SCALING PRETRAINED REPRESENTATIONS ENABLES LABEL-FREE OUT-OF-DISTRIBUTION DETECTION WITHOUT FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Models trained with deep learning often fail to signal when inputs fall outside their training data manifold, leading to unreliable predictions under distribution shift. Prior work suggests that effective out-of-distribution (OOD) detection often requires some combination of class-conditional modeling, sophisticated detectors, and specialized models obtained through supervised fine-tuning. We ask whether, in modern pretrained models, label-free OOD detection is increasingly limited by the representation being probed rather than the sophistication of the detector. Across 59 backbone–task pairings spanning vision and language, we compare two complementary label-free detectors: a global Mahalanobis estimator fit on unlabeled latent representations, and ReSCOPED, a lightweight, diffusion-based typicality estimator operating on the same features at a local level. As pretrained representations scale, both detectors improve substantially, previously challenging benchmarks approach saturation even with simple label-free detection, and the performance gap between the two methods shrinks. This same trend persists on harder, non-saturated vision and language benchmarks. These results suggest that label-free OOD detection depends strongly on the geometry exposed by frozen pretrained backbones, reducing the importance of detector choice as backbone scale increases and enabling efficient deployment directly on frozen models.

1 INTRODUCTION

Whether utilizing a Vision Transformer (ViT) Dosovitskiy et al. (2021) for image classification or a Large Language Model (LLM) for natural language processing, current architectures generate high confidence predictions even when presented with inputs that fall outside their training manifold. Unsupervised detection strategies are particularly valuable because they rely only on unlabeled in-distribution data and therefore remain task agnostic (Liu et al., 2020; Graham et al., 2023; Heng

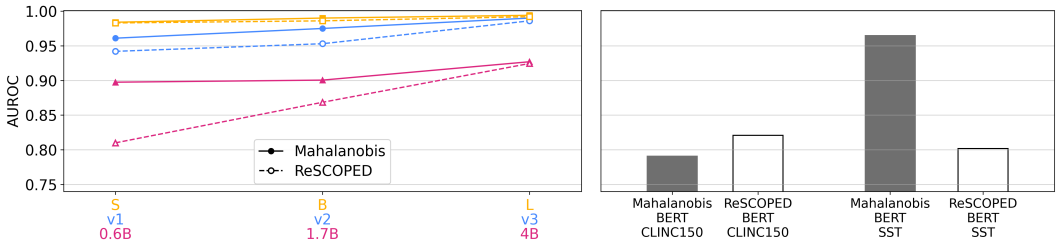


Figure 1: AUROC for Mahalanobis and ReSCOPED across DINO generation, DINOv3 model scale (S, B, L), and Qwen3 model scale (0.6, 1.7, 4.0). For DINO, AUROC is averaged across the 9 tasks in Table 3; for Qwen3, AUROC is averaged across the tasks in Table 2. Across both vision and language, scaled frozen backbones yield high-performing label-free OOD detectors while also reducing the gap between global Mahalanobis and local score-curvature detectors. In weaker frozen BERT representations, rankings are dataset-dependent: ReSCOPED is stronger on CLINC150, while Mahalanobis is stronger on SST.

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.

et al., 2024). However, traditional out-of-distribution (OOD) methods face persistent challenges: likelihood-based estimators can assign higher probabilities to OOD samples than to training data (Nalisnick et al., 2018; Choi et al., 2018), while reconstruction-based methods are sensitive to bottleneck tuning (An & Cho, 2015; Pinaya et al., 2021). In this work, we provide empirical evidence that frozen representation structure governs not only label-free OOD detection performance, but also how much detector sophistication matters. Here, representation structure refers to how readily OOD-relevant features are exposed to global and local detectors as backbone families and model scales change.

A common perspective across modalities is that strong OOD detection often requires either class labels for in-distribution (ID) data or adaptation of the learned representation. For example, Lee et al. (2018) model feature representations using class-conditional Gaussian distributions, an assumption that persists in later extensions (Liu et al., 2024; Mueller & Hein, 2025; Wei et al., 2025). Similarly, prior work has shown that near-OOD detection remains challenging when using frozen representations, and that supervised fine-tuning can substantially improve performance (Fort et al., 2021; Xu et al., 2021). At the same time, OOD detection methods have been shown to exhibit inconsistent performance across datasets, with no single detector consistently outperforming others (Tajwar et al., 2021). Together, these results suggest that OOD performance depends not only on the detector applied to a representation, but also on the scale and structure of the representation itself.

This motivates the fully unsupervised setting, where only unlabeled in-distribution data is available. While labels and fine-tuning can improve OOD detection, they limit deployment alongside frozen foundation models: class labels are often unavailable, especially in language settings where models are trained on large-scale unlabeled corpora and deployed on open-ended inputs, and fine-tuning requires maintaining task-adapted model variants. Recent work such as DiffPath has focused on making label-free OOD detection more general-purpose by using a single pretrained diffusion model across multiple tasks (Heng et al., 2024). Recent work has also observed that DINOv2 (Oquab et al., 2024) and other foundation-model representations can support competitive OOD detection with simple scores such as nearest neighbors (Zhang et al., 2024; Zhao et al., 2025). However, these observations do not establish whether this behavior generalizes across fundamentally different detectors and modalities, nor whether the inconsistencies observed in prior OOD detectors persist as representations scale. This raises a broader question: as pretrained representations improve, is label-free OOD detection still primarily limited by detector design, or by the representation being probed?

We test this question using two deliberately contrasting OOD detectors rather than attempting to identify a universally best detector. First, we use a label-free global Mahalanobis distance-based scheme that models the frozen representation distribution using only unlabeled in-distribution latents, removing the class-conditioning used in prior Mahalanobis-based methods (Lee et al., 2018; Liu et al., 2024; Mueller & Hein, 2025; Wei et al., 2025). The second is a typicality-based method, ReSCOPED, that builds upon SCOPED (Barkley et al., 2025) and evaluates local representation geometry using diffusion-based score and curvature information. In particular, diffusion-based detectors can be viewed as multi-scale filters that selectively expose structure at different granularities, allowing them to recover OOD signal when it is not globally accessible. Their substantially different biases provide a controlled test of whether representation-driven improvements persist across both a simple global model and a substantially more expressive local detector.

We validate this representation-centered view across 59 backbone-task pairings sourced from the Bonsai repository (Google, 2024) and HuggingFace Transformers (Wolf et al., 2019). As shown in Figure 1, stronger representations substantially improve label-free OOD detection across both vision and language. Previously challenging benchmarks approach saturation even with simple global Mahalanobis detection, while on harder, non-saturated benchmarks both global and local detectors improve with representation scale and their performance gap narrows. In weaker and moderately strong representations, the relative performance of global Mahalanobis and local score-curvature methods often depends on the particular datasets and backbones. As representations improve, however, OOD-relevant structure becomes increasingly accessible to both detector designs.

