# Tippett-minimum Fusion of Representation-space Diffusion Models for Multi-Encoder Out-of-Distribution Detection

**Neelkamal Bhuyan** [1]

## Abstract

We address out-of-distribution (OOD) detection across the full spectrum of distribution shifts—global domain changes, semantic divergence, texture differences, and covariate corruptions—through a multi-encoder fusion of per-encoder representation-space diffusion models (RDMs). We statistically identify each encoder's sensitivity to specific shift types from ID data alone and introduce EncMin2L—an encoder-agnostic two-level $\min(\cdot)$-gate that combines and calibrates per-encoder diffusion-based likelihood detectors without OOD labels, outperforming monolithic multi-encoder baselines at $2.3\times$ lower parameter cost. Two ID-data diagnostics: $\eta^2$ (class-conditional F-test) and $\Delta\mu$ (log-likelihood shift under synthetic corruptions)—quantify encoder specialization, while a Tippett minimum $p$-value combination aggregates per-encoder scores into a single, calibration-stable OOD signal. EncMin2L achieves $\geq 0.94$ AUROC across all four shift types simultaneously, outperforming the state-of-the-art representation-space diffusion OOD detectors across overlapping benchmarks.

## 1. Introduction

Out-of-distribution (OOD) detection asks whether a given image was generated by the same process as the true data, often known as the in-distribution (ID) data. A standard approach fits a density model to ID training data and flags a test point as OOD when its log-likelihood $\log p_\theta(z)$ falls below a threshold calibrated on held-out ID samples (Xiao et al., 2020). The challenge lies in building density models expressive enough to capture the full structure of the ID distribution yet sensitive to the specific forms of shift that arise in practice.

Deep generative models have become the dominant framework for learning complex data distributions, with diffusion models demonstrating remarkable fidelity across image generation, density estimation, and downstream tasks including anomaly detection and domain adaptation (Ho et al., 2020; Livernoche et al., 2024). Score-matching stochastic differential equations (SDEs) (Song et al., 2021b) provide a principled framework: a neural network learns the score function $\nabla_z \log p(z)$ via a denoising objective, and the associated probability flow ODE (PF-ODE) enables numerically accurate computation of $\log p_\theta(z)$ for any test point (Song et al., 2021a). This makes score-based diffusion models natural candidates for likelihood-based OOD detection.

Operating diffusion models directly in pixel space is computationally expensive, and pixel-space density estimates are notoriously unreliable for OOD (Xiao et al., 2020; Graham et al., 2023). Representation-space diffusion models (RDMs) address this by operating in the compressed representation space of a pretrained encoder—capturing underlying semantic structure by removing pixel-level variation (Rombach et al., 2022). Modern self-supervised encoders are each capable of meaningfully compressing $224\times224$ images into dense semantic vectors: CLIP (Radford et al., 2021) encodes global visual-linguistic alignment; DINOv2 (Oquab et al., 2024) captures fine-grained patch-level semantics; and ResNet-50 (He et al., 2016) penultimate features encode texture and low-level image statistics.

This compression is not lossless—each encoder emphasises a different aspect of image structure and is therefore sensitive to a different class of distribution shift. In §4 we quantify this directly: CLIP detects global domain shift but is nearly blind to covariate corruptions; DINOv2 separates fine-grained semantic classes but its features are largely invariant to image corruptions; ResNet-50 is the strongest covariate-shift detector but ranks last on semantic benchmarks. This raises a fundamental question:

> *Given a library of pretrained encoders, which subset should be deployed, and how should their signals be combined, when the type of distribution shift is unknown at test time?*

**Main Contributions**: We make two contributions that together answer this question. **(i) Quantifying encoder sen-**

[1]Georgia Institute of Technology. Correspondence to: Neelkamal Bhuyan <nbhuyan3@gatech.edu>.

