUROC	$0.9634_{\pm 0.0071}$	$0.9747_{\pm 0.0039}$	$0.9371_{\pm 0.0212}$	$0.9150_{\pm 0.0362}$	$0.8585_{\pm 0.0310}$	$0.8615_{\pm 0.0380}$
		FPR@95	$0.2267_{\pm 0.0532}$	$0.1377_{\pm 0.0224}$	$0.4686_{\pm 0.1818}$	$0.6143_{\pm 0.2476}$	$0.8626_{\pm 0.1034}$	$0.8529_{\pm 0.1320}$
	0.60	AUROC	$0.9703_{\pm 0.0055}$	$0.9721_{\pm 0.0088}$	$0.9261_{\pm 0.0207}$	$0.9368_{\pm 0.0438}$	$0.8591_{\pm 0.0355}$	$0.8790_{\pm 0.0184}$
		FPR@95	$0.1633_{\pm 0.0353}$	$0.1502_{\pm 0.0531}$	$0.5467_{\pm 0.1684}$	$0.4214_{\pm 0.3016}$	$0.8447_{\pm 0.0421}$	$0.7286_{\pm 0.0744}$
	0.70	AUROC	$0.9540_{\pm 0.0037}$	$0.9549_{\pm 0.0140}$	$0.8902_{\pm 0.0698}$	$0.9343_{\pm 0.0147}$	$0.8544_{\pm 0.0023}$	$0.9048_{\pm 0.0291}$
		FPR@95	$0.2651_{\pm 0.0385}$	$0.2557_{\pm 0.0870}$	$0.5969_{\pm 0.2661}$	$0.4086_{\pm 0.0778}$	$0.9452_{\pm 0.0444}$	$0.6327_{\pm 0.2264}$

Table 22: Performance on CIFAR-10 (ID) with MNIST as the OOD dataset under spectral normalization settings. Values are mean ± standard deviation over three seeds.

Method	ρ	Metric	WideResNet-28-10		ResNet-18		VGG-16	
			SNO	SNX	SNO	SNX	SNO	SNX
MSP	0.00	AUROC	$0.8986_{\pm 0.0365}$	$0.9064_{\pm 0.0115}$	$0.8964_{\pm 0.0193}$	$0.9101_{\pm 0.0095}$	$0.8148_{\pm 0.0153}$	$0.8221_{\pm 0.0362}$
		FPR@95	$0.5250_{\pm 0.0837}$	$0.5550_{\pm 0.0233}$	$0.5808_{\pm 0.0669}$	$0.5547_{\pm 0.0452}$	$0.6703_{\pm 0.0106}$	$0.6520_{\pm 0.0378}$
	0.01	AUROC	$0.8984_{\pm 0.0354}$	$0.9130_{\pm 0.0127}$	$0.9040_{\pm 0.0193}$	$0.9140_{\pm 0.0120}$	$0.8335_{\pm 0.0267}$	$0.8291_{\pm 0.0269}$
		FPR@95	$0.5646_{\pm 0.0923}$	$0.5440_{\pm 0.0804}$	$0.5742_{\pm 0.0278}$	$0.5250_{\pm 0.0847}$	$0.7003_{\pm 0.0368}$	$0.6732_{\pm 0.0470}$
	0.02	AUROC	$0.8962_{\pm 0.0263}$	$0.9223_{\pm 0.0083}$	$0.9186_{\pm 0.0168}$	$0.9155_{\pm 0.0127}$	$0.8533_{\pm 0.0027}$	$0.8751_{\pm 0.0216}$
		FPR@95	$0.6024_{\pm 0.0471}$	$0.4870_{\pm 0.0427}$	$0.5186_{\pm 0.0774}$	$0.5484_{\pm 0.0643}$	$0.6987_{\pm 0.0075}$	$0.6594_{\pm 0.0726}$
	0.05	AUROC	$0.9253_{\pm 0.0143}$	$0.9110_{\pm 0.0138}$	$0.9277_{\pm 0.0059}$	$0.9211_{\pm 0.0151}$	$0.8885_{\pm 0.0128}$	$0.8614_{\pm 0.0119}$
		FPR@95	$0.5192_{\pm 0.0553}$	$0.5864_{\pm 0.0719}$	$0.4864_{\pm 0.0195}$	$0.5328_{\pm 0.0933}$	$0.6596_{\pm 0.0576}$	$0.6942_{\pm 0.0346}$
	0.10	AUROC	$0.9181_{\pm 0.0151}$	$0.9111_{\pm 0.0184}$	$0.9198_{\pm 0.0019}$	$0.8939_{\pm 0.0209}$	$0.8830_{\pm 0.0095}$	$0.8848_{\pm 0.0120}$
		FPR@95	$0.5448_{\pm 0.0592}$	$0.5885_{\pm 0.0773}$	$0.5394_{\pm 0.0081}$	$0.6261_{\pm 0.0480}$	$0.6280_{\pm 0.0333}$	$0.6302_{\pm 0.0419}$
	0.20	AUROC	$0.9274_{\pm 0.0055}$	$0.9166_{\pm 0.0188}$	$0.9045_{\pm 0.0072}$	$0.9204_{\pm 0.0076}$	$0.8686_{\pm 0.0225}$	$0.8823_{\pm 0.0110}$
		FPR@95	$0.5239_{\pm 0.0351}$	$0.5724_{\pm 0.0827}$	$0.5866_{\pm 0.0237}$	$0.5291_{\pm 0.0436}$	$0.6681_{\pm 0.0286}$	$0.6372_{\pm 0.0298}$
	0.50	AUROC	$0.9286_{\pm 0.0217}$	$0.9181_{\pm 0.0195}$	$0.9163_{\pm 0.0278}$	$0.9238_{\pm 0.0032}$	$0.8757_{\pm 0.0141}$	$0.8839_{\pm 0.0195}$
