DIVERSITY AS A PARADIGM FOR OUT-OF-DISTRIBUTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection is critical for the safe deployment of machine learning systems and for scientific discovery. Existing post-hoc detectors typically rely on model confidence scores or likelihood estimates in feature space, often under restrictive distributional assumptions. In this work, we introduce a new paradigm and formulate OOD detection from a diversity perspective. We propose Vendi Score Out-Of-Distribution detection (VSOOD), an OOD detector based on the Vendi Scores (VS), a family of similarity-based diversity metrics. VSOOD quantifies how much a test sample increases the VS of the in-distribution feature set, providing a principled notion of novelty that does not require density modeling. VSOOD is linear-time and naturally combines class-conditional (local) and dataset-level (global) novelty signals. Across multiple image classification benchmarks and network architectures, VSOOD achieves state-of-the-art OOD detection performance. Remarkably, VSOOD retains this performance when computed using only 1% of the training data, enabling deployment in memory- or access-constrained settings.

1 INTRODUCTION

Modern machine learning systems have made remarkable progress across a wide variety of tasks. Yet, these systems can behave unpredictably on previously unseen out-of-distribution (OOD) inputs, often producing high-confidence, incorrect predictions (Goodfellow et al., 2014; Hendrycks & Gimpel, 2016). Knowing when a model can be trusted is essential in high-stakes settings such as autonomous driving (Shoeb et al., 2025), medical imaging (Zhang et al., 2021), and model-guided scientific discovery (Segal et al., 2025). OOD detectors provide a safeguard for these settings by identifying the inputs that are OOD and should be routed to human review or other fallback policies.

Recent methods for OOD detection span specialized models (Hendrycks et al., 2018; Katz-Samuels et al., 2022), modifications to the training process (Pinto et al., 2022; Hendrycks et al., 2022), and post-hoc methods that act independently of the training process (Hendrycks & Gimpel, 2016; Lee et al., 2018; Yang et al., 2024). As models have become increasingly large and trained on massive datasets, post-hoc methods are often the cheapest and most practical options for OOD detection. Post-hoc methods can be broadly categorized by the information from the model they exploit during inference. One class of approaches relies on the model's prediction logits (Liu et al., 2023; 2020), while another leverages the model's intermediate representations and gradients to identify deviations from the in-distribution (ID) feature geometry (Ren et al., 2021; Ammar et al., 2023; Huang et al., 2021). While this second class of methods often achieves the strongest performance, they also make strong assumptions about the geometry of the model's representations, which may not hold consistently across architectures and datasets. For example, Liu & Qin (2025) leverage the properties of neural collapse for OOD detection, which can lead to worse performance on off-the-shelf classifiers that do not exhibit neural collapse. Other similarity-based approaches like Sun et al. (2022) do not make explicit assumptions about the representation space, but incur significant test-time overhead by performing a k-nearest neighbor search to identify OOD samples. More recent work aiming to exploit the gradients of the model (Regmi, 2025) requires multiple forward passes, making them impractical in latency-sensitive settings. Designing computationally efficient OOD detectors without restrictive distributional assumptions is therefore critical for real-world deployment.

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.

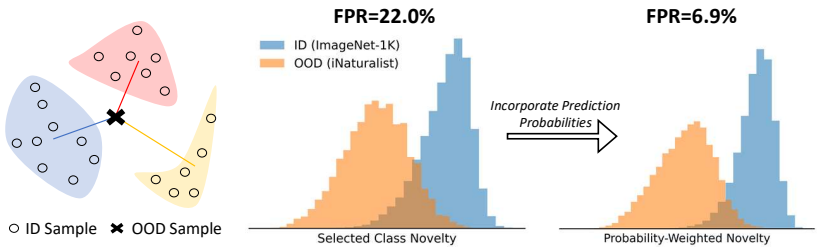

Figure 1: Conceptual overview of VSOOD. Left: Three in-distribution (ID) classes are shown as clusters in representation space (colored regions), along with a test out-of-distribution (OOD) sample (star). VSOOD computes a class-conditional novelty score for the OOD sample. Middle: Using only the class-conditional novelty score from the model's predicted class provides reasonable ID-OOD separation. Right: Aggregating class-conditional novelty scores using prediction probabilities (Equation 7) yields improved ID-OOD separation, reducing the False Positive Rate (FPR).

In this work, we introduce a new paradigm for post-hoc OOD detection based on diversity. We build on the Vendi Scores (VS), a family of similarity-based diversity metrics (Friedman & Dieng, 2022; Pasarkar & Dieng, 2023), and propose Vendi Score Out-Of-Distribution detection (VSOOD). Rather than approximating feature distributions, VSOOD quantifies novelty via a test sample's impact on representation diversity. Specifically, VSOOD computes class-conditional novelty contributions—as measured by the change in VS when the test sample is added to each training set class—and aggregates these contributions using the model's predicted class probabilities (Fig. 1).

VSOOD offers an efficient way to evaluate sample novelty and incorporates both local and global novelty signals. We provide a series of approximations that allow VSOOD to scale linearly with the data dimension and number of classes. We demonstrate the effectiveness of VSOOD across various OOD benchmarks spanning several datasets (CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), ImageNet-1K (Deng et al., 2009)) and multiple network architectures (ResNet (He et al., 2016), ViT (Dosovitskiy, 2020), Swin-T (Liu et al., 2021)). Our approach often achieves state-of-the-art detection performance, even when given access to only 1% of the training data. We also provide insights into the accuracy and efficiency of VSOOD theoretically and through various ablation studies.

2 Related Work

Post-Hoc OOD Detection. Early work on post-hoc OOD detection focused on score functions based on the model's logits. Such methods leverage characteristics of the logit distribution, such as the maximum class probabilities, energy scores, or their entropy (Hendrycks & Gimpel, 2016; Liu et al., 2020; 2023). Huang et al. (2021) measured novelty via the gradient norm of the KL divergence of the logit distribution from a uniform prior. These approaches are attractive because they do not rely on access to the training data, making them easy to deploy. However, this simplicity comes at the cost of ignoring in-distribution feature statistics, which can reduce OOD detection reliability.

Another line of work modifies the model's inference-time activations to increase separability between ID and OOD logit distributions. Activation-shaping methods clip extreme activations or apply pruning or adaptive scaling to activations (Sun et al., 2021; Djurisic et al., 2022; Xu et al., 2023). Input perturbation methods such as ODIN (Liang et al., 2017), IODIN (Regmi, 2024), and ADASCALE (Regmi, 2025) use gradient information to modify the input at test time, but require additional forward and backward passes.

Most closely related to our approach are feature-space density methods. Some of these methods use Mahalanobis distance and model features as class-conditional Gaussians (Lee et al., 2018) with a background global Gaussian score (Ren et al., 2021). Mueller & Hein (2025) found that normalizing samples prior to fitting the Gaussian densities helps improve detection robustness across architectures. Related feature-space methods such as ViM (Wang et al., 2022) combine information from logits and intermediate representations, modeling ID structure via a principal subspace of features and using the residual norm to improve OOD separation. Other feature-space methods include

FDBD, which found that OOD samples reside closer to decision boundaries than ID samples (Liu & Qin, 2023), and methods inspired by the properties of neural collapse (Ammar et al., 2023; Liu & Qin, 2025). These approaches all avoid searching over the entire training set at inference, which makes them faster than the k-nearest neighbor approach proposed in Sun et al. (2022).

Training Methods for OOD Detection. A complementary approach for OOD detection is to modify the training process of the model so that OOD samples are easier to detect. Researchers have explored modifying the architecture itself, like in DeVries & Taylor (2018), where a separate confidence-based OOD detector branch was added to the model. Alternatively, many works have explored modifications to the training data. For example, adding OOD samples and augmentations has been shown to improve detection accuracy (Zhang et al., 2023c;a; Zhu et al., 2023; Pinto et al., 2022). Other methods have explored new training objectives, such as tuning logit activations (Xu et al., 2023; Wei et al., 2022), or using contrastive learning to improve ID and OOD separation (Seifi et al., 2024; Winkens et al., 2020; Tack et al., 2020; Ming et al., 2022). VSOOD is complementary to these methods, since it operates post-hoc and can be applied to pretrained models without retraining.

Vendi Scoring for Novelty. The Vendi Score, introduced in (Friedman & Dieng, 2022), has been used as an accurate measure of diversity across domains. Prior work has used the Vendi Score to evaluate dataset diversity in machine learning and the natural sciences (Nielsen et al., 2025; Bousso Dieng & Pasarkar, 2025; Roediger et al., 2025; Priyadarshini & Ganley; Bai et al., 2025), as well as to encourage diverse outputs in simulations, generative models, and reinforcement learning (Rezaei & Dieng, 2025a;b; Pasarkar et al., 2023; Askari Hemmat et al., 2024; Lintunen, 2025). Recent work has also explored Vendi-based novelty criteria. For example, Nguyen & Dieng (2024) and Liu et al. (2024) measure the incremental change in Vendi Score induced by adding a candidate point as a selection rule in active learning. More recently, Pasarkar & Dieng (2025) proposed the Vendiscope, a method to analyze the content of large-scale datasets. The Vendiscope quantifies sample rarity by optimizing over a distribution supported on the dataset points. This formulation cannot measure an unseen sample's rarity *against* a dataset, so it is not applicable to OOD detection.