Because both detectors operate directly on latents from frozen models already used downstream, this representation-centered view also enables efficient deployment. For example, ReSCOPED can act as a lightweight LLM gatekeeper prior to decoding, adding only 5.4 ms of latency for Qwen3-4B (Yang et al., 2025). Beyond this primary scaling result, our ablations further show that the highest

performing score-curvature detection occurs across a broad and consistent diffusion-noise plateau, suggesting that OOD-relevant structure is exposed at stable scales across Transformer (Vaswani et al., 2017) backbones.

Key Contributions:

1. We show that scaling modern pretrained vision and language representations changes the operating regime of label-free OOD detection, enabling strong performance without fine-tuning or class labels and driving previously challenging benchmarks toward saturation even with simple global detection.
2. Using unsupervised Mahalanobis and ReSCOPED as illustrative global and local detectors, we show empirically that detector differences are larger and dataset-dependent for weaker representations but shrink as representations scale, including on harder non-saturated vision and language benchmarks.
3. We identify a consistent denoising-scale plateau for ReSCOPED across vision and language Transformers, suggesting OOD-relevant structure becomes accessible to the diffusion-based detector at a shared scale.
4. We demonstrate a practical deployment pathway via low-latency OOD detection on frozen models, including real-time LLM prefill-stage gatekeeping.

2 Background and Related Work

2.1 Representation Geometry and OOD Detection

Out-of-distribution (OOD) detection can be viewed as probing the geometry of a model's learned representation. Let $z = \phi(x) \in \mathbb{R}^d$ denote the representation of an input x under a frozen encoder ϕ. Different OOD methods vary in what information they require and what geometry they measure: distance-based methods estimate global or class-conditional structure (Lee et al., 2018; Mueller & Hein, 2025; Wei et al., 2025), score-based typicality methods (Heng et al., 2024; Barkley et al., 2025) capture local quantities such as gradients and curvature, and supervised or fine-tuned approaches (Xu et al., 2021; Fort et al., 2021) modify ϕ itself to expose OOD-relevant structure.

The class-conditional formulation of Lee et al. (2018) considers a labeled in-distribution dataset $\{(x_i, y_i)\}_{i=1}^N$, where x_i is an input and $y_i \in \{1, \dots, C\}$ is its class label. Let N_c denote the number of samples in class c, and let $\mathcal{I}_c = \{i : y_i = c\}$ denote the corresponding class index set. The method estimates per-class means $\mu_c = \frac{1}{N_c} \sum_{i \in \mathcal{I}_c} z_i$ and a shared covariance $\Sigma_{\text{class}} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i \in \mathcal{I}_c} (z_i - \mu_c)(z_i - \mu_c)^\top$. The OOD score is then defined as the minimum Mahalanobis distance to the class means: $S_{\text{class}}(z) = \min_c (z - \mu_c)^\top \Sigma_{\text{class}}^{-1} (z - \mu_c)$.

Later extensions continue to leverage class-conditional structure in feature space (Mueller & Hein, 2025; Wei et al., 2025). Prior work has also shown that OOD detector rankings can be inconsistent across datasets, with no single method uniformly outperforming others (Tajwar et al., 2021). Recent work has further shown that the performance of OOD detectors such as Mahalanobis distance is strongly influenced by representation geometry, including spectral properties and intrinsic dimensionality (Janiak et al., 2026). Our work connects these observations by showing that representation geometry shapes detector choice: when OOD-relevant structure is weak or only partially organized, the best-performing method can depend on the dataset, while scaled pretrained backbones can make the same structure accessible to both local and global methods.

In contrast to class-conditional methods, we focus on the fully label-free setting, where detectors are fit using only unlabeled in-distribution data and do not have access to additional class labels. This setting often arises when deploying frozen models on unlabeled task data or user prompts. In this setting, we compare two complementary probes of the same frozen representation: a global covariance probe and a local score-curvature probe.

2.2 Score-Based Typicality and Local Geometry

Models can assign higher likelihood to OOD samples than to in-distribution (ID) data (Nalisnick et al., 2018; Choi et al., 2018) because probability mass concentrates not at the mode, but in a thin shell known as the typical set (Cover & Thomas, 2006; Nalisnick et al., 2019). Consequently, high-likelihood samples may still be atypical under the true data distribution, motivating alternatives to direct density thresholding (Ren et al., 2019), including supervised fine-tuning or outlier-exposure methods (Hendrycks & Gimpel, 2016; Lee et al., 2018; Wei et al., 2025) and unsupervised latent-space methods based on frozen representations (Xu et al., 2021; Gulati et al., 2024).

Score-based diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2020) learn a time-dependent score function $s_\theta(x_t, t) \approx \nabla_x \log p_t(x_t)$ under progressive noise perturbations. The model is trained to reverse a Gaussian noising process, and p_t is the distribution of data at stage t of the diffusion process; the learned score field can be used both for generation and to estimate local geometric quantities such as curvature.

One way to incorporate local geometry is through the score-curvature ratio introduced by Barkley et al. (2025), motivated by the Fisher information identity. Let $p(x)$ denote the in-distribution density and $s(x) = \nabla_x \log p(x)$ its score function. Then $\mathbb{E}_{x\sim p}[\|s(x)\|^2] = \mathbb{E}_{x\sim p}[-\operatorname{Tr}(\nabla_x s(x))]$, which motivates the typicality statistic $T(x) = \|s(x)\|^2/\kappa(x)$, where $\kappa(x) = -\operatorname{Tr}(\nabla_x s(x))$. For d-dimensional standard Gaussian p, typical samples $x \sim p$ concentrate near $\|x\| = \sqrt{d}$ and satisfy $T(x) \approx 1$. For non-Gaussian data, $T(x)$ need not equal 1 for all points of the typical set; however, empirically, diffusion score fields associated with non-Gaussian data produce score-curvature statistics that have a consistent value for in-distribution samples and deviate for OOD samples, providing an unsupervised typicality signal.

3 Global and Local Probes of Representation Geometry

We compare two label-free probes of the same frozen representation distribution: global Mahalanobis distance fit to unlabeled in-distribution representations, and ReSCOPED, a local score-curvature detector introduced in this work. Our goal is not to identify a universally best detector, but to test whether scaling pretrained representations drives strong OOD detection performance across methods with substantially different complexity and inductive biases. Together, they test whether OOD-relevant structure is accessible through both global covariance geometry and local typicality estimates.

3.1 Global Mahalanobis Probe

Assume access to unlabeled in-distribution inputs $\{x_i\}_{i=1}^N$ and a frozen encoder ϕ, and let $z_i = \phi(x_i)$. Without labels, class means cannot be estimated. We therefore fit a global Gaussian model to the representation distribution, with mean $\mu = \frac{1}{N}\sum_{i=1}^N z_i$ and covariance $\Sigma_{\text{global}} = \frac{1}{N}\sum_{i=1}^N (z_i - \mu)(z_i - \mu)^\top + \lambda I$, where $\lambda > 0$ is a regularization parameter. The OOD score is $S_{\text{global}}(z) = (z-\mu)^\top \Sigma_{\text{global}}^{-1}(z-\mu)$. This provides a deliberately simple, label-free global probe of the structure exposed by frozen representations.