*ICML 2026 Workshop on Hypothesis Testing*, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

**sitivity from ID data.** We quantify each encoder's OOD sensitivity through ID-data diagnostics ($\eta^2$, $\Delta\mu$, and cross-encoder Spearman $\rho$) and characterise how representation space geometry governs sensitivity to covariate versus semantic distribution shifts. These diagnostics, computed entirely from ID training data, reveal strict encoder specialisation: no encoder dominates across all shift types simultaneously, and any fixed weighting of encoder contributions will underperform on at least one shift type. **(ii) Two-level $\min(\cdot)$-gate with statistical calibration.** We combine per-encoder VP-SDE density models through a two-level $\min(\cdot)$-gate grounded in Tippett's minimum $p$-value test, resulting in our **EncMin2L** framework. At each level, raw log-likelihood scores are first mapped through the ID empirical CDF to yield calibrated $p$-values on $[0, 1]$. For every encoder, a $\min(\cdot)$ operation across $n$ different input-normalization forks collapses $n$ complementary views of each encoder's representation space into a single per-encoder score (Level 1), followed by an explicit CDF-based re-scaling that corrects for the non-uniform $\mathrm{Beta}(1, n)$ null distribution of the resulting minimization operation. A second Tippett minimization across all encoders' Level 1 scores, with a joint threshold $\tau$ calibrated on held-out ID data, constitutes the final OOD decision (Level 2). The result is a statistically valid multiple testing procedure in which the input sample itself forces the most-alarmed representation-space diffusion OOD detector to make a decision—adapting to the input without requiring any knowledge of the distribution shift at test time.

We evaluate on four OOD benchmarks spanning global domain shift (SVHN), fine-grained semantic shift (CIFAR-100), texture shift (DTD), and covariate shift (CIFAR-10C). EncMin2L achieves $\geq$0.94 AUROC across all four benchmarks, outperforming monolithic multi-encoder baselines at 2.3$\times$ lower parameter cost. Leave-one-out ablations confirm that each encoder specialises on the shift type predicted by its ID diagnostic profile, providing post-hoc validation of the framework.

## 2. Related Work

**Density-based OOD detection.** Using a generative model's log-likelihood as an OOD score is a longstanding approach. Grathwohl et al. (2020) reformulate a classifier as an energy-based model; Xiao et al. (2020) propose likelihood regret for VAEs to correct for the pathological behaviour of deep generative model likelihoods on OOD inputs. Graham et al. (2023) apply diffusion models directly in pixel space via multi-step reconstruction error; Liu et al. (2023) score checkerboard-masked inpainting with LPIPS (LMD); Heng et al. (2024) extract a compact 6-dimensional descriptor from the denoising trajectory of a pixel-space DM (DiffPath), demonstrating zero-shot cross-distribution generalization but at the cost of coarse OOD resolution. Our work instead trains density models in the compressed representation space of pretrained encoders, where the density is better-behaved and computationally tractable. Järve et al. (2025) independently pursue this direction: their Variational Diffusion Model (VDM) trains a learnable-schedule representation-space diffusion model on a single ResNet-18 encoder and achieves strong AUROC on CIFAR-10 ID benchmarks. Concurrently, Ding et al. (2025) train a sub-VP SDE score model on a single pretrained encoder's representation space at ImageNet scale, with a class-conditioned variant (ConRDM) that matches supervised detectors when labels are available. Score-based SDE models (Song et al., 2021b) provide numerically accurate log-likelihoods via the probability flow ODE.

**Feature-space methods.** A complementary family of methods detects OOD samples by their position in the feature space of a pretrained or jointly trained network, without fitting an explicit density. Lee et al. (2018) use the Mahalanobis distance to class-conditional Gaussian fits in feature space. Liu et al. (2020) exploit the energy function implicit in classifier logits. Sun et al. (2022) show that $k$-NN distance in the penultimate feature space outperforms many parametric alternatives. These methods depend on a single representation space; combining multiple encoders requires an explicit fusion strategy.

**Multi-encoder and ensemble detection.** Ensemble and multi-view approaches to OOD detection are relatively unexplored. Shalev et al. (2018) show, in a text setting, that replacing a single label embedding with multiple dense semantic representations from different corpora or architectures improves OOD discrimination—an early demonstration that representation diversity across sources provides complementary detection signal. Prior works that combine multiple representation spaces in vision typically do so by concatenation or majority voting, without a theoretical account of when combination helps and when it is redundant.