3 The Vendi Score Out-of-Distribution Detector

In this section, we briefly review the Vendi Scores (VS), and introduce our proposed algorithm to make them effective for OOD detection.

The Vendi Scores. Consider a set of N data points $\boldsymbol{X} \in \mathbb{R}^{N \times D}$. Let $k(\cdot, \cdot)$ denote a positive semi-definite pairwise kernel. The VS, as defined in Friedman & Dieng (2022), is the entropy of the eigenvalues of the similarity matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ induced by the kernel k, normalized so that $\mathrm{Tr}(\boldsymbol{K}) = 1$. This can be generalized to the Rényi entropy (Pasarkar & Dieng, 2023)

$$\mathrm{VS}_q(\boldsymbol{X}, k) = \exp\left(\frac{1}{1-q} \log \sum_{i=1}^{N} \lambda_i(\boldsymbol{K})^q\right), \tag{1}$$

where $\lambda_i(\boldsymbol{K})$ is the ith eigenvalue of $\boldsymbol{K}$ and $q \geq 0$ is the order of the VS.

Computing the VS for arbitrary similarity matrices $\boldsymbol{K}$ requires computing their eigenspectrum, which requires $O(N^3)$ time. However, in the common setting where each data point is represented by an ℓ_2-normalized vector and the cosine kernel is used–as is often the case for learned embeddings–Friedman & Dieng (2022) note that we can compute the VS more cheaply. In this case, the similarity matrix is $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^T/N$, and its non-zero eigenvalues coincide with those of the matrix $\rho = \boldsymbol{X}^T\boldsymbol{X}/N \in \mathbb{R}^{D \times D}$. We can compute the VS using ρ, which reduces the time complexity of VS computation to $O(D^2N + D^3)$ and is efficient when $N \gg D$.

Certain choices of the order q have convenient forms. We highlight $q = 2$ and $q = \infty$:

$$\mathrm{VS}_2(\boldsymbol{X}, k) = \frac{1}{\mathrm{Tr}(\boldsymbol{K}^2)} = \frac{1}{\sum_{i,j=1}^{N} K_{ij}^2} \text{ and } \mathrm{VS}_\infty(\boldsymbol{X}, k) = \frac{1}{\lambda_{\max}(\boldsymbol{K})}. \tag{2}$$

Class-Conditional Novelty. Here, we describe how we use the VS to measure the novelty of a test sample with respect to a class. We consider a trained discriminative model f_θ (e.g., ResNet-50) with C classes, trained on a labeled in-distribution dataset $\mathcal{D}_{\text{train}} = \{(x_j, y_j)\}_{j=1}^{N}$. Let N_c denote the

number of training samples belonging to class c. Let $h(x) \in \mathbb{R}^D$ denote the representation of input x extracted from the penultimate layer of f_θ, and let $\boldsymbol{X}_c \in \mathbb{R}^{N_c \times D}$ be the matrix of embeddings $h(x)$ of all samples that belong to class c in $\mathcal{D}_{\text{train}}$. All representations are ℓ_2-normalized.

We first compute class-conditional matrices $\rho_c = \boldsymbol{X}_c^T \boldsymbol{X}_c / N_c \in \mathbb{R}^{D \times D}$. Given a test sample x, we consider the class-conditional matrix obtained after incorporating the representation $h(x)$ into class c, which we denote by ρ_c'. We denote by $\text{VS}_2(\rho_c)$ the VS of order 2 computed for class c under a cosine kernel. The class-conditional novelty is then defined as

$$\Delta_c(x) = \log \text{VS}_2(\rho_c') - \log \text{VS}_2(\rho_c). \tag{3}$$

$\Delta_c(x)$ measures the log-ratio increase in diversity from adding sample x to class c.

We can use Equation 2 to compute Δ_c efficiently, using the identity $\text{VS}_2(\rho_c) = \frac{1}{\text{Tr}(\rho_c^2)}$. To do so, the updated matrix after adding a test sample x can be written as $\rho_c' = \frac{N_c \rho_c + h(x)h(x)^T}{N_c + 1}$. Substituting this expression yields

$$\text{Tr}({\rho_c'}^2) = \frac{N_c^2 \,\text{Tr}(\rho_c^2) + 2N_c h(x)^T \rho_c h(x) + 1}{(N_c+1)^2}.$$

Denote the eigenvalues and eigenvectors of ρ_c as $\lambda_{c,1}, \lambda_{c,2}, \ldots, \lambda_{c,D}$ and $u_{c,1}, u_{c,2}, \ldots, u_{c,D}$. Let $\alpha_{c,j}(x) = (u_{c,j}^T h(x))^2$. We can rewrite the above equation for $\text{Tr}({\rho_c'}^2)$ in terms of this eigenbasis,

$$\text{Tr}({\rho_c'}^2) = \frac{N_c^2 \sum_{i=1}^D \lambda_{c,i}^2 + 2N_c \sum_{i=1}^D \lambda_{c,i} \alpha_{c,i}(x) + 1}{(N_c+1)^2}. \tag{4}$$

Plugging in $\text{Tr}(\rho_c'^2)$ into Equation 2 would give us an exact, closed-form, $O(D^2)$ time complexity operation to compute the class-conditional novelty scores. However, we find that maintaining the entire eigenbasis is unnecessary and can be sensitive to estimation noise in small classes. Instead, we can achieve strong results if we only maintain the top eigenvalue (Appendix Section C). Let λ_c denote the largest eigenvalue for class c, with corresponding eigenvector u_c. Then, we can compute the new trace in $O(D)$ using the rank-1 approximation

$$\alpha_c(x) := (u_c^T h(x))^2 \qquad \text{Tr}({\rho_c'}^2) \approx \frac{N_c^2 \lambda_c^2 + 2N_c \lambda_c \alpha_c(x) + 1}{(N_c+1)^2}.$$

The class-conditional novelty contribution is then

$$\Delta_c(x) = -\log\left(\frac{N_c^2 \lambda_c^2 + 2N_c \lambda_c \alpha_c(x) + 1}{(N_c+1)^2}\right) + \log(\lambda_c^2). \tag{5}$$

For each test sample x, we compute a vector of class-conditional novelty scores

$$\boldsymbol{\Delta}(x) = (\Delta_1(x), \Delta_2(x), \ldots, \Delta_C(x)). \tag{6}$$

Probability-weighted aggregation. The classifier f_θ produces a predictive distribution

$$\mathbf{p}(x) = (p_1(x), \ldots, p_C(x)), \quad p_c(x) = \Pr(y = c \mid x).$$

Rather than relying on the novelty with respect to a single predicted class, we aggregate class-conditional novelty scores using a probability-weighted scheme. Let $\mathcal{T}_K(x) \subseteq \{1, \ldots, C\}$ denote the indices of the K classes with largest predicted probabilities $p_c(x)$.

$$S_{\text{LOCAL-OOD}}(x) = \sum_{c \in \mathcal{T}_K(x)} N_c p_c(x)^\gamma \Delta_c(x), \tag{7}$$

where $\gamma \geq 0$ and $K \leq C$ are tunable hyperparameters.

Notably, equation 7 mirrors the form of GEN (Liu et al., 2023), which aggregates over $\mathcal{T}_K(x)$ using probability-based weights of the form $p_c(x)^\gamma (1 - p_c(x))^\gamma$; in contrast, we use $p_c(x)^\gamma$ to weight class-conditional novelty terms $\Delta_c(x)$. The exponent γ controls the sharpness of the aggregation: larger values emphasize high-confidence classes, while smaller values yield a softer, more global aggregation across classes. The parameter K controls the number of ID classes included in the aggregation (selected as the top-K classes by $p_c(x)$). While increasing K can incorporate information

when predictions are ambiguous, overly large values may degrade performance by including many low-probability classes whose class-conditional novelty scores are weakly related to the sample and introduce noise into the aggregation. Smaller values of K also enable cheaper computation.

We additionally rescale $\Delta_c(x)$ by N_c to account for class-size. Appendix Section A.1 shows that the class-conditional novelty score scales by $O(1/N_c)$, so we remove the effect of class-size with this multiplicative factor.

Incorporating Global Diversity. Previous work has highlighted the importance of modeling the local density of model representations as well as the global density. Ren et al. (2019) and Ren et al. (2021) showed that subtracting a global novelty or likelihood score from the local novelty score can yield more robust results. Following this idea, we construct a global background novelty score using a first-order approximation of the effect of a test sample x on the VS_∞ computed on the entire dataset.