3.2 Local Score-Curvature Probe: ReSCOPED

We use ReSCOPED, a representation-based variant of SCOPED Barkley et al. (2025), as a substantially more expressive local counterpart to the global Mahalanobis method. Rather than modeling the full representation distribution with a single covariance, ReSCOPED trains a lightweight diffusion observer on frozen representations and evaluates local score-curvature typicality. This contrast allows us to test whether representation-driven improvements persist when OOD structure is measured using fundamentally different global and local information.

Formally, given an input x, we extract a representation $z = \phi(x)$ from a frozen encoder ϕ, and train an Elucidated Diffusion Model (EDM) $s_\theta(z, \sigma)$ (Karras et al., 2022) on these representations. At test

time, we compute the SCOPED typicality statistic (Barkley et al., 2025):

$$T(z) = \frac{\|s_\theta(z,\sigma)\|^2}{\kappa(z)}, \quad \kappa(z) = -\operatorname{Tr}(\nabla_z s_\theta(z,\sigma)),$$

where the trace is approximated using Hutchinson's estimator (Hutchinson, 1989).

We obtain anomaly scores by fitting a Kernel Density Estimator (KDE) to in-distribution values of $T(z)$ and evaluating the negative log-likelihood at test time. A noise level ($\sigma \approx 0.09$) is selected via a denoising sweep on a validation task (SVHN vs. CelebA) and used across all experiments, avoiding per-dataset tuning required by prior diffusion-based approaches (Heng et al., 2024; Barkley et al., 2025; Ding et al., 2025). This value lies within a broad plateau of highest performing noise levels for OOD detection observed across models and modalities, suggesting that OOD-relevant structure in pretrained Transformer (Vaswani et al., 2017) representations becomes accessible within a stable denoising range. We analyze this phenomenon further in Section 5.2.

By combining a frozen pretrained encoder (e.g., ViT, DINOv3, or Qwen3) with a lightweight diffusion-based observer (approximately 8M parameters), ReSCOPED avoids full-model fine-tuning, which is commonly required in prior OOD detection approaches. Training completes in two minutes on a single GPU, and inference requires one encoder forward pass, one observer forward pass, and a single Jacobian-vector product, corresponding to approximately 5.4 ms of additional latency for a 2560-dimensional latent from Qwen3-4B (Yang et al., 2025) on an NVIDIA 4090.

Together, these two methods provide complementary views of the same representation distribution: Mahalanobis measures global covariance structure, while ReSCOPED measures local score-curvature typicality. Figure 2 provides a qualitative preview of this contrast. As the backbone improves, ReSCOPED scores become sharper with reduced ID/OOD overlap, while Mahalanobis separation improves because OOD samples move farther outward in global radius even though the ID distribution broadens. Thus, stronger representations do not simply improve one particular detector or make the data more Gaussian; they make OOD-relevant deviations more visible to both local and global methods.

4 Vision and Language Processing Results

We now test whether scaling pretrained representations drives label-free OOD performance across detector constructions, first in vision and then in language.

Vision benchmarks. We evaluate OOD detection on CIFAR-10 (C10) (Krizhevsky, 2009), SVHN (Netzer et al., 2011), CelebA (Liu et al., 2015), and CIFAR-100 (C100). For harder near-OOD tasks, we additionally evaluate ImageNet-200 as ID against SSB-hard and NINCO as OOD, following the OpenOOD v1.5 benchmark protocol (Zhang et al., 2024). Since the small-scale datasets have differing resolutions, we follow Heng et al. (2024) and apply a resizing procedure that avoids artifacts that could trivially simplify OOD detection. Specifically, all images are first downsampled to the lowest native resolution among the datasets (32×32 for CIFAR and SVHN), and then upsampled to the model input resolution.

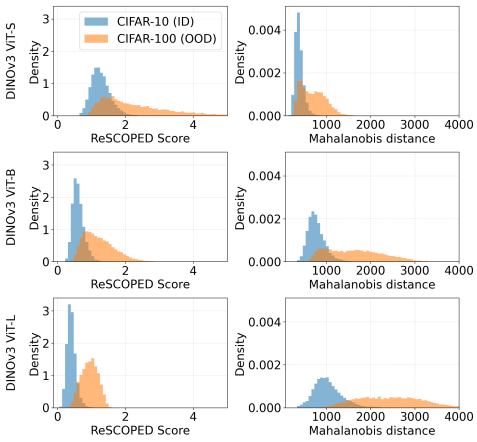


Figure 2: ID/OOD score distributions for CIFAR-10 (ID) vs. CIFAR-100 (OOD) across DINOv3 scale (rows: ViT-S/B/L; columns: ReSCOPED and Mahalanobis scores). Both methods improve with model scale.

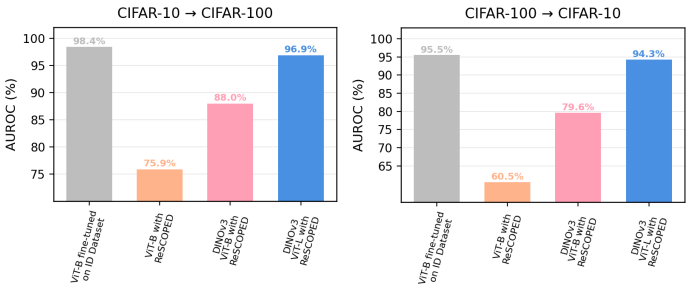
We compare against generative baselines including SCOPED (Barkley et al., 2025), DiffPath (Heng et al., 2024), and Improved CD (Du et al., 2021), as well as likelihood-based methods including Density of States Estimation (DoSE) (Morningstar et al., 2021), the typicality test of Nalisnick et al. (2019), WAIC (Choi et al., 2018), and the likelihood-ratio score of Ren et al. (2019). We also include diffusion-based baselines such as NLL (negative ELBO), DDPM-OOD (Graham et al., 2023), LMD (Liu et al., 2023),

and MSMA (Mahmood et al., 2020). Following prior work (Barkley et al., 2025), we report AUROC and computational cost in terms of the number of forward passes and Jacobian-vector products (JVPs), which together correspond to the total number of function evaluations for diffusion-based methods.

Language benchmarks. For language tasks, we evaluate on CLINC150 (Larson et al., 2019) and a cross-corpus SST benchmark following Xu et al. (2021). CLINC150 evaluates cross-intent OOD detection, while SST evaluates cross-corpus OOD detection using out-of-domain corpora. Following Xu et al. (2021), we report AUROC, detection accuracy (DTACC), and area under the precision-recall curve for both in-domain (AUIN) and out-of-domain (AUOUT) samples. All baselines use a frozen BERT (Devlin et al., 2019) backbone and differ in how in-distribution representations are used for OOD detection. This includes methods from Xu et al. (2021) which include frozen feature-based methods like (class-conditional) Euclidean-based and Mahalanobis-based distance functions (EDF and MDF, respectively) and approaches that adapt representations using in-distribution data without access to task-specific OOD or class labels like Binary Classification with Auxiliary Dataset (BCAD) and In-Domain Masked Language Modeling (IMLM). We also evaluate EC-NNK-Means (Gulati et al., 2024) under the frozen BERT setting to ensure consistency with prior label-blind evaluations.