**Multiple hypothesis testing.** The minimum $p$-value statistic for combining independent tests is classical: Tippett's test rejects the global null when $\min_k p_k < \tau$, with $\tau$ set to control the family-wise error rate (Westfall & Young, 1993). Fisher's method (Westfall & Young, 1993) provides an alternative via $-2\sum_k \log p_k \sim \chi^2_{2K}$, which is more powerful when multiple tests simultaneously reject; Tippett's test (and our $\min(\cdot)$-gate) is more powerful when a single test dominates. Given the encoder specialization we observe—each encoder covers a distinct shift type—Tippett's approach is the natural choice: one encoder typically dominates for any given OOD shift type.

# 3. Preliminaries

## 3.1. Representation Space and Notation

Let $\{E_1, \ldots, E_K\}$ be a fixed set of pretrained image encoders, $E_k : \mathcal{X} \to \mathbb{R}^{d_k}$. We use $K = 3$: CLIP ViT-B/32 ($d_1 = 512$), DINOv2 ViT-B/14 ($d_2 = 768$), and ResNet-50 penultimate layer ($d_3 = 2048$). For each encoder we define $n = 2$ *normalization forks*:

$$z_{k,\mathrm{n}} = (E_k(x) - \mu_k)/\sigma_k, \qquad z_{k,\mathrm{u}} = E_k(x),$$

where $\mu_k, \sigma_k$ are per-dimension statistics estimated on the ID training set only. Normalization amplifies fine-grained structure within each encoder's distribution; the unnormed fork retains the raw geometry.

*Remark* 1. *Within each encoder, the normed fork is more sensitive to covariate shifts and the unnormed fork holds a marginal edge on semantic OOD—a consequence of per-dimension Z-score normalization expanding the dynamic range of score differences under perturbations.*

## 3.2. Score Network Architecture and Per-Encoder Density Models

**Per-encoder models.** For each fork $(k, f)$ we train an independent VP-SDE score network (Song et al., 2021b) with linear noise schedule $\beta(t) = \beta_{\min} + t(\beta_{\max} - \beta_{\min})$, $\beta_{\min} = 0.1$, $\beta_{\max} = 20$. The score network is an MLP with hidden dimension $H_k = 2d_k$ and depth $L = 6$ residual layers—capacity proportional to the encoder's intrinsic complexity. All models are trained with Adam (Song et al., 2021b, lr=$2\times10^{-4}$, up to 500 epochs with early stopping), batch size 512, gradient clipping (max norm 1.0). The probability flow ODE (PF-ODE) corresponding to the VP-SDE provides an exact log-likelihood via the instantaneous change-of-variables formula and Hutchinson trace estimation with $M$ random probes:

$$\ell_{k,f}(z) \triangleq \log p_\theta^{(k,f)}(z) = \log p_T(\tilde{z}) - \int_0^T \nabla \cdot f_\theta(\tilde{z}(t), t)\, dt, \tag{1}$$

where $f_\theta$ is the PF-ODE drift and $p_T$ is the terminal Gaussian prior. We use $M = 10$ Rademacher probe vectors and an adaptive ODE solver with relative and absolute tolerances of $10^{-5}$. Let $\hat{F}_{k,f}$ be the empirical CDF of $\{\ell_{k,f}(z_i)\}$ over ID validation samples. The transformed score

$$r_{k,f}(z) = \hat{F}_{k,f}(\ell_{k,f}(z)) \tag{2}$$

is a $p$-value: $r_{k,f}(z) \approx \mathrm{Uniform}(0, 1)$ for ID samples by the probability integral transform; small values indicate OOD.