Let $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ denote the matrix whose rows are the ℓ_2-normalized embeddings $h(x)$ of all training samples, with $\rho_{\text{global}} = \boldsymbol{X}^T\boldsymbol{X}/N \in \mathbb{R}^{D\times D}$. Given a test sample x, we denote the updated global matrix as ρ'_{global} and define the global novelty as

$$\Delta_{\text{global}}(x) = \log \text{VS}_\infty(\rho'_{\text{global}}) - \log \text{VS}_\infty(\rho_{\text{global}}). \tag{8}$$

VS_∞ only depends on the maximum eigenvalue of ρ_g (Equation 2). We approximate the effect of x on the largest eigenvalue $\lambda_{\max}$ using the following proposition, with proof in Section A.2.

Proposition 3.1 (Accuracy of the Max Eigenvalue Update) *Let $\rho_{global} \in \mathbb{R}^{D\times D}$ be a symmetric positive semidefinite matrix with a unique largest eigenvalue $\lambda_{\max}(\rho_{global})$ and corresponding unit-norm eigenvector $u_{\max}$. For a unit-norm vector $h(x) \in \mathbb{R}^D$ and dataset size N, define the rank-one updated matrix*

$$\rho'_{global} = \frac{1}{N+1}\left(N\rho_{global} + h(x)h(x)^T\right).$$

Then the largest eigenvalue of ρ'_{global} admits the first-order expansion

$$\lambda_{\max}(\rho'_{global}) = \frac{N\lambda_{\max}(\rho_{global}) + (u_{\max}^T h(x))^2}{N+1} + O\left(\frac{1}{N^2}\right).$$

Equivalently, the estimator

$$\widehat{\lambda}_{\max}(\rho'_{global}) := \frac{N\lambda_{\max}(\rho_{global}) + (u_{\max}^T h(x))^2}{N+1}$$

approximates $\lambda_{\max}(\rho'_{global})$ with error $O(1/N^2)$.

Using the estimator $\widehat{\lambda}_{\max}(\rho'_{\text{global}})$ from Proposition 3.1, we have the following approximation for the change in VS_∞

$$\Delta_{\text{global}}(x) \approx -\log\left(\widehat{\lambda}_{\max}(\rho'_{\text{global}})\right) + \log\lambda_{\max}(\rho_{\text{global}}) \tag{9}$$

We note that Equation 5 used for the local diversity and Equation 9 used for the global diversity have seemingly similar forms, though Equation 5 is better suited for local diversity measurements because it better handles the variability in class's max eigenvalues. Appendix Section A.3 further discusses the differences in these two approaches.

We also scale the global novelty by multiplying by the total dataset size N:

$$S_{\text{GLOBAL-OOD}}(x) = -N\log\left(\frac{N\lambda_{\max}(\rho_{\text{global}}) + \alpha(x)}{N+1}\right) + N\log\lambda_{\max}(\rho_{\text{global}}).$$

Because subtracting global density does not always yield gains, we introduce a binary coefficient variable $g \in \{0,1\}$, yielding the full VSOOD detector score:

$$\text{VSOOD}(x) := S_{\text{LOCAL-OOD}}(x) - gS_{\text{GLOBAL-OOD}}(x).$$

Method	CIFAR-10 OpenOOD Benchmark							CIFAR-100 OpenOOD Benchmark						
	Near-OOD		Far-OOD				AVG	Near-OOD		Far-OOD				AVG
	CIFAR-100	TIN	MNIST	SVHN	Texture	Place365		CIFAR-10	TIN	MNIST	SVHN	Texture	Place365	
Evaluation under FPR (↓)														
MSP	53.10	43.26	23.64	25.81	34.96	42.47	37.04	**58.90**	50.70	57.24	59.07	61.88	56.62	57.40
ASH	87.31	86.29	70.00	83.64	84.60	77.86	81.62	68.07	63.37	66.60	**45.97**	61.29	62.94	61.37
SCALE	81.78	79.12	48.69	70.54	80.39	70.51	71.84	59.11	52.24	51.64	49.27	58.44	56.98	54.61
ADASCALE	78.93	74.89	42.20	59.97	74.75	65.90	66.11	59.27	51.94	51.25	47.41	58.39	56.41	54.11
GEN	58.77	48.57	23.00	28.14	40.73	47.06	41.05	**58.87**	49.97	53.93	55.45	61.22	56.25	55.95
IODIN	61.32	50.20	25.86	31.80	43.36	49.19	43.64	59.08	51.57	52.91	54.07	62.07	57.46	56.19
KPCA	41.07	32.83	19.79	22.68	**22.96**	30.76	28.35	73.54	57.01	**45.43**	**40.09**	**49.11**	61.09	54.38
FDBD	39.60	31.04	**19.33**	22.89	24.28	**29.12**	27.71	63.89	**47.89**	51.35	53.80	53.65	57.16	54.62
NCI	52.47	42.92	28.93	31.71	27.58	35.64	36.54	63.59	**48.59**	51.14	48.36	**47.76**	**53.93**	**52.23**
KNN	**37.62**	**30.38**	20.04	**22.62**	24.06	30.38	**27.52**	72.81	49.66	**48.57**	51.75	53.56	60.70	56.18
RMDS++	41.11	31.27	21.99	23.60	24.60	**28.64**	28.53	61.43	48.90	79.37	52.49	52.40	54.78	53.82
VSOOD	**39.48**	**30.27**	**16.78**	**19.68**	**23.44**	31.74	**26.90**	60.34	48.82	50.00	49.32	52.75	**54.04**	**52.55**
Evaluation under AUROC (↑)														
MSP	87.19	88.87	92.63	91.46	89.89	88.92	89.83	78.47	82.07	76.09	78.42	77.32	79.23	78.60
ASH	74.10	76.44	83.16	73.45	77.45	79.89	77.42	76.47	79.92	77.23	**85.60**	80.72	78.76	79.78
SCALE	81.27	83.84	90.58	84.63	83.94	86.41	85.11	**79.26**	82.71	80.27	84.45	80.50	80.47	81.28
ADASCALE	82.69	85.16	91.57	87.32	85.61	87.49	86.64	79.22	82.82	80.39	85.27	81.01	80.68	81.57
GEN	87.21	89.20	93.83	91.97	90.14	89.46	90.30	**79.38**	83.25	78.29	81.41	78.74	80.28	80.23
IODIN	86.87	88.96	93.45	91.55	89.78	89.16	89.96	79.24	82.96	78.89	81.56	78.48	79.83	80.16
KPCA	89.41	91.28	94.46	**92.83**	**93.77**	**92.05**	**92.30**	71.22	75.71	**84.64**	**89.12**	**86.17**	76.30	80.53
FDBD	89.56	**91.60**	**94.71**	92.80	93.13	92.01	**92.30**	78.35	**83.97**	79.05	80.48	81.18	79.85	80.48
NCI	87.84	89.50	92.08	90.67	91.97	90.36	90.43	78.31	83.55	79.89	83.01	**83.75**	80.86	81.56
KNN	**89.73**	91.56	94.26	92.67	93.16	91.77	92.19	77.02	83.34	**82.36**	84.15	83.66	79.43	**81.66**
RMDS++	88.83	90.68	92.99	92.61	92.15	91.36	91.44	78.08	82.72	79.37	83.95	82.56	**83.01**	81.62
VSOOD	**89.97**	**92.39**	**95.94**	**94.43**	**93.90**	**92.17**	**93.13**	78.70	**84.09**	80.99	84.47	83.06	**82.15**	**82.29**

Table 1: VSOOD achieves state-of-the-art performance on CIFAR-10 and is competitive on CIFAR-100 OpenOOD benchmarks. **Orange**, **blue**, and **purple** text mark the best, second-best, and third-best results.

We use VSOOD(x) as a continuous OOD score, where larger values indicate more novel samples. To perform OOD detection, this score can be thresholded; throughout our experiments, we choose a threshold that achieves a fixed true positive rate of 95%, following standard evaluation practice. Additional details are provided in Section 4.

Time and Space Complexity. Due to our rank-1 approximation in Equation 5, VSOOD has an inference space and time complexity of $O(CD)$, which scales linearly with the number of classes C and the dimension of the data D. Logit-based baselines such as MSP (Hendrycks & Gimpel, 2016), as well as NCI (Liu & Qin, 2025), which leverages neural-collapse structure in the feature space, are still faster. However, VSOOD is more efficient than Mahalanobis-style methods, which are $O(CD^2)$ (Lee et al., 2018; Ren et al., 2021; Mueller & Hein, 2025).

4 Experiments

In this section, we compare VSOOD against a diverse set of state-of-the-art OOD detection algorithms. We benchmark all algorithms on three image classification datasets: CIFAR-10, CIFAR-100, and ImageNet-1K, as well as on ResNet, Swin-T, and ViT architectures. Overall, VSOOD exhibits strong and consistent performance across datasets and models.