		FPR@95	$0.5432_{\pm 0.1557}$	$0.6177_{\pm 0.1086}$	$0.5566_{\pm 0.1230}$	$0.5277_{\pm 0.0355}$	$0.6643_{\pm 0.0503}$	$0.6305_{\pm 0.0686}$
	0.60	AUROC	$0.9172_{\pm 0.0178}$	$0.9258_{\pm 0.0103}$	$0.9331_{\pm 0.0124}$	$0.9220_{\pm 0.0167}$	$0.8829_{\pm 0.0144}$	$0.8773_{\pm 0.0061}$
		FPR@95	$0.6195_{\pm 0.0889}$	$0.5922_{\pm 0.0997}$	$0.4520_{\pm 0.0696}$	$0.5395_{\pm 0.0693}$	$0.6207_{\pm 0.0620}$	$0.6413_{\pm 0.0103}$
	0.70	AUROC	$0.9235_{\pm 0.0182}$	$0.9302_{\pm 0.0188}$	$0.9215_{\pm 0.0161}$	$0.9220_{\pm 0.0100}$	$0.8622_{\pm 0.0292}$	$0.8755_{\pm 0.0054}$
		FPR@95	$0.6051_{\pm 0.1319}$	$0.5535_{\pm 0.1361}$	$0.5397_{\pm 0.0914}$	$0.5388_{\pm 0.0632}$	$0.6320_{\pm 0.0496}$	$0.6727_{\pm 0.0186}$
Energy	0.00	AUROC	$0.8977_{\pm 0.0447}$	$0.9088_{\pm 0.0150}$	$0.8775_{\pm 0.0238}$	$0.8922_{\pm 0.0279}$	$0.8048_{\pm 0.0167}$	$0.8045_{\pm 0.0487}$
		FPR@95	$0.4488_{\pm 0.1185}$	$0.4506_{\pm 0.0359}$	$0.5466_{\pm 0.0815}$	$0.5448_{\pm 0.1079}$	$0.6016_{\pm 0.0244}$	$0.6004_{\pm 0.0510}$
	0.01	AUROC	$0.8990_{\pm 0.0405}$	$0.9151_{\pm 0.0213}$	$0.8918_{\pm 0.0196}$	$0.9053_{\pm 0.0128}$	$0.8259_{\pm 0.0230}$	$0.8203_{\pm 0.0440}$
		FPR@95	$0.4860_{\pm 0.1308}$	$0.4640_{\pm 0.1050}$	$0.5576_{\pm 0.0071}$	$0.4850_{\pm 0.0931}$	$0.6324_{\pm 0.0467}$	$0.5886_{\pm 0.0891}$
	0.02	AUROC	$0.9018_{\pm 0.0360}$	$0.9292_{\pm 0.0061}$	$0.9102_{\pm 0.0359}$	$0.9135_{\pm 0.0135}$	$0.8485_{\pm 0.0092}$	$0.8842_{\pm 0.0236}$
		FPR@95	$0.4886_{\pm 0.0838}$	$0.3850_{\pm 0.0353}$	$0.4720_{\pm 0.1605}$	$0.4705_{\pm 0.0592}$	$0.6393_{\pm 0.0270}$	$0.5539_{\pm 0.1071}$
	0.05	AUROC	$0.9376_{\pm 0.0119}$	$0.9169_{\pm 0.0137}$	$0.9316_{\pm 0.0111}$	$0.9198_{\pm 0.0255}$	$0.8989_{\pm 0.0187}$	$0.8765_{\pm 0.0171}$
		FPR@95	$0.3820_{\pm 0.0479}$	$0.4804_{\pm 0.0899}$	$0.3974_{\pm 0.0562}$	$0.4772_{\pm 0.1457}$	$0.5598_{\pm 0.0919}$	$0.5926_{\pm 0.0656}$
	0.10	AUROC	$0.9314_{\pm 0.0119}$	$0.9220_{\pm 0.0165}$	$0.9214_{\pm 0.0025}$	$0.8993_{\pm 0.0366}$	$0.9040_{\pm 0.0114}$	$0.9042_{\pm 0.0208}$
		FPR@95	$0.4180_{\pm 0.0540}$	$0.4701_{\pm 0.0761}$	$0.4634_{\pm 0.0069}$	$0.5205_{\pm 0.1035}$	$0.5114_{\pm 0.0471}$	$0.5185_{\pm 0.0915}$
	0.20	AUROC	$0.9436_{\pm 0.0080}$	$0.9318_{\pm 0.0221}$	$0.9113_{\pm 0.0037}$	$0.9378_{\pm 0.0054}$	$0.8968_{\pm 0.0152}$	$0.9118_{\pm 0.0063}$
		FPR@95	$0.3794_{\pm 0.0497}$	$0.4338_{\pm 0.1268}$	$0.5061_{\pm 0.0218}$	$0.3897_{\pm 0.0444}$	$0.5572_{\pm 0.0492}$	$0.4947_{\pm 0.0488}$
	0.50	AUROC	$0.9479_{\pm 0.0203}$	$0.9398_{\pm 0.0163}$	$0.9431_{\pm 0.0260}$	$0.9473_{\pm 0.0080}$	$0.9068_{\pm 0.0089}$	$0.9137_{\pm 0.0167}$
		FPR@95	$0.3537_{\pm 0.1723}$	$0.4221_{\pm 0.1287}$	$0.3737_{\pm 0.1771}$	$0.3568_{\pm 0.0838}$	$0.5195_{\pm 0.0438}$	$0.4899_{\pm 0.0886}$
	0.60	AUROC	$0.9433_{\pm 0.0172}$	$0.9566_{\pm 0.0055}$	$0.9530_{\pm 0.0093}$	$0.9492_{\pm 0.0195}$	$0.9116_{\pm 0.0171}$	$0.9141_{\pm 0.0008}$