**Why monolithic concatenated models underperform.** A single score network trained on the full concatenated representation $z = [z_1; \ldots; z_K] \in \mathbb{R}^D$, $D = \sum_k d_k$, must enforce block-diagonal coupling in order to preserve the individual geometry of each encoder's representation space: cross-encoder weight sharing conflates the distinct distributional structures of different encoders. Under this constraint, the score-matching loss $\mathbb{E}\left[\|s_\theta(z+\sigma\epsilon, t) + \epsilon/\sigma\|^2\right]$ is minimized by a score that acts independently on each partition—the optimal score $s^*(z_k, t)$ in partition $k$ depends only on $z_k$—so the model converges to a product of per-encoder marginals $\log p_\theta(z) = \sum_k \log p_k(z_k)$, equivalent to separate per-encoder models but at $2.3\times$ more parameters.

*Table 1.* Parameter counts per model. $H_k = 2d_k$ for per-encoder models; the monolithic Score Network uses $H_k = 4d_k$ per encoder block.

| Model | $H$ | Params |
|---|---|---|
| E1 (CLIP, $d_1 = 512$) | 1024 | 7.5M |
| E2 (DINOv2, $d_2 = 768$) | 1536 | 16.7M |
| E3 (ResNet, $d_3 = 2048$) | 4096 | 118.0M |
| Per-encoder suite (6 models) | — | **285M** |
| Monolithic Score Network | 13312 | 647M |

**Monolithic baseline.** For comparison we train a monolithic score network on the full concatenated representation $z \in \mathbb{R}^{3328}$. It uses a block-diagonal MLP with three encoder-aligned blocks of hidden dimension $H_k = 4d_k$ (yielding block widths 2048, 3072, 8192 for CLIP, DINOv2, and ResNet-50 respectively), a total hidden dimension of 13,312, and $L = 6$ residual layers—trained with the same VP-SDE objective as the per-encoder models. As shown in Table 3, the $H_k = 2d_k$ model fails on three of four benchmarks (SVHN 0.799, CIFAR-100 0.617, CIFAR-10C 0.631); $H_k = 4d_k$ is the minimum scale at which the monolithic model reaches $\geq 0.90$ AUROC across all shift types, ruling out the possibility that the cost gap is an artefact of an over-parameterised baseline.

**Model capacities.** Table 1 compares parameter counts. The six per-encoder models total 285M parameters; the monolithic Score Network baseline totals 647M. Per-encoder training is also parallelizable and can target smaller GPU budgets.

# 4. Encoder Analysis and Combination Design

## 4.1. ID-Data Diagnostics: Characterising Test Power

Different near-OOD benchmarks present distinct pain points—global domain shift, fine-grained semantic shift, and covariate corruption— where prior methods most often fail, each demanding a different kind of encoder sensitivity. The following diagnostics surface these differences from ID data alone.

*Table 2.* ID-data diagnostics for all per-encoder models alongside observed Near-OOD performance. $\eta^2_c$: ANOVA effect size on clean ID images. $\eta^2_{corr}$: same, under synthetic corruptions. $\Delta\mu$: mean log-likelihood shift under synthetic corruptions. All from ID data only.

| Model | $\eta^2_c$ | $\eta^2_{corr}$ | $\Delta\mu$ | SVHN | CIFAR-100 | CIFAR-10C |
|---|---|---|---|---|---|---|
| E1 normed-fork (CLIP) | 0.096 | 0.092 | 282 | 0.977 | 0.841 | 0.907 |
| E1 unnormed-fork (CLIP) | 0.133 | 0.106 | 180 | 0.975 | 0.830 | 0.887 |
| E2 normed-fork (DINOv2) | 0.243 | 0.112 | 81 | 0.859 | 0.963 | 0.568 |
| E2 unnormed-fork (DINOv2) | 0.154 | 0.084 | 93 | 0.868 | 0.964 | 0.569 |
| E3 normed-fork (ResNet) | 0.109 | 0.046 | 4171 | 0.762 | 0.770 | 0.981 |
| E3 unnormed-fork (ResNet) | 0.071 | 0.015 | 691 | 0.737 | 0.697 | 0.926 |

**$\eta^2$: class-conditional F-test.** Group $\{\ell_{k,f}(z_i)\}$ on the ID validation set by class label. The ANOVA effect size

$$\eta^2_{k,f} = \frac{\mathrm{SS}_{\mathrm{between}}}{\mathrm{SS}_{\mathrm{total}}} \tag{3}$$

measures the fraction of log-likelihood variance explained by class identity. High $\eta^2$ means the density model reflects fine-grained class structure, predicting high sensitivity to semantic OOD (e.g. CIFAR-100). Conversely, a low-$\eta^2$ model has learned a broad ID density that is not overfit to class detail, predicting sensitivity to global domain shift (SVHN) where class distinctions are irrelevant. We find Spearman $\rho = 0.771$ between $\eta^2$ and CIFAR-100 AUROC across all six per-encoder models.