Datasets and Models. Following the OpenOOD benchmark (Zhang et al., 2023b), we use a standardized set of OOD test sets for each ID benchmark. For CIFAR-10 and CIFAR-100, we use six OOD test sets comprising two Near-OOD (hard) and four Far-OOD datasets (easy). For CIFAR-10, we use Tiny ImageNet (TIN) (Le & Yang, 2015), MNIST (Deng, 2012), SVHN (Netzer et al., 2011), Texture (Cimpoi et al., 2014), and Places365 (Zhou et al., 2017), and treat CIFAR-100 as an OOD test set. For CIFAR-100, we use the same five OOD datasets and treat CIFAR-10 as an OOD test set. For both benchmarks, we evaluate using three pretrained ResNet-18 models from OpenOOD and report the average across models. For ImageNet-1K, we use two Near-OOD datasets (SSB-Hard (Vaze et al., 2021) and NINCO (Bitterwolf et al., 2023)) and three Far-OOD datasets (iNaturalist (Van Horn et al., 2018), Texture (Cimpoi et al., 2014), and OpenImage-O (Wang et al., 2022)). We evaluate pretrained ResNet-50 (He et al., 2016), ViT-B/16 (Dosovitskiy, 2020), and Swin-T (Liu et al., 2021) checkpoints, using 2048-dimensional penultimate-layer features from ResNet-50 and 768-dimensional features from ViT-B/16 and Swin-T.

Method	ImageNet-1K OpenOOD Benchmark (Average Across Architectures)					
	Near-OOD		Far-OOD			AVG
	SSB-Hard	Ninco	iNaturalist	Textures	OpenImage-O	
Evaluation under FPR (↓)						
MSP	80.60	65.12	41.06	59.41	51.87	59.61
ASH	87.62	80.91	68.45	69.98	72.57	75.91
SCALE	83.00	76.89	57.00	62.63	68.53	69.61
ADASCALE	**76.43**	58.96	25.30	38.64	33.61	46.59
GEN	79.04	54.08	23.18	43.03	34.23	46.71
IODIN	81.14	65.78	36.63	54.59	47.65	57.16
KPCA	95.40	83.98	55.79	41.97	69.01	69.23
FDBD	84.75	58.11	27.91	38.77	35.47	49.00
NCI	82.99	57.54	25.64	36.70	37.06	47.99
KNN	88.51	62.36	41.06	**27.50**	47.16	53.32
RMDS++	83.19	**49.79**	**18.64**	41.62	**32.24**	**45.10**
VSOOD	**76.12**	**47.13**	**12.88**	**30.26**	**28.58**	**38.99**
Evaluation under AUROC (↑)						
MSP	70.94	79.93	88.82	83.59	85.16	81.69
ASH	57.55	61.01	64.77	62.36	64.80	62.10
SCALE	66.15	73.06	83.44	81.68	79.22	76.71
ADASCALE	71.66	82.56	93.00	90.53	**90.98**	85.74
GEN	71.62	83.13	93.40	88.66	90.03	85.37
IODIN	71.07	80.13	89.79	85.56	86.61	82.64
KPCA	44.41	62.44	79.04	86.79	72.53	69.04
FDBD	66.65	80.81	91.53	89.95	89.60	83.71
NCI	68.83	82.15	93.00	90.79	90.15	84.98
KNN	57.45	76.45	84.38	**93.69**	83.61	79.12
RMDS++	**72.59**	**86.09**	**95.30**	89.26	90.91	**86.83**
VSOOD	**75.17**	**86.16**	**96.71**	**91.96**	**92.57**	**88.51**

Table 2: VSOOD achieves state-of-the-art performance on the OpenOOD ImageNet-1K benchmark. Results are averaged across 3 models (ViT-B/16, ResNet-50, Swin-T). **Orange** marks best, **blue** marks 2nd, and **purple** marks 3rd.

Model	GEN	NCI	ADASCALE	RMDS++	VSOOD
Swin-T Latency	1.19	1.19	4.72	18.54	1.48
ResNet-50 Latency	0.59	0.59	2.50	64.66	0.78
ViT-B/16 Latency	2.20	2.20	9.85	14.19	2.39

Table 3: Latency (ms per image) comparison of top-performing post-hoc OOD detectors on ImageNet-1K models. Latency includes model inference, with optimal batch sizes picked for each method. VSOOD has latency comparable to logit-only methods while outperforming them on the tested benchmarks. Measurements taken on single NVIDIA A6000 GPU.

Baselines. We compare against 11 OOD detection algorithms: MSP (Hendrycks & Gimpel, 2016), ASH (Djurisic et al., 2022), SCALE (Xu et al., 2023), ADASCALE (Regmi, 2025), GEN (Liu et al., 2023), IODIN (Regmi, 2024), FDBD (Liu & Qin, 2023), NCI (Liu & Qin, 2025), KNN (Sun et al., 2022), KPCA (Fang et al., 2024), and RMDS++ (Mueller & Hein, 2025). These baselines span a range of widely used post-hoc OOD paradigms: activation shaping (ASH, SCALE, ADASCALE, and IODIN), logit confidences (MSP, GEN), and feature geometry (KPCA, FDBD, NCI, KNN, and RMDS++). We provide a description of each algorithm, and how they compare to VSOOD in Appendix Section B.1.

Hyperparameter Selection. MSP, FDBD, and RMDS++ are all hyperparameter-free methods and do not require tuning. For all other methods, we follow prior work (Liu & Qin, 2025) and the OpenOOD benchmark to automatically select hyperparameters. Specifically, we construct a validation set consisting of ID samples from the training data mixed with synthetic OOD Gaussian noise

Benchmark		Method			
Dataset	**Architecture**	**GEN+ReAct**	**KPCA+GEN**	**Hybrid RMDS++**	**VSOOD**
CIFAR-10	ResNet-18	88.59	**92.51**	91.21	**93.13**
CIFAR-100	ResNet-18	80.47	**82.80**	81.55	**82.29**
ImageNet-1K	ResNet-50	87.07	87.43	**88.10**	**89.26**
ImageNet-1K	ViT-B/16	85.19	84.74	**88.35**	**87.48**
ImageNet-1K	Swin-T	85.73	85.43	**88.54**	**88.79**

Table 4: Average AUROC of hybrid OOD detectors on all tested OpenOOD benchmarks. **Orange** and **blue** mark the best and second-best methods in each row.

images, where each pixel is sampled from $\mathcal{N}(0, 1)$. This protocol provides a proxy for selecting hyperparameters without access to real OOD data.

VSOOD introduces three hyperparameters: K, the number of classes in Equation 7, γ, the probability exponent in Equation 7, and g, a binary variable for if the global density correction should be applied. For K, we sweep over $\{|C|/100, |C|/40, |C|/20, |C|/10\}$, for γ, we search over $\{0.1, 0.5, 1., 2.\}$ and for g, we search over $\{0, 1\}$. We use the hyperparameter triplet (γ, K, g) with the highest AUROC on the validation set.

Evaluation Metrics. We mix each OOD test set with the ID test set, and ask each postprocessor to compute an ID confidence score. Using the scores computed from each method, we compute (1) False Positive Rate under a true positive rate of 95% (FPR@95) and (2) Area under the ROC curve (AUROC), which measures the probability that a randomly chosen ID sample is deemed less novel than a randomly chosen OOD sample.

4.1 OOD Detection Performance Results

CIFAR-10 and CIFAR-100. On CIFAR-10, VSOOD achieves the best average performance across OOD test sets in terms of both FPR@95 and AUROC (Table 1). FDBD and KNN provide the second and third best performance on CIFAR-10. On CIFAR-100, VSOOD and RMDS++ (Mueller & Hein, 2025) are the only methods that place among the top performances across both metrics, with VSOOD reducing FPR@95 by 2% relative to RMDS++.

ImageNet-1K. Table 2 summarizes ImageNet-1K OOD detection results. We average across all tested architectures, with results for individual models shown in Tables 6, 7, and 8. On average, VSOOD reduces the FPR@95 by 13% relative to the second-best performing algorithm. We note that on individual architectures, other algorithms may perform better. For example, the activation shaping methods SCALE (Xu et al., 2023) and ADASCALE (Regmi, 2025) perform better than VSOOD on the ResNet architecture (Table 7). However, these gains do not consistently transfer across architectures. Xu et al. (2023) argue for the effectiveness of scaling activations under the assumption that the mean of ID activations is higher than that of OOD activations. This claim does not necessarily hold for transformer-based architectures, where these methods struggle.