		FPR@95	$0.3815_{\pm 0.1296}$	$0.2975_{\pm 0.0587}$	$0.3057_{\pm 0.0714}$	$0.3209_{\pm 0.1359}$	$0.5093_{\pm 0.0905}$	$0.5156_{\pm 0.0145}$
	0.70	AUROC	$0.9525_{\pm 0.0147}$	$0.9601_{\pm 0.0122}$	$0.9487_{\pm 0.0094}$	$0.9520_{\pm 0.0106}$	$0.9002_{\pm 0.0170}$	$0.9148_{\pm 0.0067}$
		FPR@95	$0.3351_{\pm 0.1334}$	$0.2510_{\pm 0.1015}$	$0.3619_{\pm 0.0848}$	$0.3274_{\pm 0.0948}$	$0.5280_{\pm 0.0444}$	$0.5213_{\pm 0.0203}$
Mahalanobis	0.00	AUROC	$0.9408_{\pm 0.0224}$	$0.9425_{\pm 0.0096}$	$0.9330_{\pm 0.0172}$	$0.9487_{\pm 0.0145}$	$0.8888_{\pm 0.0170}$	$0.8923_{\pm 0.0147}$
		FPR@95	$0.3793_{\pm 0.1412}$	$0.4036_{\pm 0.0548}$	$0.4304_{\pm 0.1290}$	$0.3229_{\pm 0.0950}$	$0.6388_{\pm 0.0797}$	$0.6055_{\pm 0.0255}$
	0.01	AUROC	$0.9473_{\pm 0.0187}$	$0.9397_{\pm 0.0021}$	$0.9348_{\pm 0.0230}$	$0.9363_{\pm 0.0395}$	$0.8778_{\pm 0.0252}$	$0.8936_{\pm 0.0100}$
		FPR@95	$0.3585_{\pm 0.1414}$	$0.3957_{\pm 0.0168}$	$0.3967_{\pm 0.1209}$	$0.3941_{\pm 0.2492}$	$0.6534_{\pm 0.1209}$	$0.5912_{\pm 0.0790}$
	0.02	AUROC	$0.9347_{\pm 0.0161}$	$0.9419_{\pm 0.0128}$	$0.9353_{\pm 0.0222}$	$0.9374_{\pm 0.0184}$	$0.8860_{\pm 0.0076}$	$0.8940_{\pm 0.0217}$
		FPR@95	$0.4393_{\pm 0.0942}$	$0.3685_{\pm 0.0542}$	$0.4117_{\pm 0.1400}$	$0.4080_{\pm 0.1253}$	$0.6583_{\pm 0.0588}$	$0.6156_{\pm 0.0785}$
	0.05	AUROC	$0.9541_{\pm 0.0169}$	$0.9493_{\pm 0.0075}$	$0.9271_{\pm 0.0102}$	$0.9472_{\pm 0.0192}$	$0.9053_{\pm 0.0247}$	$0.8706_{\pm 0.0114}$
		FPR@95	$0.3150_{\pm 0.1330}$	$0.3523_{\pm 0.0587}$	$0.4742_{\pm 0.0466}$	$0.3428_{\pm 0.1331}$	$0.5683_{\pm 0.1605}$	$0.6799_{\pm 0.0666}$
	0.10	AUROC	$0.9411_{\pm 0.0208}$	$0.9312_{\pm 0.0311}$	$0.9271_{\pm 0.0020}$	$0.9078_{\pm 0.0360}$	$0.8744_{\pm 0.0228}$	$0.9089_{\pm 0.0169}$
		FPR@95	$0.3685_{\pm 0.1124}$	$0.4552_{\pm 0.1881}$	$0.4644_{\pm 0.0050}$	$0.5634_{\pm 0.1940}$	$0.6594_{\pm 0.0935}$	$0.4942_{\pm 0.0892}$
	0.20	AUROC	$0.9410_{\pm 0.0108}$	$0.9247_{\pm 0.0137}$	$0.9013_{\pm 0.0247}$	$0.8882_{\pm 0.0173}$	$0.8786_{\pm 0.0137}$	$0.8711_{\pm 0.0284}$
		FPR@95	$0.3880_{\pm 0.0624}$	$0.4904_{\pm 0.1024}$	$0.5904_{\pm 0.1043}$	$0.6598_{\pm 0.0730}$	$0.6273_{\pm 0.0086}$	$0.6521_{\pm 0.0972}$
	0.50	AUROC	$0.9020_{\pm 0.0228}$	$0.9207_{\pm 0.0124}$	$0.7895_{\pm 0.0344}$	$0.7675_{\pm 0.0260}$	$0.8188_{\pm 0.0024}$	$0.8507_{\pm 0.0349}$
		FPR@95	$0.6332_{\pm 0.1044}$	$0.5722_{\pm 0.0477}$	$0.8966_{\pm 0.0516}$	$0.9280_{\pm 0.0342}$	$0.8470_{\pm 0.0213}$	$0.7385_{\pm 0.1258}$
	0.60	AUROC	$0.8716_{\pm 0.0160}$	$0.8489_{\pm 0.0256}$	$0.7062_{\pm 0.0407}$	$0.7492_{\pm 0.0502}$	$0.8269_{\pm 0.0437}$	$0.7529_{\pm 0.0085}$
		FPR@95	$0.7773_{\pm 0.0302}$	$0.8136_{\pm 0.0671}$	$0.9628_{\pm 0.0127}$	$0.9512_{\pm 0.0252}$	$0.7842_{\pm 0.1475}$	$0.9142_{\pm 0.0346}$
	0.70	AUROC	$0.8662_{\pm 0.0215}$	$0.8626_{\pm 0.0335}$	$0.6410_{\pm 0.0376}$	$0.7058_{\pm 0.0213}$	$0.8013_{\pm 0.0159}$	$0.7292_{\pm 0.0744}$