4.1 Near-OOD Detection Without Labels or Fine-Tuning

We first isolate a concrete question raised by prior near-OOD work: whether strong near-OOD detection requires supervised adaptation of the representation. Fort et al. (2021) argue that full-parameter supervised fine-tuning of large Vision Transformers is important for strong near-OOD performance. Their evaluation uses class-conditional Mahalanobis detectors, which rely on labels to define the OOD score (Lee et al., 2018; Mueller & Hein, 2025; Wei et al., 2025). We test whether improvements in modern pretrained representations can instead provide this OOD-relevant structure directly in the fully label-free setting, where the encoder is frozen and only unlabeled in-distribution representations are available.



(a) CIFAR-10 to CIFAR-100 (b) CIFAR-100 to CIFAR-10

Figure 3: Comparison of ReSCOPED AUROC performance using frozen encoders of varying representational scale and structure against fine-tuning methods across near-OOD datasets.

As shown in Figure 3, representations from modern frozen backbones achieve competitive near-OOD performance when paired with ReSCOPED. For example, on CIFAR-100 (ID) versus CIFAR-10 (near-OOD), ReSCOPED achieves 94.3% AUROC using a frozen DINOv3-L backbone (Siméoni et al., 2025), closely matching fine-tuned models (Fort et al., 2021). This suggests that improvements in the frozen representation itself can recover OOD-relevant structure that prior work obtained through supervision or adaptation. We next ask whether this structure is specific to ReSCOPED on vision tasks, or whether it also appears for simpler global detectors and in other modalities.

4.2 Main Vision Results

Our vision experiments ask how detector behavior changes as modern frozen representations improve. Here, we study three empirical questions. First, how well can simple methods perform for OOD detection with modern representations, compared to methods using class labels or fine-tuning? Second, at smaller model scales, is relative detector performance dataset-dependent? Third, as modern pretrained representations scale, does the gap between simple global detectors (Mahalanobis) and more sophisticated local detectors (ReSCOPED) shrink? If so, this would suggest that representation geometry, rather than detector complexity alone, drives OOD performance in modern frozen backbones.

Table 1: Performance comparison across vision in- vs out-of-distribution detection tasks as measured by AUROC score. Higher is better. **Bold** is best and underline is second best. The first set of baselines and the reporting format are reproduced from Barkley et al. (2025). We additionally report OpenOOD's ImageNet-200 ID results against SSB-hard and NINCO OOD datasets for the DINOv3 methods. As in prior work, we report computational cost as #F + #J, where F denotes a forward pass and J a Jacobian–vector product.

Method	C10 vs			CelebA vs			Avg	ImageNet-200 vs			#F + #J
	SVHN	CelebA	C100	C10	SVHN	C100		SSB-hard	NINCO	Avg	
Improved CD (Du et al., 2021)	0.910	–	0.830	–	–	–	–	–	–	–	–
DoSE (Morningstar et al., 2021)	0.955	0.995	0.571	0.949	0.997	0.956	0.904	–	–	–	–
WAIC (Choi et al., 2018)	0.143	0.928	0.532	0.507	0.139	0.535	0.464	–	–	–	–
TT (Nalisnick et al., 2019)	0.870	0.848	0.548	0.634	0.982	0.671	0.759	–	–	–	–
LR (Ren et al., 2019)	0.064	0.914	0.520	0.323	0.028	0.357	0.368	–	–	–	–
Diffusion-based											
NLL	0.091	0.574	0.521	0.814	0.105	0.786	0.482	–	–	–	1000F + 0J
MSMA (Mahmood et al., 2020)	0.957	**1.000**	0.615	0.910	0.996	0.927	0.901	–	–	–	10F + 0J
DDPM-OOD (Graham et al., 2023)	0.390	0.659	0.536	0.795	0.636	0.778	0.632	–	–	–	350F + 0J
LMD (Liu et al., 2023)	0.992	0.557	0.604	0.989	**1.000**	0.979	0.854	–	–	–	10^4F + 0J
Curvature and Diffusion-Based											
DiffPath (Heng et al., 2024)	0.910	0.897	0.590	0.998	**1.000**	**0.998**	0.899	–	–	–	10F + 0J
SCOPED (Barkley et al., 2025)	0.814	0.940	0.477	0.925	0.994	0.962	0.852	–	–	–	2F + 2J
Label-Free Mahalanobis DINOv3 ViT-S	**0.997**	**1.000**	0.864	**1.000**	**1.000**	**0.998**	0.977	0.714	0.700	0.707	**1F + 0J**
Label-Free Mahalanobis DINOv3 ViT-B	0.990	**1.000**	0.923	**1.000**	**1.000**	0.997	0.985	0.795	0.800	0.797	**1F + 0J**
Label-Free Mahalanobis DINOv3 ViT-L	0.983	**1.000**	**0.970**	**1.000**	0.999	0.997	**0.991**	**0.865**	0.908	**0.887**	**1F + 0J**
ReSCOPED DINOv3 ViT-S	0.994	0.999	0.864	0.999	**1.000**	0.993	0.975	0.617	0.612	0.614	2F + 1J
ReSCOPED DINOv3 ViT-B	0.968	0.996	0.919	**1.000**	**1.000**	0.996	0.980	0.829	0.838	0.834	2F + 1J
ReSCOPED DINOv3 ViT-L	0.969	0.997	0.969	**1.000**	0.997	0.995	0.988	0.855	**0.910**	0.882	2F + 1J

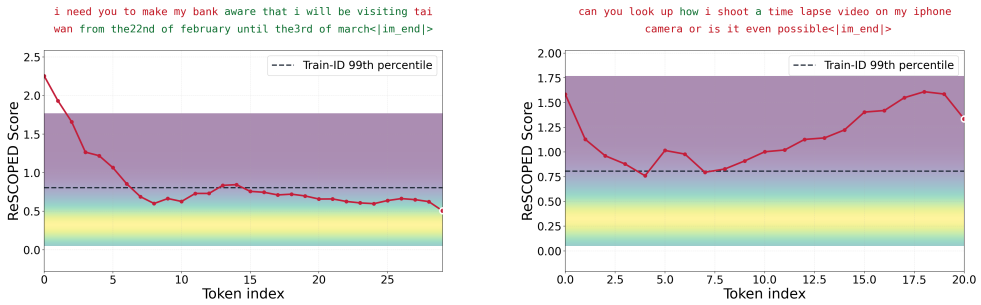
The first question is answered in Table 1: modern frozen DINO representations enable high-performing label-free OOD detection across the small-scale vision benchmarks. ReSCOPED achieves strong AUROC while requiring far fewer function evaluations than prior diffusion-based methods. However, the DINOv3 results also answer the third question: this performance is not unique to ReSCOPED, since a simple label-free Mahalanobis detector, fit only on unlabeled representations, matches or slightly exceeds ReSCOPED across DINOv3 scales (S, B, L). Notably, several benchmarks that previously separated more specialized OOD methods approach saturation with these modern frozen representations. This suggests that in this regime, performance is governed less by detector complexity and more by the geometry of the frozen representation, which makes OOD-relevant structure accessible to both global and local detectors.