**$\Delta\mu$: synthetic corruption probe.** Apply a fixed synthetic corruption to clean ID test images (GaussianBlur, kernel 3, $\sigma = 1.0$; additive noise $\sigma = 0.1$) and compute:

$$\Delta\mu_{k,f} = \mu_{\mathrm{clean}} - \mu_{\mathrm{corrupt}}. \tag{4}$$

All corrupted images are derived from ID data; no OOD dataset is consulted. Large $\Delta\mu$ indicates corruption-sensitive features, predicting high power against covariate shift. We find Spearman $\rho = 1.0$ between $\Delta\mu$ and CIFAR-10C AUROC. DINOv2 is designed for corruption invariance (Oquab et al., 2024): its $\Delta\mu \approx 87$ (vs. 4171 for ResNet) correctly predicts its CIFAR-10C AUROC of 0.568 vs. 0.981.

**$\rho$: cross-encoder correlation screening.** Spearman rank correlations between per-encoder log-likelihood vectors on the ID test set quantify redundancy. Within-encoder fork pairs (normed/unnormed) yield $\rho \approx 0.92$: nearly redundant on ID data, but differing in OOD-space geometry (the normed fork amplifies corruption sensitivity; the unnormed fork preserves raw semantic geometry). Cross-encoder pairs yield $\rho \in [0.17, 0.39]$: genuinely complementary, justifying multi-encoder combination.

*Remark 2. The $\eta^2$ and $\Delta\mu$ diagnostics reveal a strict encoder specialisation: no single encoder dominates across all shift types simultaneously (Figure 1, Table 2). CLIP maximises global domain-shift detection but underperforms on covariate shift; DINOv2 leads on semantic OOD but is nearly blind to corruptions ($\Delta\mu \approx 87$ vs. 4171 for ResNet); ResNet-50 dominates CIFAR-10C but ranks last on SVHN and CIFAR-100.*

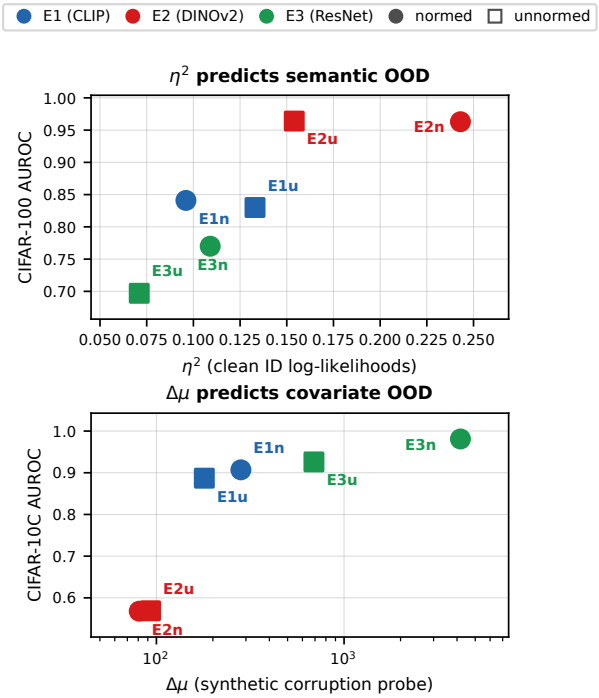


*Figure 1.* ID-data diagnostics predict OOD capability without observing any OOD samples. **Left**: $\eta^2$ (class structure in ID log-likelihoods) vs. CIFAR-100 AUROC (Spearman $\rho = 0.771$). **Right**: $\Delta\mu$ (score shift under synthetic corruptions of ID images) vs. CIFAR-10C AUROC (Spearman $\rho = 1.0$). Points are labelled E1n/E1u/E2n/E2u/E3n/E3u (n = normed, u = unnormed).