		FPR@95	$0.8063_{\pm 0.0834}$	$0.7960_{\pm 0.1219}$	$0.9851_{\pm 0.0091}$	$0.9822_{\pm 0.0084}$	$0.8413_{\pm 0.0331}$	$0.8703_{\pm 0.0806}$
DDU	0.00	AUROC	$0.9135_{\pm 0.0275}$	$0.9126_{\pm 0.0093}$	$0.9345_{\pm 0.0132}$	$0.9506_{\pm 0.0119}$	$0.8639_{\pm 0.0311}$	$0.8703_{\pm 0.0226}$
		FPR@95	$0.4907_{\pm 0.1195}$	$0.5634_{\pm 0.0329}$	$0.4357_{\pm 0.1014}$	$0.3250_{\pm 0.0759}$	$0.6694_{\pm 0.0697}$	$0.5981_{\pm 0.0543}$
	0.01	AUROC	$0.9216_{\pm 0.0264}$	$0.9112_{\pm 0.0049}$	$0.9383_{\pm 0.0155}$	$0.9423_{\pm 0.0311}$	$0.8614_{\pm 0.0294}$	$0.8771_{\pm 0.0090}$
		FPR@95	$0.4599_{\pm 0.1220}$	$0.5404_{\pm 0.0240}$	$0.3973_{\pm 0.0633}$	$0.3768_{\pm 0.2145}$	$0.6291_{\pm 0.0831}$	$0.6112_{\pm 0.0831}$
	0.02	AUROC	$0.9042_{\pm 0.0164}$	$0.9235_{\pm 0.0055}$	$0.9443_{\pm 0.0035}$	$0.9475_{\pm 0.0055}$	$0.8625_{\pm 0.0203}$	$0.8799_{\pm 0.0243}$
		FPR@95	$0.5626_{\pm 0.0452}$	$0.4833_{\pm 0.0407}$	$0.3699_{\pm 0.0332}$	$0.3681_{\pm 0.0524}$	$0.6702_{\pm 0.0919}$	$0.6351_{\pm 0.0390}$
	0.05	AUROC	$0.9271_{\pm 0.0279}$	$0.9085_{\pm 0.0117}$	$0.9491_{\pm 0.0078}$	$0.9497_{\pm 0.0094}$	$0.8907_{\pm 0.0238}$	$0.8411_{\pm 0.0135}$
		FPR@95	$0.4257_{\pm 0.1644}$	$0.5536_{\pm 0.0441}$	$0.3439_{\pm 0.0485}$	$0.3455_{\pm 0.0469}$	$0.6086_{\pm 0.1694}$	$0.7138_{\pm 0.0752}$
	0.10	AUROC	$0.9249_{\pm 0.0176}$	$0.8910_{\pm 0.0350}$	$0.9503_{\pm 0.0016}$	$0.9283_{\pm 0.0064}$	$0.8730_{\pm 0.0231}$	$0.8921_{\pm 0.0203}$
		FPR@95	$0.4703_{\pm 0.0660}$	$0.6334_{\pm 0.1572}$	$0.3495_{\pm 0.0228}$	$0.4929_{\pm 0.0949}$	$0.6062_{\pm 0.0854}$	$0.5637_{\pm 0.0843}$
	0.20	AUROC	$0.9018_{\pm 0.0267}$	$0.9032_{\pm 0.0234}$	$0.9319_{\pm 0.0193}$	$0.9271_{\pm 0.0147}$	$0.8641_{\pm 0.0331}$	$0.8681_{\pm 0.0106}$
		FPR@95	$0.6205_{\pm 0.0998}$	$0.6163_{\pm 0.0842}$	$0.4467_{\pm 0.1400}$	$0.4681_{\pm 0.0966}$	$0.6178_{\pm 0.0282}$	$0.6467_{\pm 0.1260}$
	0.50	AUROC	$0.8895_{\pm 0.0071}$	$0.8738_{\pm 0.0166}$	$0.8764_{\pm 0.0134}$	$0.8683_{\pm 0.0107}$	$0.8363_{\pm 0.0034}$	$0.8613_{\pm 0.0230}$
		FPR@95	$0.7452_{\pm 0.0332}$	$0.7869_{\pm 0.0437}$	$0.7951_{\pm 0.0235}$	$0.8177_{\pm 0.0238}$	$0.8877_{\pm 0.0366}$	$0.7563_{\pm 0.1163}$
	0.60	AUROC	$0.8322_{\pm 0.0228}$	$0.8178_{\pm 0.0193}$	$0.8674_{\pm 0.0137}$	$0.8531_{\pm 0.0090}$	$0.8460_{\pm 0.0143}$	$0.7946_{\pm 0.0098}$
		FPR@95	$0.8809_{\pm 0.0661}$	$0.9233_{\pm 0.0334}$	$0.8237_{\pm 0.0250}$	$0.8725_{\pm 0.0102}$	$0.8173_{\pm 0.1164}$	$0.9220_{\pm 0.0024}$
	0.70	AUROC	$0.8095_{\pm 0.0127}$	$0.7839_{\pm 0.0420}$	$0.8350_{\pm 0.0257}$	$0.8408_{\pm 0.0272}$	$0.8170_{\pm 0.0161}$	$0.7846_{\pm 0.0424}$
		FPR@95	$0.9093_{\pm 0.0045}$	$0.9275_{\pm 0.0402}$	$0.8746_{\pm 0.0479}$	$0.8820_{\pm 0.0457}$	$0.8794_{\pm 0.0537}$	$0.8714_{\pm 0.0691}$

Table 23: Geometric representation metrics on CIFAR-10 under backbone and spectral-normalization settings. Values are mean ± standard deviation over three seeds.