To test whether this convergence is merely an artifact of saturated benchmarks, Table 1 also includes harder OpenOOD's (Zhang et al., 2024) ImageNet-200 near-OOD evaluations against SSB-hard and NINCO. These settings are substantially less saturated: DINOv3-L label-free Mahalanobis obtains 0.865 AUROC on SSB-hard and 0.908 on NINCO, compared with near-perfect AUROC on many smaller reproduced benchmarks. Nevertheless, the same scale-dependent pattern persists. On ImageNet-200, the average AUROC gap between Mahalanobis and ReSCOPED decreases from 0.093 for DINOv3-S and 0.037 for DINOv3-B to only 0.005 for DINOv3-L. Thus, the harder OpenOOD results support detector convergence even when absolute detection remains nontrivial.

Finally, these results address the second question by connecting detector variability to scale. Prior work shows that detector rankings can depend strongly on the structure of the ID and OOD datasets, with no method consistently dominating across dataset pairs (Tajwar et al., 2021). For small models, our results match this prediction of noisy rankings. At larger scales, however, the performance converges, suggesting these models learn representations that even simple detectors can exploit. This convergence foreshadows the empirical regularities discussed in Section 5.2, where we show that Transformer representations expose OOD-relevant structure across a stable range of denoising scales.

4.3 Language Results

Language provides a cross-modal test of our representation-centered hypothesis: if the regime-dependent behavior observed in our vision results reflects representation geometry rather than vision-specific structure, then weaker frozen language representations should also show dataset-dependent detector rankings, while modern scaled language backbones should reduce the gap between detector types. For both Mahalanobis and ReSCOPED detectors, we use a standard sequence mean pooling (SMP) strategy, averaging token-level hidden states into a single utterance representation before



(a) In-distribution utterance

(b) Out-of-distribution utterance

Figure 4: ReSCOPED prefix-level OOD scores from a frozen Qwen3-4B model on CLINC150. At each token, ReSCOPED scores the mean hidden representation of the prefix ending at that token. Token color denotes anomaly score (green lower, red higher), and the dotted line marks the 99th percentile of the in-distribution ReSCOPED scores. The in-distribution utterance falls below the threshold as context accumulates, while the OOD utterance remains above it across most of the query.

applying each detector. As summarized in Table 2, the BERT results support the first prediction. On CLINC150, ReSCOPED improves over the label-free Mahalanobis detector using the same SMP representation, while on the SST benchmark, label-free Mahalanobis performs better. Thus, in weaker frozen language representations, no single detector is consistently better across datasets.

The Qwen3 results provide a cross-modal confirmation of the scaling behavior observed in vision and show that scaled language backbones reduce the performance gap between global covariance and local score-curvature detectors. On CLINC150, label-free Mahalanobis and ReSCOPED are closely matched across Qwen3 scales. On SST, ReSCOPED initially lags behind Mahalanobis at smaller Qwen3 scales, but closely matches it once Qwen3-4B is used. Thus, scaled language representations make OOD-relevant structure accessible to both detector types. Together, the language results mirror the behavior observed in Table 1: detector choice matters more when the representation is weaker or less structured, but becomes less important as modern pretrained representations scale. With strong frozen language backbones, both label-free detectors achieve high-performing OOD detection that rivals or outperforms methods using fine-tuning and class labels.

5 Downstream Applications and Implications

5.1 LLM Gatekeeping and Prefill Efficiency

Representation-level OOD detection fits naturally into modern LLM inference pipelines as a gatekeeper. This deployment view is related to selective generation and abstention, where distribution-shift or reliability signals are used to decide whether a language model should proceed with generation or not (Ren et al., 2022; 2023; Azaria & Mitchell, 2023). We show that an OOD detector (operating on latents, not directly on tokens) can be used during the prefill stage to decide whether to reject generation and/or route to a fallback system. Since the key-value cache is constructed from intermediate representations of the prompt, these latents can be probed without additional transformer forward passes. In Figure 4, we test this idea with ReSCOPED, but the same prefill-stage gate can use other representation-level methods; for example, our Qwen3 results show that global Mahalanobis performs comparably in modern language representations.

At inference time, ReSCOPED adds only a single forward pass through the lightweight $\sim$8M parameter diffusion model and one Jacobian-vector product, taking approximately 5.4 ms to complete for the Qwen3-4.0B penultimate representation (2560 dimensions) on a single NVIDIA 4090 GPU. With JAX-based compilation (Bradbury et al., 2018), this enables immediate rejection decisions prior to autoregressive decoding, providing a lightweight gating signal at the prefill stage. In principle, multiple OOD methods could also be applied with minimal additional cost.

Figure 4 illustrates this behavior on CLINC150 using a frozen Qwen3-4B backbone. ReSCOPED is trained on full-utterance SMP representations, but at inference time we evaluate prefix-level SMP

representations obtained by averaging hidden states from the beginning of the prompt through each token; the full procedure is given in Appendix C. Early prefix scores can be unstable because they are computed from only a few tokens, but become more reliable as the prefix representation approaches the full-utterance representation used during training. Thus, ReSCOPED can provide repeated prefill-stage estimates of whether an accumulating query appears in- or out-of-distribution, suggesting a path toward decoding-stage OOD monitoring during autoregressive generation.

5.2 Empirical Regularities Across Transformer Backbones

Beyond comparing detector performance, ReSCOPED probes the scale at which OOD-relevant structure is exposed in frozen representations. As shown in Figure 5, sweeping the diffusion noise level across 59 configurations reveals a consistent performance plateau across both vision and language, with a broad highest performing region centered near $t = 5$. Despite differing modalities and training objectives, OOD-relevant structure becomes accessible over a common intermediate range.

This plateau suggests that OOD separation is not tied to a task-specific noise level, but to a common geometric organization of pretrained representations. At low noise levels, ReSCOPED remains sensitive to fine-scale variation, while at high noise levels, the diffusion process removes too much structure. The intermediate regime suppresses unimportant representation variation while preserving structure useful for OOD detection.

These results provide empirical evidence for a shared granularity regime in which OOD-relevant structure in pretrained Transformer representations becomes accessible. While we do not claim universality beyond the models and datasets studied, the consistency of the plateau suggests that modern representations expose OOD-relevant information at similar levels of abstraction. Understanding the theoretical basis for this regularity remains an important direction for future work.

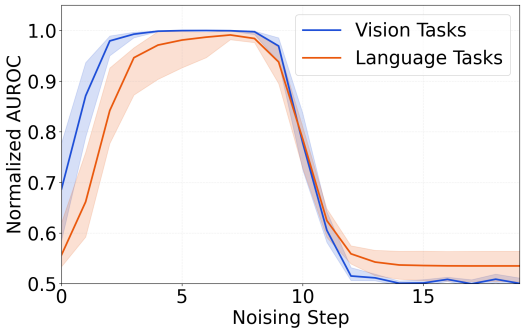


Figure 5: OOD timestep convergence across modalities. Interquartile mean (IQM) normalized AUROC and 95% bootstrapped confidence intervals across 51 vision backbone-task pairings and 8 language backbone-task pairings. The aggregate curves show a broad region of highest performance.