Table 3 shows that VSOOD is computationally efficient despite using representation-level information. Across tested architectures, VSOOD is up to 80× faster than the second-most accurate method RMDS++, and only slightly slower than the fastest detectors GEN and NCI. Thus, VSOOD retains the accuracy benefits of representation-based OOD detection while avoiding the large test-time overhead of feature-search methods. The efficiency of VSOOD is possible due to its rank-1 approximation. We show in Section C that using only rank-1 preserves accuracy. We also provide hyperparameter ablations for VSOOD in Appendix Section C.

4.2 Comparison of Hybrid Approaches

VSOOD provides a hybrid approach of combining predictive uncertainty from logits with a class-conditional novelty score computed in feature space. As many other baselines only leverage one set of information, we evaluate VSOOD against an additional set of hybrid detectors.

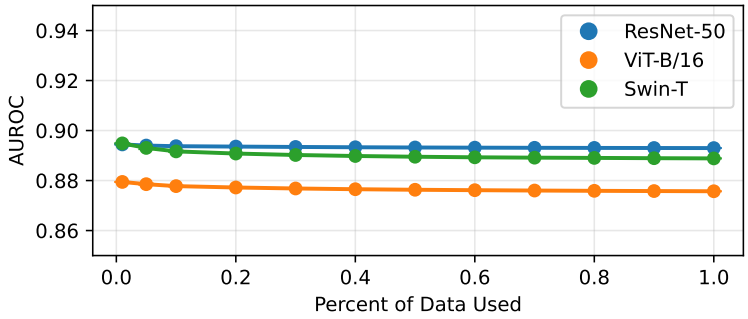

Figure 2: VSOOD provides accurate OOD detection even with access to only 1% of training data. AUROC is shown as a function of the percent of the training data used. Results across three ImageNet-1K models are displayed, averaged across three reproductions.

One baseline is to combine RMDS++ with our probability-weighting scheme. We follow Mueller & Hein (2025) and compute a Mahalanobis distance between a normalized test sample to the class mean. The novelty score function then becomes

$$s_{\text{Maha}}(x) = \sum_{i \in \mathcal{T}_K(x)} p_c(x)^\gamma d_{\text{Maha}}(x, \mu_c) - g d_{\text{Maha}}(x, \mu),$$

where d_{Maha} denotes the Mahalanobis distance. We also compare VSOOD against a hybrid of KPCA and the logit-only method GEN, multiplying the scores from the two methods, and a hybrid of GEN and ReAct (Sun et al., 2021). Results are shown in Table 4. The Hybrid-RMDS++ approach provides the most competitive performance to VSOOD, although there is a larger performance gap on the CIFAR-10/100 benchmarks and RMDS++ can be over $80\times$ slower than VSOOD (Table 3).

4.3 Data-efficient OOD Detection

A major advantage of activation shaping and logit-based methods over other approaches is that they do not require access to a model's training data. This is useful in contexts where the training data is not available, or is too expensive to store. While VSOOD cannot operate without access to the training data, we show that VSOOD can operate effectively with limited access. We sample varying percentages of samples from each training ImageNet class and measure VSOOD's performance after being fit on this subset, using the same hyperparameters throughout. In Figure 2, we see that even at tiny fractions, VSOOD provides accurate OOD detection. With 1% of the data, VSOOD achieves an average AUROC of 88.96 ± 0.72 across models, compared to 88.59 ± 0.74 when given all of the data. VSOOD's robustness to dataset size is likely due to two factors: (1) VSOOD incorporates logit information, which is unaffected by dataset size and (2) VSOOD only utilizes the largest eigenvalue from each class, which can be reliably estimated from small sample sizes.

5 Conclusion

This work introduces the Vendi Score Out-of-Distribution detector (VSOOD), a post-hoc OOD detector that leverages the Vendi Scores to quantify novelty from a diversity perspective. VSOOD scores a test sample by measuring its class-conditional novelty using an efficient rank-1 approximation of the Vendi Score of order 2, aggregates these signals via a probability-weighted top-K scheme, and applies a data-dependent global background correction to improve robustness. Across multiple image classification benchmarks and architectures, VSOOD often achieves state-of-the-art detection performance. On ImageNet-1K, VSOOD reduces average FPR@95 by 13% relative to the strongest tested baseline, and remains effective when computed using only 1% of the training set.

Limitations. VSOOD introduces three tunable hyperparameters, which we select using an automatically generated validation set of random noise images. However, this validation set may not be representative of the samples encountered during deployment. In addition, for scalability, VSOOD relies on the cosine kernel, whereas alternative kernels may improve results.

AI USE STATEMENT

In this work, we used generative AI tools to help implement methods and create our evaluation setup. We did not use these tools for developing the theoretical models or conceptual frameworks, formulating or proving mathematical claims, assisting in the writing of proofs, proposing hypothesis, designing the methodology, performing qualitative and thematic data analysis, or for result interpretation. Other required disclosure tasks such as generating synthetic datasets, assisting with translation, or cleaning datasets are not relevant to this work. Additionally, we used generative AI tools to help create the cleaned code repository. We have reviewed all AI-assisted code by ensuring that for all benchmarks, we are able to reproduce publicly available performance metrics. Code was also tested for correctness by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work improves OOD detection, helping ML systems identify inputs beyond their competence and trigger safer responses such as abstention, human escalation, or fallback procedures. However, they are not complete safety solutions and can fail under adversarial or shifting conditions. OOD detectors should therefore be used within broader safety protocols with monitoring, human oversight, and evaluation across diverse environments.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, all code used to generate our benchmark results is provided alongside our submission. Our code follows the structure of the OpenOOD benchmark to make it easier for researchers to integrate it into their workflows. This also allows us to follow standard protocols for accessing model checkpoints, data processing, and comparing baseline methods. For all mathematical claims in our paper, we provide the relevant proofs in the appendix.

REFERENCES

Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. *arXiv preprint arXiv:2310.06823*, 2023.

Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized vendi score guidance. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2024.

Yatong Bai, Jonah Casebeer, Somayeh Sojoudi, and Nicholas J Bryan. Dragon: Distributional rewards optimize diffusion generative models. *arXiv preprint arXiv:2504.15217*, 2025.

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023.

Adji Bousso Dieng and Amey Pasarkar. A unified and predictive measure of functional diversity. *arXiv e-prints*, pp. arXiv–2509, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

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. Ieee, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kun Fang, Qinghua Tao, Kexin Lv, Mingzhen He, Xiaolin Huang, and Jie Yang. Kernel pca for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:134317–134344, 2024.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

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

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16783–16792, 2022.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pp. 10848–10865. PMLR, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Erik M Lintunen. Vendirl: A framework for self-supervised reinforcement learning of diversely diverse skills. *arXiv preprint arXiv:2509.02930*, 2025.

Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. *arXiv preprint arXiv:2312.11536*, 2023.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15424–15433, 2025.

Tsung-Wei Liu, Quan Nguyen, Adji Bousso Dieng, and Diego A Gómez-Gualdrón. Diversity-driven, efficient exploration of a mof design space to optimize mof properties. *Chemical Science*, 15(45):18903–18919, 2024.

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

Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23946–23955, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022.

Maximilian Mueller and Matthias Hein. Mahalanobis++: Improving ood detection via feature normalization. *arXiv preprint arXiv:2505.18032*, 2025.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, 2011.

Quan Nguyen and Adji Bousso Dieng. Quality-weighted vendi scores and their application to diverse experimental design. *arXiv preprint arXiv:2405.02449*, 2024.

Bjarke Frost Nielsen, Amey P Pasarkar, Qiqi Yang, Bryan T Grenfell, and Adji Bousso Dieng. Applications of the vendi score in genomic epidemiology. *arXiv preprint arXiv:2509.22520*, 2025.

Amey P Pasarkar and Adji Bousso Dieng. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. *arXiv preprint arXiv:2310.12952*, 2023.

Amey P Pasarkar and Adji Bousso Dieng. The vendiscope: An algorithmic microscope for data collections. *arXiv preprint arXiv:2502.10828*, 2025.

Amey P Pasarkar, Gianluca M Bencomo, Simon Olsson, and Adji Bousso Dieng. Vendi sampling for molecular simulations: Diversity as a force for faster convergence and better exploration. *The Journal of chemical physics*, 159(14), 2023.

Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. *Advances in neural information processing systems*, 35:14608–14622, 2022.

Maitreyee Sharma Priyadarshini and Connor Ganley. Diversity-driven training of machine-learned force fields. In *AI for Accelerated Materials Design-NeurIPS 2025*.

Sudarshan Regmi. Image-based outlier synthesis with training data. *arXiv preprint arXiv:2411.10794*, 2024.

Sudarshan Regmi. Adascale: Adaptive scaling for ood detection. *arXiv preprint arXiv:2503.08023*, 2025.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

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.

Mohammad Reza Rezaei and Adji Bousso Dieng. The alpha-alternator: Dynamic adaptation to varying noise levels in sequences using the vendi score for improved robustness and performance. *arXiv preprint arXiv:2502.04593*, 2025a.