Metric	ρ	WideResNet-28-10		ResNet-18		VGG-16	
		SNO	SNX	SNO	SNX	SNO	SNX
WithinVar	0.00	$0.8655_{\pm 0.0244}$	$0.9524_{\pm 0.1090}$	$1.3084_{\pm 0.0989}$	$1.2826_{\pm 0.0253}$	$0.8570_{\pm 0.0659}$	$0.8259_{\pm 0.0304}$
	0.01	$0.9261_{\pm 0.0629}$	$1.0262_{\pm 0.0807}$	$1.4725_{\pm 0.0642}$	$1.5165_{\pm 0.0820}$	$1.1101_{\pm 0.1386}$	$1.0430_{\pm 0.0537}$
	0.02	$1.0805_{\pm 0.0862}$	$1.0807_{\pm 0.0798}$	$1.7025_{\pm 0.0494}$	$1.7268_{\pm 0.1107}$	$1.4743_{\pm 0.1105}$	$1.4688_{\pm 0.0531}$
	0.05	$1.2640_{\pm 0.1631}$	$1.4300_{\pm 0.0910}$	$2.3415_{\pm 0.0658}$	$2.2731_{\pm 0.2644}$	$2.7008_{\pm 0.1704}$	$2.4209_{\pm 0.1508}$
	0.10	$1.7433_{\pm 0.1295}$	$1.7592_{\pm 0.0550}$	$2.9947_{\pm 0.1434}$	$3.1762_{\pm 0.0241}$	$2.1435_{\pm 0.1183}$	$2.1078_{\pm 0.0670}$
	0.20	$2.0108_{\pm 0.0214}$	$2.0171_{\pm 0.0500}$	$3.0706_{\pm 0.0789}$	$2.7717_{\pm 0.1360}$	$1.3636_{\pm 0.1022}$	$1.3310_{\pm 0.0942}$
	0.50	$2.4303_{\pm 0.0142}$	$2.3845_{\pm 0.0017}$	$4.2795_{\pm 0.0831}$	$3.9377_{\pm 0.1116}$	$2.9106_{\pm 0.2442}$	$2.9409_{\pm 0.1282}$
	0.60	$2.8391_{\pm 0.0845}$	$2.8358_{\pm 0.0316}$	$4.3733_{\pm 0.2242}$	$4.2953_{\pm 0.4377}$	$2.7787_{\pm 0.4992}$	$3.0236_{\pm 0.3212}$
	0.70	$3.3350_{\pm 0.0910}$	$3.2381_{\pm 0.0713}$	$4.8353_{\pm 0.4370}$	$4.4372_{\pm 0.0914}$	$2.2976_{\pm 0.3898}$	$3.0385_{\pm 0.1290}$
InterDist	0.00	$7.3106_{\pm 0.3599}$	$7.5098_{\pm 0.0780}$	$7.0188_{\pm 0.0246}$	$7.0012_{\pm 0.0380}$	$9.8223_{\pm 0.0564}$	$9.8739_{\pm 0.0370}$
	0.01	$7.4538_{\pm 0.0153}$	$7.5189_{\pm 0.0430}$	$7.0129_{\pm 0.0097}$	$6.9660_{\pm 0.0213}$	$9.8038_{\pm 0.0660}$	$9.7738_{\pm 0.0484}$
	0.02	$7.4644_{\pm 0.0632}$	$7.4630_{\pm 0.0417}$	$7.0202_{\pm 0.0063}$	$6.9968_{\pm 0.0141}$	$9.7281_{\pm 0.0293}$	$9.7369_{\pm 0.0628}$
	0.05	$7.1803_{\pm 0.3163}$	$7.3428_{\pm 0.0306}$	$7.0158_{\pm 0.0237}$	$7.0621_{\pm 0.0271}$	$9.5440_{\pm 0.0348}$	$9.5192_{\pm 0.0365}$
	0.10	$7.1992_{\pm 0.0944}$	$7.2117_{\pm 0.0364}$	$7.0275_{\pm 0.0631}$	$7.2276_{\pm 0.3390}$	$9.1657_{\pm 0.0175}$	$9.1094_{\pm 0.0470}$
	0.20	$6.9563_{\pm 0.0558}$	$6.9771_{\pm 0.0346}$	$6.9505_{\pm 0.4183}$	$7.3805_{\pm 0.1632}$	$8.8062_{\pm 0.0349}$	$8.7681_{\pm 0.1158}$
	0.50	$6.8398_{\pm 0.0089}$	$6.8122_{\pm 0.0500}$	$6.7109_{\pm 0.1016}$	$6.8054_{\pm 0.0093}$	$8.1212_{\pm 0.1636}$	$8.2407_{\pm 0.1162}$
	0.60	$6.8387_{\pm 0.0716}$	$6.7653_{\pm 0.0645}$	$6.4154_{\pm 0.0863}$	$6.7071_{\pm 0.1371}$	$7.6344_{\pm 0.2629}$	$7.7840_{\pm 0.1155}$
	0.70	$6.7350_{\pm 0.0993}$	$6.6446_{\pm 0.0776}$	$5.8342_{\pm 0.4225}$	$5.7983_{\pm 0.2615}$	$6.7348_{\pm 0.5783}$	$7.4152_{\pm 0.1140}$