6 Conclusion

We study label-free OOD detection in frozen pretrained representations across vision and language, comparing a simple global Mahalanobis method with ReSCOPED, a lightweight local score-curvature typicality detector. Our results show that scaling pretrained representations drives strong label-free OOD performance while reducing the importance of detector sophistication: previously challenging benchmarks approach saturation even with simple global detection, while the performance gap between Mahalanobis and ReSCOPED also contracts on harder non-saturated vision and language tasks. Across weaker representations, detector rankings remain dataset-dependent, while scaled pretrained backbones make OOD-relevant structure increasingly accessible to both local and global methods. Together with the consistent denoising-scale performance plateau and our lightweight LLM gatekeeping example, these results suggest that modern pretrained representations increasingly expose OOD-relevant geometry in increasingly accessible ways across models and modalities. Thus, effective OOD detection in scaled foundation models may depend less on detector choice and more on understanding when frozen representations already expose OOD-relevant geometry to methods of different construction.

REFERENCES

Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. 2015. URL https://api.semanticscholar.org/CorpusID:36663713.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Brett Barkley, Preston Culbertson, and David Fridovich-Keil. Scoped: Score-curvature out-of-distribution proximity evaluator for diffusion, 2025. URL https://arxiv.org/abs/2510.01456.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2 edition, 2006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Yifan Ding, Arturas Aleksandraus, Amirhossein Ahmadian, Jonas Unger, Fredrik Lindsten, and Gabriel Eilertsen. Revisiting likelihood-based out-of-distribution detection by modeling representations. In *Scandinavian Conference on Image Analysis*, pp. 166–179. Springer, 2025.

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, 2021. URL https://arxiv.org/abs/2010.11929.

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In *International Conference on Machine Learning*, pp. 2837–2848. PMLR, 2021.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in neural information processing systems*, 34:7068–7081, 2021.

Google. Bonsai: Minimal, lightweight JAX implementations of popular models. https://github.com/jax-ml/bonsai, 2024.

Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2948–2957, 2023.

Aryan Gulati, Xingjian Dong, Carlos Hurtado, Sarath Shekkizhar, Swabha Swayamdipta, and Antonio Ortega. Out-of-distribution detection through soft clustering with non-negative kernel regression. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12943–12959, 2024.

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

Alvin Heng, Harold Soh, et al. Out-of-distribution detection with a single unconditional diffusion model. *Advances in Neural Information Processing Systems*, 37:43952–43974, 2024.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics–Simulation and Computation*, 18(3):1059–1076, 1989.

Denis Janiak, Jakub Binkowski, and Tomasz Kajdanowicz. A geometry-based view of mahalanobis ood detection, 2026. URL https://arxiv.org/abs/2510.15202.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1311–1316, 2019.

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.

Bo Liu, Liming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. How good are llms at out-of-distribution detection?, 2024. URL https://arxiv.org/abs/2308.10261.

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.

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *International Conference on Machine Learning*, pp. 22528–22538. PMLR, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Ahsan Mahmood, Junier Oliva, and Martin Styner. Multiscale score matching for out-of-distribution detection. *arXiv preprint arXiv:2010.13132*, 2020.

Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alexander A Alemi, and Joshua V Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pp. 3232–3240. PMLR, 2021.

Maximilian Mueller and Matthias Hein. Mahalanobis++: Improving ood detection via feature normalization, 2025. URL https://arxiv.org/abs/2505.18032.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019.

Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011. URL https://api.semanticscholar.org/CorpusID:16852518.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sébastien Ourselin, and M Jorge Cardoso. Unsupervised brain anomaly detection and segmentation with transformers. *arXiv preprint arXiv:2102.11650*, 2021.

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. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*, 2022.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models, 2023. URL https://arxiv.org/abs/2312.09300.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Fahim Tajwar, Ananya Kumar, Sang Michael Xie, and Percy Liang. No true state-of-the-art? ood detection methods are inconsistent across datasets, 2021. URL https://arxiv.org/abs/2109.05554.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tong Wei, Bo-Lin Wang, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. X-mahalanobis: Transformer feature mixing for reliable OOD detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ewyR20zwqA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. Unsupervised out-of-domain detection via pre-trained transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1052–1061, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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. *Journal of Data-centric Machine Learning Research*, 2024.

Shizhen Zhao, Jiahui Liu, Xin Wen, Haoru Tan, and Xiaojuan Qi. Equipping vision foundation model with mixture of experts for out-of-distribution detection, 2025. URL https://arxiv.org/abs/2510.10584.

A AI USE STATEMENT

In this work, we used generative AI tools for help in implementing methods, creating and editing software code, creating and modifying scientific figures, identifying research gaps, to aid in and polish writing, and for providing feedback for the refinement of hypotheses and research methodology and experiments. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, assist with translation, clean or reformat datasets, support qualitative and thematic data analysis, or to interpret results. We have reviewed all AI-assisted work. We manually checked the implemented methods, code, and writing suggestions/polish, sanity checked results against prior reporting across the literature, checked our figures for correctness, and ensured the refined hypotheses, methodology, and experiments are accurate and reasonable. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

B RESCOPED SCHEMATIC

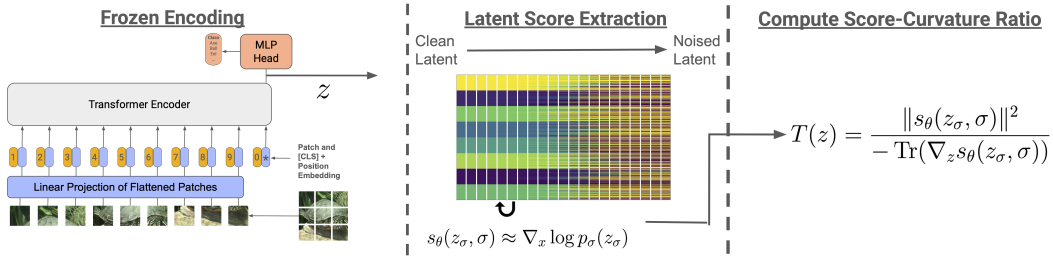


Figure 6: The ReSCOPED framework. A frozen transformer maps inputs to a latent representation z, where a diffusion-based score function s_θ estimates the local log-density gradient. The score-curvature ratio (Barkley et al., 2025) $T(z)$ is then computed to determine the typicality of the sample and whether it is in- or out-of-distribution.

C SEQUENCE MEAN POOLING (SMP) AND PREFIX-LEVEL SMP

Sequence mean pooling. Let a tokenized input sequence have hidden states $H = [h_1, \ldots, h_L] \in \mathbb{R}^{L\times d}$ from a frozen language backbone, with attention mask $m \in \{0,1\}^L$ indicating non-padding tokens. Sequence mean pooling computes a single utterance representation

$$z_{\mathrm{SMP}}(x) = \frac{\sum_{i=1}^{L} m_i h_i}{\sum_{i=1}^{L} m_i} \in \mathbb{R}^d.$$

This representation is then used as the input to both label-free Mahalanobis and ReSCOPED.