Table 2 further adds the $\eta^2$ under synthetic corruption ($\eta^2_{corr}$). Under corruptions, DINOv2's $\eta^2$ halves (0.243→0.112): class structure dissolves because DINOv2 features are explicitly trained to be corruption-invariant. CLIP's $\eta^2$ is essentially stable (0.096→0.092), consistent with its known sensitivity to class-discriminative corruption pat-

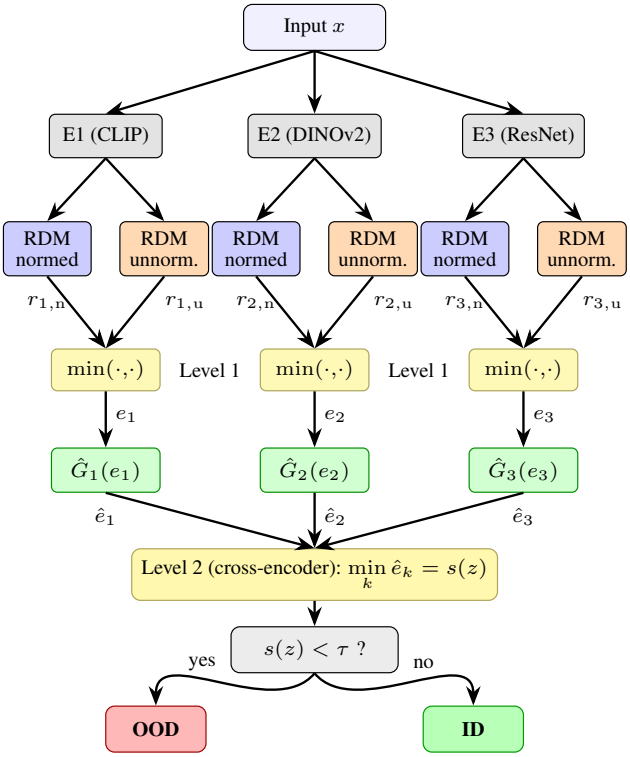


*Figure 2.* Architecture of the two-level $\min(\cdot)$-gate (EncMin2L). Each encoder produces two $p$-values: $r_{k,\mathrm{n}}$ (blue, normed repr.) and $r_{k,\mathrm{u}}$ (orange, unnormed repr.). Level 1 takes the within-encoder minimum (Eq. 5), re-CDF recalibration (Proposition 1) restores uniformity, and Level 2 takes the cross-encoder Tippett minimum (Eq. 6) to produce the final score $s(z)$.

terns. ResNet's $\eta^2$ drops sharply (0.109→0.046) because low-level texture features—the most corruption-sensitive—are class-agnostic.

### 4.2. Two-Level $\min(\cdot)$-gate Combination

Because the type of distribution shift is unknown at test time, any *fixed* weighting of encoder contributions—including a monolithic concatenated model—will underperform on at least one shift type. The combination must therefore be *input-driven*: deferring to whichever encoder raises the strongest alarm for a given test point, without committing to any encoder in advance.

**Level 1 (within-encoder fork aggregation).**

$$e_k(z) = \min\big(r_{k,\mathrm{n}}(z),\ r_{k,\mathrm{u}}(z)\big). \tag{5}$$

This declares OOD if *either* fork of encoder $k$ raises an alarm. The distribution of $e_k$ under $H_0$ is no longer uniform:

**Proposition 1** (Re-calibration between levels)**.** *Let* $U_1, U_2 \overset{iid}{\sim} \mathrm{Uniform}(0,1)$. *Then* $M = \min(U_1, U_2) \sim \mathrm{Beta}(1,2)$ *with* $\mathbb{E}[M] = \frac{1}{3}$ *and PDF* $f_M(m) = 2(1-m)$. *More generally,* $\min_k U_k \sim \mathrm{Beta}(1,K)$. *Applying the empirical CDF* $\hat{G}_k$ *of* $\{e_k(z_i)\}_i$ *(ID validation) re-maps* $e_k$ *to an approximately uniform distribution.*

We therefore set $\hat{e}_k(z) = \hat{G}_k(e_k(z)) \in [0,1]$.