Anisotropy	0.00	$0.2797_{\pm 0.0273}$	$0.2908_{\pm 0.0117}$	$0.2322_{\pm 0.0091}$	$0.2309_{\pm 0.0007}$	$0.3152_{\pm 0.0070}$	$0.3113_{\pm 0.0227}$
	0.01	$0.2814_{\pm 0.0111}$	$0.2751_{\pm 0.0112}$	$0.2018_{\pm 0.0063}$	$0.2012_{\pm 0.0068}$	$0.3391_{\pm 0.0074}$	$0.3441_{\pm 0.0199}$
	0.02	$0.2484_{\pm 0.0108}$	$0.2458_{\pm 0.0115}$	$0.1845_{\pm 0.0040}$	$0.1828_{\pm 0.0037}$	$0.3257_{\pm 0.0170}$	$0.3236_{\pm 0.0144}$
	0.05	$0.2031_{\pm 0.0011}$	$0.2030_{\pm 0.0036}$	$0.1874_{\pm 0.0022}$	$0.1974_{\pm 0.0119}$	$0.3126_{\pm 0.0123}$	$0.3262_{\pm 0.0029}$
	0.10	$0.2102_{\pm 0.0010}$	$0.2195_{\pm 0.0055}$	$0.2039_{\pm 0.0023}$	$0.2101_{\pm 0.0126}$	$0.3656_{\pm 0.0084}$	$0.3474_{\pm 0.0052}$
	0.20	$0.2290_{\pm 0.0008}$	$0.2281_{\pm 0.0044}$	$0.2412_{\pm 0.0197}$	$0.2658_{\pm 0.0084}$	—	$0.5149_{\pm 0.0166}$
	0.50	$0.2678_{\pm 0.0089}$	$0.2741_{\pm 0.0068}$	$0.2920_{\pm 0.0029}$	$0.3047_{\pm 0.0127}$	$0.6437_{\pm 0.0052}$	$0.6820_{\pm 0.0390}$
	0.60	$0.2824_{\pm 0.0031}$	$0.2836_{\pm 0.0026}$	$0.3080_{\pm 0.0019}$	$0.3092_{\pm 0.0199}$	$0.6481_{\pm 0.0120}$	—
	0.70	$0.2935_{\pm 0.0107}$	$0.2918_{\pm 0.0063}$	$0.3217_{\pm 0.0196}$	$0.2978_{\pm 0.0076}$	—	$0.6621_{\pm 0.0148}$
EffRank	0.00	$11.8506_{\pm 0.6437}$	$11.5863_{\pm 0.1941}$	$13.6690_{\pm 0.1965}$	$13.8061_{\pm 0.0695}$	$8.5327_{\pm 0.2801}$	$8.8796_{\pm 0.3253}$
	0.01	$11.6274_{\pm 0.3868}$	$11.9750_{\pm 0.3534}$	$14.2552_{\pm 0.0367}$	$14.2908_{\pm 0.1417}$	$7.8640_{\pm 0.4013}$	$8.1141_{\pm 0.4836}$
	0.02	$12.1296_{\pm 0.5708}$	$12.0481_{\pm 0.0897}$	$14.4314_{\pm 0.0496}$	$14.5188_{\pm 0.1578}$	$8.0303_{\pm 0.4203}$	$8.1517_{\pm 0.1632}$
	0.05	$12.3514_{\pm 0.2394}$	$12.3538_{\pm 0.0796}$	$14.4099_{\pm 0.1424}$	$14.1384_{\pm 0.3301}$	$8.0347_{\pm 0.0192}$	$8.1082_{\pm 0.2098}$
	0.10	$12.6358_{\pm 0.0689}$	$12.1764_{\pm 0.2298}$	$14.3578_{\pm 0.1870}$	$14.2291_{\pm 0.2700}$	$7.0925_{\pm 0.3906}$	$7.1778_{\pm 0.3869}$
	0.20	$12.7013_{\pm 0.2215}$	$12.5211_{\pm 0.0825}$	$14.2816_{\pm 1.0634}$	$13.2197_{\pm 0.3327}$	—	$5.4640_{\pm 0.0761}$
	0.50	$12.4605_{\pm 0.2444}$	$12.3937_{\pm 0.1965}$	$15.3257_{\pm 1.2270}$	$14.5775_{\pm 1.9266}$	$4.2335_{\pm 0.1084}$	$3.7129_{\pm 0.5532}$
	0.60	$12.1434_{\pm 0.3828}$	$12.2818_{\pm 0.1204}$	$14.8692_{\pm 1.5178}$	$15.2799_{\pm 2.6095}$	$4.0460_{\pm 0.0789}$	—
	0.70	$12.3875_{\pm 0.3689}$	$12.2557_{\pm 0.2060}$	$14.3276_{\pm 2.4464}$	$16.1290_{\pm 0.2384}$	—	$3.7972_{\pm 0.1606}$

Table 24: Neural-collapse metrics on CIFAR-10 under backbone and spectral-normalization settings. Values are mean ± standard deviation over three seeds.