Mohammad Reza Rezaei and Adji Bousso Dieng. Vendi-rag: Adaptively trading-off diversity and quality significantly improves retrieval augmented generation with llms. *arXiv preprint arXiv:2502.11228*, 2025b.

Sven Roediger, Matthew Sigman, and Abigail Doyle. Machine learning-guided scope selection to balance performance and substrate similarity. 2025.

Nofit Segal, Aviv Netanyahu, Kevin P Greenman, Pulkit Agrawal, and Rafael Gómez-Bombarelli. Known unknowns: Out-of-distribution property prediction in materials and molecules. *npj Computational Materials*, 11(1):345, 2025.

Soroush Seifi, Daniel Olmeda Reino, Nikolay Chumerin, and Rahaf Aljundi. Ood aware supervised contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1956–1966, 2024.

Youssef Shoeb, Azarm Nowzad, and Hanno Gottschalk. Out-of-distribution segmentation in autonomous driving: Problems and state of the art. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4310–4320, 2025.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34:144–157, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

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? 2021.

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.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pp. 23631–23644. PMLR, 2022.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. *arXiv preprint arXiv:2310.00227*, 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5531–5540, 2023a.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023b.

Oliver Zhang, Jean-Benoit Delbrouck, and Daniel L Rubin. Out of distribution detection for medical images. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pp. 102–111. Springer, 2021.

Yifan Zhang, Zhiquan Tan, Jingqin Yang, Weiran Huang, and Yang Yuan. Matrix information theory for self-supervised learning. *arXiv preprint arXiv:2305.17326*, 2023c.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Jianing Zhu, Yu Geng, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Bo Han. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. *Advances in neural information processing systems*, 36:22702–22734, 2023.

A Additional Algorithm Details for VSOOD

A.1 Scaling Novelty Scores to Account for Dataset Size

When computing the local and global novelty scores in Equation 5 and Equation 9, we scale by the number of samples. This removes the effect of dataset size because the increment Δ scales according to $\frac{1}{N}$. We show this is true for both the local Δ_c in 5 and Δ_{global} in Equation 9.

For the class-conditional novelty Δ_c, we can use a Taylor expansion around the terms in Equation 5 to show the scaling behavior of the score.

$$\begin{aligned}
\Delta_c(x) &= -\log\left(\frac{N_c^2\lambda_c^2 + 2N_c\lambda_c\alpha_c(x) + 1}{(N_c+1)^2}\right) + \log(\lambda_c^2) \\
&= -\log\left(\frac{N_c^2 + 2N_c\alpha(x)/\lambda_c + 1/\lambda_c^2}{(N_c+1)^2}\right) \\
&= 2\log(1+\frac{1}{N_c}) - \log\left(1 + \frac{2\alpha(x)}{N_c\lambda_c} + \frac{1}{N_c^2\lambda^2}\right) \\
&= 2\left(\frac{1}{N_c} - \frac{1}{2N_c^2} + O(\frac{1}{N_c^3})\right) - \left(\frac{2\alpha(x)}{\lambda_c N_c} + (\frac{1}{\lambda^2} - \frac{2\alpha(x)^2}{\lambda_c^2})\frac{1}{N_c^2} + O(\frac{1}{N_c^3})\right) \\
&= (2 - \frac{2\alpha(x)}{\lambda_c})\frac{1}{N_c} + O(\frac{1}{N_c^2}),
\end{aligned}$$

where we used the two Taylor expansions $\log(1+u) = u - \frac{u^2}{2} + O(u^3)$ and $\log(1+Au+Bu^2) = Au + (B - \frac{A^2}{2})u^2 + O(u^3)$. With the above, we have

$$N_c\Delta_c(x) = 2 - \frac{2\alpha(x)}{\lambda_c} + O(\frac{1}{n_c}),$$

which is independent of the dataset size at the leading order.

For the global novelty, we again use a Taylor expansion around the terms in Equation 9 to show the scaling behavior of the score.

$$\begin{aligned}
\Delta_{\text{global}} &= -\log\left(\frac{N\lambda_{\max}(\rho) + \alpha(x)}{N+1}\right) + \log\lambda_{\max}(\rho) \\
&= \log\left(1 + \frac{1}{N}\right) - \log\left(1 + \frac{\alpha(x)}{N\lambda}\right) \\
&= \frac{1}{N} - \frac{\alpha(x)}{N\lambda} + O(\frac{1}{N^2})
\end{aligned}$$

Therefore, scaling by a factor of N gives us a score independent of the dataset size at the leading order.

$$N\Delta_{\text{global}} = 1 - \frac{\alpha(x)}{\lambda} + O(\frac{1}{N})$$

A.2 Proof of Proposition 3.1

As defined in Eq. 2, the log-VS of order ∞ is given by

$$\log \text{VS}_\infty(\rho_{\text{global}}) = -\log\lambda_{\max}(\rho_{\text{global}}),$$

where $\lambda_{\max}(\rho_{\text{global}})$ is the largest eigenvalue of $\rho_{\text{global}} = X^TX/N \in \mathbb{R}^{D\times D}$. Let $u_{\max}$ be the corresponding eigenvector to $\lambda_{\max}(\rho_{\text{global}})$. Adding a new sample x to ρ_{global} gives an updated density matrix $\rho'_{\text{global}} = \frac{1}{N+1}(N\rho_{\text{global}} + h(x)h(x)^T)$. The top eigenvalue of ρ' satisfies the Rayleigh principle

$$\lambda_{\max}(\rho'_{\text{global}}) = \max_{||v||=1} v^T\rho'_{\text{global}}v.$$

Because the old top eigenvector has $||u_{\max}|| = 1$, we can use it as a lower bound:

$$\begin{aligned}\lambda_{\max}(\rho'_{\text{global}}) &\geq u_{\max}^T \rho' u_{\max} \\ &= u_{\max}^T \left(\frac{N\rho_{\text{global}} + h(x)h(x)^T}{N+1} \right) u_{\max} \\ &= \frac{N u_{\max}^T \rho_{\text{global}} u_{\max} + u_{\max}^T h(x)h(x)^T u_{\max}}{N+1} \\ &= \frac{N\lambda_{\max}(\rho_{\text{global}}) + \alpha(x)}{N+1},\end{aligned}$$

We plug the above expression into Eq. 8 to reach Eq. 9.

We can use a first-order perturbation expansion to derive the accuracy of our approximation. Let $\delta = \rho'_{\text{global}} - \rho_{\text{global}} = \frac{1}{N+1}\left(h(x)h(x)^T - \rho_{\text{global}}\right)$. Then,

$$\begin{aligned}\lambda_{\max}(\rho'_{\text{global}}) &= \lambda_{\max}(\rho_{\text{global}}) + u_{\max}^T \delta u_{\max} + O(||\delta||^2) \\ &= \lambda_{\max}(\rho_{\text{global}}) + \frac{(u_{\max}^T h(x))^2 - \lambda_{\max}(\rho_{\text{global}})}{N+1} + O(\frac{1}{N^2}) \\ &= \frac{N\lambda_{\max}(\rho_{\text{global}}) + (u_{\max}^T h(x))^2}{N+1} + O(\frac{1}{N^2}).\end{aligned}$$

So, our approximation to the max eigenvalue is accurate to the order of $O(1/N^2)$.

A.3 Comparison of Local and Global Novelty Scores

Here we compare Equation 5 against Equation 9. We show that despite their reliance on only the maximum eigenvector and eigenvalue, our VS_2 rank-1 approximation for measuring local diversity is better suited than our first-order VS_∞ approximation, while VS_∞ provides an accurate global correction.

In our rank-1 approximation in Equation 5, the data-dependent part of Δ_c can be written as

$$\Delta_c^{(2)}(x) \;=\; -\log\left(1 + \frac{2\alpha_c(x)}{N_c\lambda_c} + \frac{1}{N_c^2\lambda_c^2}\right) + \text{Constant} \tag{10}$$

whereas the analogous $q = \infty$ change (under our first-order approximation) has the form

$$\Delta_c^{(\infty)}(x) \;=\; -\log\left(1 + \frac{\alpha_c(x)}{N_c\lambda_c}\right) + \text{Constant} \tag{11}$$

To understand why $q = 2$ is preferable locally, we compare derivatives with respect to the quadratic alignment $\alpha_c(x) = (u_c^T h(x))^2$. Let $b_c := (N_c\lambda_c)^{-1}$. Differentiating equation 10 gives

$$\left|\frac{\partial}{\partial\alpha}\Delta_c^{(2)}\right| = \frac{2b_c}{1 + 2b_c\alpha + b_c^2}, \qquad \text{and in particular} \qquad \left|\frac{\partial}{\partial\alpha}\Delta_c^{(2)}\right|_{\alpha=0} = \frac{2b_c}{1+b_c^2}. \tag{12}$$

In contrast, differentiating equation 11 yields

$$\left|\frac{\partial}{\partial\alpha}\Delta_c^{(\infty)}\right| = \frac{b_c}{1 + b_c\alpha}, \qquad \text{and in particular} \qquad \left|\frac{\partial}{\partial\alpha}\Delta_c^{(\infty)}\right|_{\alpha=0} = b_c. \tag{13}$$

The crucial distinction is that at $\alpha = 0$, the $q = \infty$ local sensitivity is linear in $b_c = 1/(N_c\lambda_c)$ and can therefore vary substantially across classes due to variation in λ_c (and, to a lesser extent, N_c), making the local novelty overly sensitive to eigenvalue estimation error in small classes. By contrast, the $q = 2$ local sensitivity is self-normalized:

$$\left|\frac{\partial}{\partial\alpha}\Delta_c^{(2)}\right|_{\alpha=0} = \frac{2b_c}{1+b_c^2} \;\leq\; 1,$$

which suppresses both extremes $b_c \ll 1$ and $b_c \gg 1$. Equivalently, $q = 2$ yields a bounded and smoother response as λ_c varies, preventing a small subset of classes with atypical λ_c from dominating the probability-weighted aggregation. This robustness is particularly valuable for local scoring because class-level statistics are estimated from substantially fewer samples than the global statistics.