**Level 2 (cross-encoder aggregation).**

$$s(z) = \min_k\ \hat{e}_k(z). \tag{6}$$

Flag OOD if $s(z) < \tau$, where $\tau$ is calibrated on the ID validation set at target FPR $\alpha = 0.05$. This is Tippett's minimum $p$-value combination (Westfall & Young, 1993): reject the global null if any individual test rejects. By Proposition 1, $\hat{e}_k$ is approximately uniform on ID data, so $s(z)$ is well-calibrated for threshold setting.

## 5. Main Results and Experiments

EncMin2L is our proposed two-level $\min(\cdot)$-gate combination (§4.2): it aggregates per-encoder log-likelihood $p$-values via within-encoder minimum (Level 1), re-CDF recalibration (Proposition 1), and cross-encoder Tippett minimum (Level 2), calibrated on ID validation data alone with no OOD labels at any stage. We evaluate EncMin2L against classifier-based, feature-space, and diffusion-based baselines in the following experimental setting.

### 5.1. Setup

**ID dataset.** CIFAR-10 (32×32 RGB, bicubic-upsampled to 224×224 for encoding; 50K train, 10K test with 10% held out for validation).

**OOD benchmarks.** Five datasets covering distinct shift types: SVHN (Netzer et al., 2011) (global domain shift), CIFAR-100 (Krizhevsky et al.) (fine-grained semantic shift), CIFAR-10C (Hendrycks & Dietterich, 2019) (covariate shifts), DTD (Cimpoi et al., 2014) (texture shifts), and CelebA (Liu et al., 2015) (face domain).

**Metrics.** AUROC (↑) and FPR at 95% TPR (FPR@95, ↓; in parentheses throughout). Higher AUROC and lower FPR indicate better detection.

**Baselines.** *Per-encoder*: each of the six single-encoder models (E1/E2/E3 × normed/unnormed) as individual detectors. *Monolithic*: Score Network trained on the full $D = 3328$ concatenated representation in normed, unnormed, and dual-fork configurations. *Classifier-based*: Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2017), Mahalanobis distance (Lee et al., 2018), and energy score (Liu et al., 2020) applied to a ResNet-50 classifier trained on CIFAR-10. *Feature-space*: $k$-NN distance (Sun et al., 2022) ($k = 50$) in the representation space of each pretrained encoder.

### 5.2. Main Results

Table 3 reports results for all evaluated methods. EncMin2L achieves $\geq$ 0.94 AUROC on all four benchmarks simultaneously—a bar no individual encoder or monolithic baseline clears—outperforming the best monolithic competitor (dual-fork Score Network) by $+0.024$ on SVHN and $+0.047$ on CIFAR-100 while using $2.3\times$ fewer parameters.

The encoder specialisation predicted by the §4.1 diagnostics is directly visible in Table 3. CLIP (E1) leads on global domain shift (SVHN 0.977) but underperforms where semantic class structure matters, consistent with its low $\eta^2$. DINOv2 (E2) leads on semantic OOD (CIFAR-100 0.964) yet collapses on covariate shift (CIFAR-10C 0.568), exactly as its near-zero $\Delta\mu$ predicted. ResNet-50 (E3) inverts this: its large $\Delta\mu$ flags it as the covariate specialist, and it peaks at CIFAR-10C 0.981 while trailing on SVHN and CIFAR-100. Because each encoder has a structural blind spot on at least one shift type, EncMin2L's $\min(\cdot)$-gate defers to whichever encoder raises the strongest alarm per test point—and its worst-case AUROC of 0.951 (SVHN) reflects that no shift type is left uncovered. DTD and CelebA are saturated across all our models (AUROC $\geq 0.999$ on DTD; 1.000 on CelebA with FPR 0.000), reflecting the extreme perceptual distance of texture patches and face images from CIFAR-10 objects; we discuss both jointly in §5.3. The monolithic dual-fork score network is the strongest monolithic competitor, yet EncMin2L surpasses it on both SVHN and CIFAR-100.