Metric	ρ	WideResNet-28-10		ResNet-18		VGG-16	
		SNO	SNX	SNO	SNX	SNO	SNX
$\mathcal{NC}_1$	0.00	$0.0221_{\pm 0.0009}$	$0.0246_{\pm 0.0032}$	$0.0346_{\pm 0.0030}$	$0.0340_{\pm 0.0002}$	$0.0101_{\pm 0.0010}$	$0.0096_{\pm 0.0005}$
	0.01	$0.0240_{\pm 0.0015}$	$0.0265_{\pm 0.0025}$	$0.0384_{\pm 0.0020}$	$0.0401_{\pm 0.0027}$	$0.0135_{\pm 0.0020}$	$0.0125_{\pm 0.0008}$
	0.02	$0.0278_{\pm 0.0016}$	$0.0280_{\pm 0.0026}$	$0.0436_{\pm 0.0016}$	$0.0447_{\pm 0.0034}$	$0.0202_{\pm 0.0019}$	$0.0196_{\pm 0.0008}$
	0.05	$0.0342_{\pm 0.0044}$	$0.0385_{\pm 0.0035}$	$0.0583_{\pm 0.0023}$	$0.0556_{\pm 0.0077}$	$0.0411_{\pm 0.0035}$	$0.0367_{\pm 0.0028}$
	0.10	$0.0474_{\pm 0.0050}$	$0.0484_{\pm 0.0020}$	$0.0715_{\pm 0.0037}$	$0.0701_{\pm 0.0087}$	$0.0365_{\pm 0.0028}$	$0.0354_{\pm 0.0021}$
	0.20	$0.0549_{\pm 0.0014}$	$0.0555_{\pm 0.0007}$	$0.0667_{\pm 0.0068}$	$0.0531_{\pm 0.0024}$	$0.0279_{\pm 0.0017}$	$0.0259_{\pm 0.0011}$
	0.50	$0.0639_{\pm 0.0004}$	$0.0640_{\pm 0.0003}$	$0.0895_{\pm 0.0096}$	$0.0818_{\pm 0.0076}$	$0.0783_{\pm 0.0017}$	$0.0756_{\pm 0.0016}$
	0.60	$0.0757_{\pm 0.0022}$	$0.0767_{\pm 0.0004}$	$0.1059_{\pm 0.0128}$	$0.0966_{\pm 0.0212}$	$0.0889_{\pm 0.0028}$	$0.0919_{\pm 0.0037}$
	0.70	$0.0908_{\pm 0.0007}$	$0.0904_{\pm 0.0025}$	$0.1458_{\pm 0.0031}$	$0.1292_{\pm 0.0092}$	$0.1018_{\pm 0.0033}$	$0.1063_{\pm 0.0036}$
$\mathcal{NC}_2^{\mathcal{A}}$	0.00	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0001}$	$-0.1111_{\pm 0.0000}$	$-0.1110_{\pm 0.0000}$
	0.01	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0001}$	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0001}$	$-0.1110_{\pm 0.0001}$	$-0.1111_{\pm 0.0000}$
	0.02	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0001}$	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0000}$
	0.05	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0001}$	$-0.1110_{\pm 0.0001}$	$-0.1110_{\pm 0.0001}$
	0.10	$-0.1109_{\pm 0.0001}$	$-0.1109_{\pm 0.0000}$	$-0.1110_{\pm 0.0000}$	$-0.1110_{\pm 0.0001}$	$-0.1109_{\pm 0.0001}$	$-0.1110_{\pm 0.0001}$
	0.20	$-0.1107_{\pm 0.0000}$	$-0.1108_{\pm 0.0000}$	$-0.1108_{\pm 0.0001}$	$-0.1107_{\pm 0.0002}$	$-0.1105_{\pm 0.0000}$	$-0.1106_{\pm 0.0000}$
	0.50	$-0.1106_{\pm 0.0001}$	$-0.1105_{\pm 0.0001}$	$-0.1096_{\pm 0.0005}$	$-0.1095_{\pm 0.0007}$	$-0.1099_{\pm 0.0002}$	$-0.1101_{\pm 0.0001}$
	0.60	$-0.1104_{\pm 0.0001}$	$-0.1104_{\pm 0.0001}$	$-0.1086_{\pm 0.0009}$	$-0.1094_{\pm 0.0002}$	$-0.1081_{\pm 0.0027}$	$-0.1092_{\pm 0.0004}$
	0.70	$-0.1102_{\pm 0.0000}$	$-0.1101_{\pm 0.0002}$	$-0.1078_{\pm 0.0009}$	$-0.1077_{\pm 0.0003}$	$-0.1076_{\pm 0.0016}$	$-0.1089_{\pm 0.0004}$
$\mathcal{NC}_2$	0.00	$0.0864_{\pm 0.0028}$	$0.0964_{\pm 0.0019}$	$0.1034_{\pm 0.0043}$	$0.1024_{\pm 0.0016}$	$0.1371_{\pm 0.0082}$	$0.1394_{\pm 0.0036}$
	0.01	$0.0972_{\pm 0.0025}$	$0.1003_{\pm 0.0028}$	$0.1104_{\pm 0.0008}$	$0.1122_{\pm 0.0021}$	$0.1568_{\pm 0.0008}$	$0.1512_{\pm 0.0039}$
	0.02	$0.1070_{\pm 0.0024}$	$0.1099_{\pm 0.0019}$	$0.1261_{\pm 0.0021}$	$0.1243_{\pm 0.0016}$	$0.1674_{\pm 0.0077}$	$0.1691_{\pm 0.0078}$