At the global-level, $\frac{1}{N\lambda}$ is reliably estimated so we do not need the self-normalization behavior of VS_2.

B Additional Experiment Details

B.1 Baselines

Here we briefly describe the 11 baselines, and how they compare in approach to VSOOD. All baselines were run on a single NVIDIA A6000 GPU with 4 Intel Xeon Gold 5320, 2.20GHz CPUs. All experimental results were generated over approximately 200 GPU hours.

MSP. The maximum softmax probability (MSP) detector from Hendrycks & Gimpel (2016) uses the maximum class probability prediction as a measure of model confidence. MSP does not incorporate any information about the feature geometry like VSOOD.

ASH. The activation shaping (ASH) method from Djurisic et al. (2022) prunes the majority of a sample's activations and scales the rest. These perturbed activations are then fed into the Energy score from (Liu et al., 2020). The time complexity of ASH is $O(D \log D)$. ASH, unlike VSOOD, does not require access to the training data. However, this approach can effect ID accuracy unless a second forward pass in the model is made.

SCALE. Xu et al. (2023) found that ASH does not require any pruning, and benefits from only scaling the activations. SCALE also has time complexity $O(D \log D)$, but does not affect ID accuracy. SCALE can be used on any model without accessing training data.

ADASCALE. Regmi (2025) improved on SCALE by noting that using a fixed scaling factor across all samples can be overly rigid. Instead, ADASCALE uses a gradient perturbation method to estimate an ideal, per-sample scaling factor. This approach requires multiple forward passes as well as a backward pass to compute the scaling factor. There are two variants, ADASCALE-A and ADASCALE-L, that differ in how the scaling is applied. We compare against ADASCALE-A in this paper because it provides better performance on the tested benchmarks.

IODIN. Regmi (2024) used an input perturbation approach to achieve better ID-OOD logit separation. In particular, they find the regions of an image that are most informative to a model's prediction and apply a perturbation to them before using the Energy score (Liu et al., 2020). Like ADASCALE, IODIN requires the ability to compute gradients on the model and multiple forward passes for each test sample.

GEN. Liu et al. (2023) introduced the Generalized Entropy score, a logit-only OOD detector. It measures the entropy of the class predictions over the top-K classes, with OOD samples having higher uncertainty. The entropy is given as $s_{\text{GEN}}(x) = \sum_{i \in \mathcal{T}_K(x)} p_i(x)^\gamma (1 - p_i(x))^\gamma$. We incorporate this form into our probability-weighted aggregation to reflect the model uncertainty alongside our class novelty scores.

KPCA. Kernel-PCA for OOD detection was proposed by Fang et al. (2024) to measure OOD based on the KPCA reconstruction error. The authors found that using a Gaussian kernel on the ℓ_2 normalized embeddings, and then approximating this kernel with Random Fourier Features (RFFs) provided a scalable and accurate alternative to standard PCA-reconstruction error pipelines. While KPCA and VSOOD both rely on kernels, VSOOD operates in a kernel/feature space and rely on the eigenspectrum of a similarity operator, but there are two primary differences. (1) KPCA constructs a single, global kernel matrix, whereas VSOOD constructs local, class-conditional kernel matrices (2) KPCA measures reconstruction error, while VSOOD measures the change in class diversity via eigenvalue concentration. The primary hyperparameters for KPCA are the RFF dimension, the number of principal components, and the gaussian kernel bandwidth. We use the author recommended defaults for these parameters across datasets, as reconstructing the RFFs based on the dimension or the kernel bandwidth would be too expensive to sweep over.

FDBD. The Fast Decision Boundary Detector (FDBD) introduced in Liu & Qin (2023) exploits the observation that OOD samples lie closer to decision boundaries than ID samples in the embedding space. The paper provides a hyperparameter-free, fast $O(C + D)$ algorithm for estimating these

Ablation	ResNet-50	ViT-B/16	Swin-T
Original	89.26	87.48	88.79
No probability-weighting	84.98	87.28	87.30
No global correction ($g = 0$)	89.25	87.48	86.73
Global correction ($g = 1$)	89.26	87.45	88.79
$\gamma = 1$	87.45	86.75	86.98
$\gamma = 0.5$	88.59	87.01	88.79
$\gamma = 0.1$	89.26	87.48	87.53
$K = 1000$	88.25	87.47	88.56
$K = 100$	89.26	86.72	87.94
$K = 10$	88.38	87.48	**88.94**
rank $= 5$	**89.29**	87.54	88.86
rank $= 10$	89.28	87.56	88.87
rank $= 50$	89.28	**87.59**	88.89

Table 5: Ablation study of VSOOD on ImageNet-1K OOD Average AUROC detection across ResNet-50, ViT-B/16, and Swin-T. Best performance in each column is bolded.

distances. The authors also show that FDBD is theoretically accurate under the assumptions of neural collapse.

NCI. The Neural Collapse Inspired OOD Detector (NCI) directly leverages the properties of neural collapse to identify OOD samples Liu & Qin (2025). In particular, the authors leverage 4 characteristics of Neural Collapse: (1) variability within a class goes to 0 (2) class means are arranged on an Equiangular Tight frame (3) the last layer linear classifier converges to the class means and (4) the classifier acts as a nearest center classifier. With these characteristics, NCI measures the cosine similarity between the centered feature and the class weights. NCI is also fast: it runs in $O(D)$ additional time. Like FDBD, NCI is also hyperparameter-free. VSOOD does not make assumptions about Neural Collapse.

KNN. The KNN OOD detector from Sun et al. (2022) measures the distance between a test sample and its kth nearest neighbor in a normalized feature space. This non-parametric method requires each test sample to be compared against the entire training set. At inference time, VSOOD does not search over the entire training dataset.

RMDS++. Given ℓ_2-normalized embeddings $x \in \mathbb{R}^D$ with $\|x\|_2 = 1$, RMDS++ measures OOD uncertainty by comparing how well x fits a class-conditional Gaussian model versus the global feature distribution (Ren et al., 2021; Mueller & Hein, 2025). Let μ_c, Σ denote the mean and covariance of class c, and let μ_g, Σ_g denote the mean and covariance computed over the full training set. Σ is shared across all classes. RMDS++ uses a *relative* Mahalanobis distance score as

$$s_{\text{RMDS++}}(x) = \min_{c \in [C]} \left[(x - \mu_c)^\top \Sigma^{-1} (x - \mu_c) - (x - \mu_g)^\top \Sigma_g^{-1} (x - \mu_g) \right], \tag{14}$$

where smaller values indicate that x is substantially closer to some class distribution than to the global density. Mueller & Hein (2025) had found that using normalized embeddings vastly improved detector performance across a wide variety of architectures.