Prefix-level sequence mean pooling. For prefix-level scoring, let $H_{1:k} = [h_1, \ldots, h_k]$ denote the hidden states available up to prefix length k. We compute

$$z^{(k)}_{\mathrm{SMP}}(x) = \frac{\sum_{i=1}^{k} m_i h_i}{\sum_{i=1}^{k} m_i}, \qquad k = 1, \ldots, L.$$

The resulting prefix representations can be scored by the same OOD detector used for full-sequence representations.

D VISION AUROC COMPARISONS FOR ALL BACKBONES, SIZES, AND DATASETS FROM HENG ET AL. (2024)

Table 3: Performance comparison across vision in- vs out-of-distribution detection tasks as measured by AUROC score. Higher is better. **Bold** is best and underline is second best. This set of baselines and the reporting format is reproduced from Barkley et al. (2025). As in prior work, we report computational cost as #F + #J, where F denotes a forward pass and J a Jacobian–vector product.

Method	C10 vs			SVHN vs			CelebA vs			Avg	#F + #J
	SVHN	CelebA	C100	C10	CelebA	C100	C10	SVHN	C100		
Improved CD (Du et al., 2021)	0.910	–	0.830	–	–	–	–	–	–	–	–
DoSE (Morningstar et al., 2021)	0.955	0.995	0.571	0.962	**1.000**	0.965	0.949	0.997	0.956	0.928	–
WAIC (Choi et al., 2018)	0.143	0.928	0.532	0.802	0.991	0.831	0.507	0.139	0.535	0.601	–
TT (Nalisnick et al., 2019)	0.870	0.848	0.548	0.970	**1.000**	0.965	0.634	0.982	0.671	0.832	–
LR (Ren et al., 2019)	0.064	0.914	0.520	0.819	0.912	0.779	0.323	0.028	0.357	0.524	–
Diffusion-based											
NLL	0.091	0.574	0.521	0.990	0.999	0.992	0.814	0.105	0.786	0.652	1000F + 0J
MSMA (Mahmood et al., 2020)	0.957	**1.000**	0.615	0.976	0.995	0.980	0.910	0.996	0.927	0.928	10F + 0J
DDPM-OOD (Graham et al., 2023)	0.390	0.659	0.536	0.951	0.986	0.945	0.795	0.636	0.778	0.742	350F + 0J
LMD (Liu et al., 2023)	0.992	0.557	0.604	0.919	0.890	0.881	0.989	**1.000**	0.979	0.868	10^4F + 0J
Curvature and Diffusion-Based											
DiffPath (Heng et al., 2024)	0.910	0.897	0.590	0.939	0.979	0.953	0.998	**1.000**	**0.998**	0.918	10F + 0J
SCOPED (Barkley et al., 2025)	0.814	0.940	0.477	0.971	0.996	0.959	0.925	0.994	0.962	0.892	2F + 2J
Unsupervised Mahalanobis DINOv1 ViT-B	0.761	0.998	0.895	1.000	1.000	1.000	1.000	1.000	0.997	0.961	**1F + 0J**
Unsupervised Mahalanobis DINOv2 ViT-B	0.829	1.000	0.952	1.000	1.000	1.000	1.000	0.999	0.997	0.975	**1F + 0J**
Unsupervised Mahalanobis DINOv3 ViT-B	0.990	1.000	0.923	1.000	1.000	1.000	1.000	1.000	0.997	0.990	**1F + 0J**
ReSCOPED DINOv1 ViT-B	0.683	0.985	0.819	1.000	1.000	1.000	1.000	0.999	0.996	0.942	2F + 1J
ReSCOPED DINOv2 ViT-B	0.698	0.991	0.894	1.000	1.000	0.999	1.000	0.995	0.996	0.953	2F + 1J
ReSCOPED DINOv3 ViT-B	0.968	0.996	0.919	1.000	1.000	1.000	1.000	1.000	0.996	0.986	2F + 1J
Unsupervised Mahalanobis ViT-B	0.942	0.998	0.869	0.999	1.000	0.998	1.000	0.998	0.996	0.989	**1F + 0J**
Label-Free Mahalanobis Dinov3 ViT-S	**0.997**	1.000	0.864	1.000	1.000	1.000	1.000	1.000	0.998	0.984	**1F + 0J**
Label-Free Mahalanobis Dinov3 ViT-B	0.990	1.000	0.923	1.000	1.000	1.000	1.000	1.000	0.997	0.990	**1F + 0J**
Label-Free Mahalanobis Dinov3 ViT-L	0.983	1.000	**0.970**	1.000	1.000	1.000	1.000	0.999	0.997	**0.994**	**1F + 0J**
ReSCOPED ViT-B	0.885	0.987	0.759	0.998	0.999	0.997	0.999	0.999	0.995	0.958	2F + 1J
ReSCOPED DINOv3 ViT-S	0.994	0.999	0.864	**1.000**	**1.000**	**1.000**	0.999	**1.000**	0.993	0.983	2F + 1J
ReSCOPED DINOv3 ViT-B	0.968	0.996	0.919	1.000	1.000	1.000	1.000	1.000	0.996	0.986	2F + 1J
ReSCOPED DINOv3 ViT-L	0.969	0.997	0.969	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.995	0.992	2F + 1J

E LABEL-FREE MAHALANOBIS AND RESCOPED HYPERPARAMETERS

E.1 GLOBAL LABEL-FREE MAHALANOBIS

Given ID training latents $\{z_i\}_{i=1}^n$, we estimate

$$\mu = \frac{1}{n}\sum_{i=1}^{n} z_i, \qquad \Sigma_{\text{global}} = \frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)(z_i-\mu)^\top + \lambda I,$$

with $\lambda = 10^{-4}$. The global Mahalanobis score is

$$S_{\text{global}}(z) = (z-\mu)^\top \Sigma_{\text{global}}^{-1}(z-\mu).$$

Larger values indicate greater distance from the fitted ID representation distribution.

E.2 RESCOPED

ReSCOPED trains a latent diffusion observer on ID training latents. The EDM normalizer is fit on ID training latents and reused for scoring. At evaluation time, we use a fixed noise level $\sigma \approx 0.099$.

Let $s_\theta(z,\sigma)$ denote the learned diffusion score. ReSCOPED computes

$$T(z;\sigma) = \frac{\|s_\theta(z,\sigma)\|_2^2}{-\operatorname{Tr}(\nabla_z s_\theta(z,\sigma)) + \epsilon}, \qquad \epsilon = 10^{-8}.$$

We fit a one-dimensional Gaussian-kernel KDE to ID training values $\{T_i\}_{i=1}^n$ with bandwidth $h = 0.2$.