	0.05	$0.1385_{\pm 0.0084}$	$0.1486_{\pm 0.0066}$	$0.1676_{\pm 0.0023}$	$0.1737_{\pm 0.0095}$	$0.2422_{\pm 0.0040}$	$0.2410_{\pm 0.0006}$
	0.10	$0.1993_{\pm 0.0084}$	$0.2086_{\pm 0.0056}$	$0.2293_{\pm 0.0014}$	$0.2313_{\pm 0.0118}$	$0.3716_{\pm 0.0088}$	$0.3668_{\pm 0.0099}$
	0.20	$0.2769_{\pm 0.0012}$	$0.2734_{\pm 0.0061}$	$0.3227_{\pm 0.0200}$	$0.3471_{\pm 0.0077}$	$0.4575_{\pm 0.0149}$	$0.4640_{\pm 0.0051}$
	0.50	$0.3480_{\pm 0.0025}$	$0.3499_{\pm 0.0135}$	$0.4118_{\pm 0.0033}$	$0.3972_{\pm 0.0026}$	$0.5709_{\pm 0.0189}$	$0.5761_{\pm 0.0214}$
	0.60	$0.3769_{\pm 0.0027}$	$0.3784_{\pm 0.0045}$	$0.4321_{\pm 0.0130}$	$0.4190_{\pm 0.0331}$	$0.5916_{\pm 0.0143}$	$0.6071_{\pm 0.0194}$
	0.70	$0.4055_{\pm 0.0059}$	$0.4127_{\pm 0.0068}$	$0.4795_{\pm 0.0285}$	$0.4521_{\pm 0.0085}$	$0.6717_{\pm 0.0253}$	$0.6546_{\pm 0.0151}$
$\mathcal{NC}_3$	0.00	$0.1315_{\pm 0.0465}$	$0.1606_{\pm 0.0032}$	$0.1482_{\pm 0.0001}$	$0.1449_{\pm 0.0040}$	$0.1571_{\pm 0.0036}$	$0.1619_{\pm 0.0023}$
	0.01	$0.1639_{\pm 0.0037}$	$0.1670_{\pm 0.0054}$	$0.1513_{\pm 0.0025}$	$0.1553_{\pm 0.0013}$	$0.1582_{\pm 0.0069}$	$0.1526_{\pm 0.0108}$
	0.02	$0.1721_{\pm 0.0028}$	$0.1745_{\pm 0.0032}$	$0.1653_{\pm 0.0065}$	$0.1651_{\pm 0.0022}$	$0.1729_{\pm 0.0101}$	$0.1647_{\pm 0.0060}$
	0.05	$0.1816_{\pm 0.0319}$	$0.2010_{\pm 0.0075}$	$0.1925_{\pm 0.0009}$	$0.1942_{\pm 0.0044}$	$0.2175_{\pm 0.0091}$	$0.2019_{\pm 0.0073}$
	0.10	$0.2429_{\pm 0.0015}$	$0.2444_{\pm 0.0049}$	$0.2343_{\pm 0.0074}$	$0.2384_{\pm 0.0044}$	$0.3088_{\pm 0.0093}$	$0.2996_{\pm 0.0160}$
	0.20	$0.2954_{\pm 0.0008}$	$0.2972_{\pm 0.0015}$	$0.2991_{\pm 0.0171}$	$0.3134_{\pm 0.0042}$	$0.3949_{\pm 0.0192}$	$0.3956_{\pm 0.0173}$
	0.50	$0.3354_{\pm 0.0005}$	$0.3392_{\pm 0.0057}$	$0.3940_{\pm 0.0085}$	$0.3869_{\pm 0.0115}$	$0.4663_{\pm 0.0065}$	$0.4672_{\pm 0.0148}$
	0.60	$0.3596_{\pm 0.0031}$	$0.3613_{\pm 0.0020}$	$0.4249_{\pm 0.0105}$	$0.3979_{\pm 0.0189}$	$0.5189_{\pm 0.0645}$	$0.5141_{\pm 0.0152}$
	0.70	$0.3791_{\pm 0.0030}$	$0.3861_{\pm 0.0066}$	$0.4826_{\pm 0.0152}$	$0.4718_{\pm 0.0099}$	$0.5709_{\pm 0.0259}$	$0.5395_{\pm 0.0013}$
$\mathcal{NC}_4$	0.00	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$
	0.01	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$
	0.02	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$
	0.05	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$
	0.10	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$1.0000_{\pm 0.0000}$	$0.9999_{\pm 0.0000}$	$0.9993_{\pm 0.0001}$	$0.9993_{\pm 0.0000}$
	0.20	$0.9994_{\pm 0.0001}$	$0.9994_{\pm 0.0001}$	$0.9962_{\pm 0.0007}$	$0.9952_{\pm 0.0004}$	$0.9908_{\pm 0.0002}$	$0.9914_{\pm 0.0001}$
	0.50	$0.9918_{\pm 0.0007}$	$0.9916_{\pm 0.0001}$	$0.9809_{\pm 0.0004}$	$0.9809_{\pm 0.0017}$	$0.9647_{\pm 0.0009}$	$0.9662_{\pm 0.0003}$
	0.60	$0.9891_{\pm 0.0002}$	$0.9897_{\pm 0.0003}$	$0.9752_{\pm 0.0010}$	$0.9765_{\pm 0.0004}$	$0.9588_{\pm 0.0016}$	$0.9588_{\pm 0.0010}$
	0.70	$0.9866_{\pm 0.0005}$	$0.9865_{\pm 0.0003}$	$0.9670_{\pm 0.0031}$	$0.9689_{\pm 0.0012}$	$0.9517_{\pm 0.0009}$	$0.9517_{\pm 0.0013}$