C VSOOD Ablation Analysis

We perform hyperparameter ablations of VSOOD on the tested benchmarks to justify the role of each component as well as the choice of our rank-1 approximation. In Table 5, we test how VSOOD performs if we do not use the probability-weighted aggregation and instead only use the minimum class-conditional score. For other tests, we fix one hyperparameter and do the automatic selection on the validation set for the rest. The three hyperparameters are: g, the global correction term, γ, a class-probability weighting exponent used in Equation 7, and K, the number of classes to consider. Table 5 shows that the probability-weighting scheme is important for our performance, which is

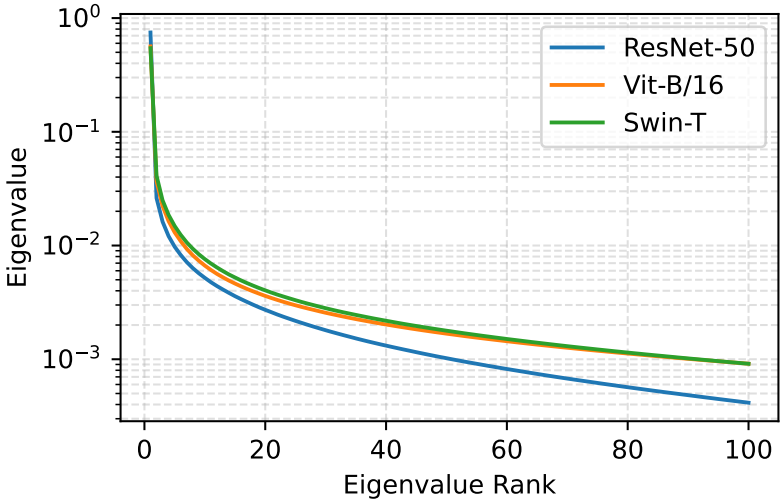

Figure 3: Average eigenspectrum of class-conditional density matrices ρ_c across models is dominated by the leading eigenvalue. The average values of the top-100 eigenvalues of ρ_c are shown (log-scale y-axis), across three ImageNet-1K models.

unsurprising given we only track the principal eigenvector and eigenvalue in each class. Leveraging information from multiple classes helps us compensate for using less class-specific information. For different choices of hyperparameters, we can achieve competitive, and sometimes even better, performance than the selected parameters.

In Table 5, we also include average AUROC for using different choices of rank from Equation 5. In general, only using a rank-1 approximation provides very similar performance to using a rank-50 approximation. In Figure 3, we observe that class density matrices exhibit a large spectral gap: the leading eigenvalue is on average an order of magnitude larger than the second. Since VS_2 depends on the squared spectrum (Equation 2), this gap implies that the rank-1 component captures most of the VS_2-relevant information, whereas smaller components contribute weakly and are more sensitive to noise.

Method	ImageNet-1K OpenOOD Benchmark (ViT-B/16)					AVG
	Near-OOD		Far-OOD			
	SSB-Hard	Ninco	iNaturalist	Textures	OpenImage-O	
Evaluation under FPR (↓)						
MSP	86.42	77.28	42.40	56.46	56.19	63.75
ASH	93.50	95.37	97.02	98.50	94.79	95.84
SCALE	92.35	94.66	86.73	84.69	89.50	89.59
ADASCALE	85.89	62.79	36.45	54.01	43.65	56.56
GEN	**82.23**	**59.33**	**22.92**	38.30	**35.47**	**47.65**
IODIN	86.41	77.33	42.47	56.45	56.18	63.77
KPCA	92.00	72.32	47.94	49.20	53.66	63.02
FDBD	87.03	60.67	33.14	45.07	39.56	53.09
NCI	86.14	61.46	34.84	48.86	41.96	54.65
KNN	89.04	63.90	45.62	**21.41**	50.97	54.19
RMDS++	84.82	45.07	13.95	35.71	27.23	**41.36**
VSOOD	**83.79**	**45.31**	**16.53**	**34.37**	**26.42**	**41.28**
Evaluation under AUROC (↑)						
MSP	68.94	78.11	88.19	85.06	84.86	81.03
ASH	53.90	52.51	50.62	48.53	55.51	52.21
SCALE	56.46	61.33	73.69	78.88	72.54	68.58
ADASCALE	67.47	80.55	89.81	86.77	87.86	82.49
GEN	**70.09**	**82.51**	**93.54**	90.23	**90.27**	**85.33**
IODIN	68.94	78.11	88.19	85.06	84.86	81.03
KPCA	51.62	70.10	80.36	82.08	77.78	72.39
FDBD	65.80	80.75	90.46	88.50	89.01	82.90
NCI	66.30	80.80	90.20	87.85	88.46	82.72
KNN	55.64	76.06	83.28	**96.28**	83.18	78.89
RMDS++	73.72	88.31	96.82	89.11	92.39	**88.07**
VSOOD	**72.99**	**86.23**	**95.58**	**90.38**	**92.23**	**87.48**

Table 6: VSOOD provides competitive performance with state-of-the-art methods on the OpenOOD ImageNet-1K benchmark with a ViT-B/16 model. RMDS++ achieves similar performance. **Orange** marks best, **blue** marks 2nd, and **purple** marks 3rd.

Method	ImageNet-1K OpenOOD Benchmark (Resnet-50)					AVG
	Near-OOD		Far-OOD			
	SSB-Hard	Ninco	iNaturalist	Textures	OpenImage-O	
Evaluation under FPR (↓)						
MSP	74.48	56.87	43.46	60.89	50.13	57.15
ASH	73.66	52.99	14.10	15.28	29.19	37.04
SCALE	67.71	51.81	9.51	11.90	28.17	33.82
ADASCALE	**57.87**	**45.76**	7.60	**10.42**	**20.59**	**28.45**
GEN	75.71	54.88	26.11	46.23	34.52	47.49
IODIN	76.14	58.78	30.06	46.40	37.46	49.75
KPCA	98.31	92.13	58.93	23.22	82.89	73.57
FDBD	81.00	54.28	23.23	28.96	31.85	43.86
NCI	75.76	53.72	14.84	18.08	32.19	38.92
KNN	89.04	63.90	45.62	21.41	50.97	54.19
RMDS++	80.70	53.20	29.85	48.42	39.88	50.41
VSOOD	74.55	54.16	**6.89**	13.48	33.63	36.54
Evaluation under AUROC (↑)						
MSP	72.09	79.95	88.41	82.43	84.86	81.55
ASH	72.89	83.45	97.06	96.90	93.26	88.71
SCALE	77.35	85.37	98.02	97.63	93.95	90.46
ADASCALE	**81.66**	**87.14**	98.31	**97.88**	**95.62**	**92.12**
GEN	72.01	81.70	92.44	87.60	89.26	84.60
IODIN	72.49	80.57	91.33	88.35	89.21	84.39
KPCA	40.03	58.13	82.36	95.70	73.65	69.97
FDBD	68.22	81.55	93.22	92.99	90.32	85.26
NCI	72.62	83.20	96.86	96.53	92.69	88.38
KNN	55.64	76.06	83.28	96.28	83.18	78.89
RMDS++	71.64	83.52	91.99	90.46	88.73	85.27
VSOOD	73.82	84.32	**98.35**	97.15	92.71	89.27

Table 7: VSOOD ranks third on the OpenOOD ImageNet-1K benchmark with a ResNet-50 model. Activation shaping methods SCALE and ADASCALE perform best. **Orange** marks best, **blue** marks 2nd, and **purple** marks 3rd.

Method	ImageNet-1K OpenOOD Benchmark (Swin-T)					AVG
	Near-OOD		Far-OOD			
	SSB-Hard	Ninco	iNaturalist	Textures	OpenImage-O	
Evaluation under FPR (↓)						
MSP	80.90	61.21	37.31	60.87	49.30	57.92
ASH	95.70	94.38	94.24	96.16	93.74	94.84
SCALE	88.94	84.20	74.76	91.30	87.91	85.42
ADASCALE	85.54	68.33	31.84	51.49	36.59	54.76
GEN	**79.19**	**48.04**	20.51	44.57	32.69	45.00
IODIN	80.88	61.24	37.35	60.93	49.30	57.94
KPCA	95.88	87.49	60.50	53.50	70.49	73.57
FDBD	86.23	59.39	27.35	42.29	35.00	50.05
NCI	87.07	57.44	27.25	43.15	37.04	50.39
KNN	87.44	59.27	31.93	**39.68**	39.55	51.57
RMDS++	84.05	51.10	**12.11**	**40.74**	**29.60**	**43.52**
VSOOD	**70.01**	**41.93**	**15.22**	42.94	**25.69**	**39.16**
Evaluation under AUROC (↑)						
MSP	71.78	81.72	89.86	83.27	85.77	82.48
ASH	45.87	47.07	46.62	41.66	45.64	45.37
SCALE	64.64	72.47	78.60	68.52	71.18	71.08
ADASCALE	65.84	79.98	90.87	86.93	89.45	82.61
GEN	**72.75**	85.18	94.23	88.15	90.56	86.17
IODIN	71.79	81.72	89.86	83.28	85.77	82.48
KPCA	41.59	59.08	74.40	82.58	66.17	64.76
FDBD	65.93	80.12	90.91	**88.35**	89.48	82.96
NCI	67.58	82.44	91.95	87.99	89.29	83.85
KNN	61.06	77.23	86.58	**88.51**	84.47	79.57
RMDS++	72.42	**86.45**	**97.08**	88.21	**91.62**	**87.16**
VSOOD	**78.71**	**87.93**	**96.20**	88.34	**92.76**	**88.79**

Table 8: VSOOD achieves state-of-the-art performance on the OpenOOD ImageNet-1K benchmark with a Swin-T model. **Orange** marks best, **blue** marks 2nd, and purple marks 3rd.