F Frozen Pretrained Models Used in This Work

Latent extraction. For each modality, we extract a single frozen embedding $\phi(x)$ and use it as the input to both label-free Mahalanobis and ReSCOPED. The DINOv3, ViT-B/16, and Qwen3 checkpoints are sourced from Hugging Face but evaluated through Bonsai/JAX (Google, 2024). DINOv1, DINOv2, and BERT are evaluated through Hugging Face `transformers` (Wolf et al., 2019). For vision models, we use a CLS-style global image representation. In the Bonsai/JAX ViT and DINOv3 models, this corresponds to the post-final-layer-normalization CLS token; for DINOv3, this vector is exposed through the pooled-output field. For BERT, we compute an utterance embedding by sequence mean-pooling token hidden states. For Qwen3, we compute an utterance embedding by sequence mean-pooling non-padding token hidden states. Thus, all backbones provide one fixed-length representation per input, but the pooling mechanism differs across modalities: vision uses a designated global token, while language uses explicit sequence mean pooling over token states.

G Datasets and Preprocessing

Small-scale vision datasets. We use `torchvision.datasets` with canonical train/test splits for CIFAR-10, CIFAR-100, SVHN, and CelebA. Mahalanobis and ReSCOPED are fit using ID train latents and evaluated on ID test versus OOD test.

Vision preprocessing. Images are converted to RGB, resized to the model input resolution, scaled to $[0, 1]$, and normalized with ImageNet mean and standard deviation. For CelebA in the small-scale vision comparisons, images are first resized to 32×32 before being resized to the model input resolution, matching the comparison protocol of Heng et al. (2024).

OpenOOD ImageNet-200. For ImageNet-200 experiments, we use the OpenOOD image lists for ImageNet-200 ID train/test and SSB-hard or NINCO OOD test sets. Images are decoded to RGB, resized to 224×224, and ImageNet-normalized. Detectors are fit on ImageNet-200 train latents and evaluated on ImageNet-200 test versus OOD test.

CLINC150. We use `datasets.load_dataset("clinc_oos", "plus")`. In-domain train examples are used for fitting the detector, and evaluation is performed on in-domain test examples versus OOS test examples. The validation split is not used for detector training or reported test metrics.

SST cross-corpus. For SST cross-corpus evaluation, SST train examples are used for fitting, SST test examples are used as ID test data, and OOD examples are drawn from the specified cross-corpus sources. We sample 500 examples from each OOD source with random seed 42 as in Xu et al. (2021).

H Per Backbone Reproduction of Figure 5 Without Normalization

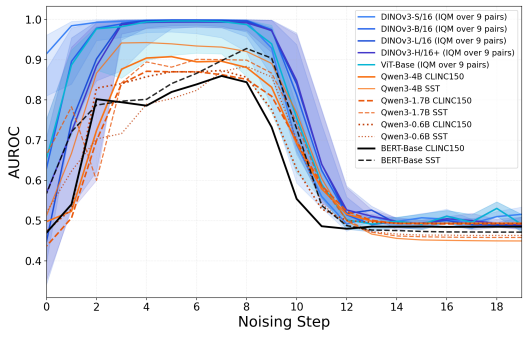


Figure 7: AUROC scores for all of the backbones that comprise Figure 5.

	# features	CLINC150 AUROC	CLINC150 DTACC	CLINC150 AUIN	CLINC150 AUOUT	SST AUROC	SST DTACC	SST AUIN	SST AUOUT
Frozen Backbone Approaches									
EDF	12	0.553	0.552	0.843	0.203	0.901	0.848	0.928	0.842
MDF	12	0.767	0.711	0.934	0.382	0.933	0.875	0.949	0.891
EC-NNK-Means	768	0.672	0.818	0.895	0.289	–	–	–	–
BERT + Label-free Mahalanobis + SMP	1	0.791	0.828	0.939	0.447	0.965	0.907	0.972	0.953
BERT + ReSCOPED + SMP	1	0.815	0.828	0.949	0.463	0.802	0.730	0.822	0.778
Fine-Tuning Approaches									
IMLM + MDF	12	0.778	0.722	0.938	0.391	0.936	0.881	0.975	0.894
BCAD + MDF	12	0.812	0.745	0.946	0.474	0.970	0.945	0.980	0.948
IMLM + BCAD + MDF	12	0.821	0.756	0.950	0.476	**0.981**	**0.954**	**0.987**	**0.959**
Scaling Behavior (Modern Backbones)									
Qwen3-0.6B + Label-free Mahalanobis + SMP	1	0.856	0.859	0.960	0.617	0.939	0.871	0.941	0.938
Qwen3-1.7B + Label-free Mahalanobis + SMP	1	0.867	0.858	0.965	0.614	0.934	0.859	0.941	0.928
Qwen3-4.0B + Label-free Mahalanobis + SMP	1	**0.903**	**0.872**	**0.976**	**0.683**	0.951	0.885	0.954	0.947
Qwen3-0.6B + ReSCOPED + SMP	1	0.856	0.825	0.951	0.643	0.764	0.698	0.757	0.763
Qwen3-1.7B + ReSCOPED + SMP	1	0.861	0.856	0.963	0.599	0.876	0.811	0.886	0.853
Qwen3-4.0B + ReSCOPED + SMP	1	0.900	**0.872**	0.974	0.669	0.949	0.881	0.952	0.945

Table 2: OOD detection performance on CLINC150 and SST. Larger values indicate better performance. BERT baselines are reproduced from Xu et al. (2021). The Qwen3 rows evaluate scaling behavior in modern frozen language backbones using label-free Mahalanobis and ReSCOPED on mean-pooled hidden representations.

Table 4: Hyperparameters for global label-free Mahalanobis.

Quantity	**Value**
Input data	Unlabeled ID training latents
Covariance ridge λ	10^{-4}
Evaluation	ID test vs. OOD test

Table 5: Hyperparameters for ReSCOPED.

Quantity	**Value**
Input data	Unlabeled ID training latents
Training steps	150,000
Batch size	256
Optimizer	Adam
Learning rate	3×10^{-4} with cosine decay
Denoiser	Residual MLP, width 1024, 6 layers
Time embedding dimension	256
Scoring noise level	$\sigma \approx 0.099$
KDE bandwidth	0.2
Evaluation	ID test vs. OOD test

Table 6: Frozen pretrained backbones and latent representations used by the OOD detectors.

Backbone	**Checkpoint**	**Inference**	**Latent representation**
DINOv3 ViT-S/B/L	`facebook/dinov3-{vits16,vitb16,vitl16}-pretrain-lvd1689m`	Bonsai/JAX	post-norm CLS/global image vector
ViT-B/16	`google/vit-base-patch16-224`	Bonsai/JAX	post-norm CLS vector before classifier
DINOv2 ViT-B	`facebook/dinov2-base`	Hugging Face Transformers	CLS/global image vector
DINOv1 ViT-B/16	`facebook/dino-vitb16`	Hugging Face Transformers	CLS/global image vector
BERT	`bert-base-uncased`	Hugging Face Transformers	mean-pooled token hidden states
Qwen3	`Qwen/Qwen3-{0.6B,1.7B,4B}`	Bonsai/JAX	mean-pooled non-padding token hidden states