### 5.3. Comparison with Feature-Space and Prior Methods

**Classifier-based methods** use a supervised ResNet-50 trained on CIFAR-10. Mahalanobis distance matches or exceeds EncMin2L on SVHN (0.957) and CIFAR-10C (0.981): class-conditional Gaussian distances capture large-scale domain and covariate shift well when ID class structure is informative of the shift. On CIFAR-100, however, Mahalanobis collapses to 0.830—the 10-class covariance model is too coarse to separate 100 semantically adjacent categories, exactly the limitation the $\eta^2$ framing exposes: a covariance model can only be as fine-grained as the label set it was trained with. MSP and energy score are weaker overall. None of the three achieves $\geq 0.94$ simultaneously across all four benchmarks; EncMin2L closes this gap without using class labels at any stage.

**Feature-space $k$-NN** ($k$=50) on the same pretrained features uses identical representations but replaces the learned density with a non-parametric distance score. DINOv2 $k$-NN (E2) exceeds EncMin2L on SVHN (0.978 vs. 0.951) and CIFAR-100 (0.976 vs. 0.942), yet collapses on CIFAR-10C (0.867). The diagnostics explain this directly: DINOv2's corruption-invariant features ($\Delta\mu = 28$) render training neighbours nearly indistinguishable from corrupted images, so nearest-neighbour proximity is not a reliable covariate-shift signal. ResNet-50 $k$-NN inverts the pattern—strong on SVHN (0.970) and CIFAR-10C (0.936) but poor on CIFAR-100 (0.775)—its large $\Delta\mu$ conferring covariate sensitivity at the cost of fine-grained semantic discrimination. No single-encoder $k$-NN achieves $\geq 0.94$ on all four benchmarks.

**DTD and CelebA: saturation gap.** All per-encoder models and EncMin2L achieve AUROC $\geq 0.999$ on DTD and 1.000 on CelebA (FPR 0.000), as do CLIP and DINOv2 $k$-NN. Classifier-based methods do not: MSP 0.856, Energy 0.809, Mahalanobis 0.772—a supervised ResNet-50 trained on CIFAR-10 object classes does not reliably flag face images because they have no equivalent class boundary to cross. Pixel-space diffusion methods similarly fail to saturate: EigenScore 0.873, DiffPath 0.897, LMD 0.834, DM Dual Threshold 0.648 on CelebA; EigenScore 0.756 and Projection Regret 0.910 on DTD. Modelling the ID distribution in representation space, where face and texture images have no support, is what makes the gap disappear.

**Head-to-head: Representation-space diffusion on CIFAR-10.** Järve et al. (2025) independently apply the same core hypothesis on the same CIFAR-10 ID dataset, training a Variational Diffusion Model (VDM) on ResNet-18 penultimate features (512-d). EncMin2L outperforms their best likelihood-based score (VDM log $p_T$) by $+0.012$ on SVHN (0.951 vs. 0.939) and $+0.057$ on CIFAR-100 (0.942 vs. 0.885), consistent with encoder diversity: CLIP and DINOv2 capture semantic cues inaccessible to a single ResNet-18. They also report TKDL—a class-conditioned score requiring labels at inference—as their best single-encoder method, reaching 0.900 on SVHN. EncMin2L surpasses TKDL by $+0.051$ while remaining fully unsupervised, showing that multi-encoder combination recovers gains otherwise achievable only through supervision.

*Remark* 3. *Ding et al. (2025) independently apply a sub-VP SDE score model on a single pretrained encoder's representation space at ImageNet scale, validating the same core hypothesis. Architecturally, this is equivalent to the single-encoder models in Table 3—a setting where EncMin2L consistently outperforms every individual encoder across all four benchmarks.*

**Head-to-head: Pixel-space diffusion on CIFAR-10.** Heng et al. (2024) propose DiffPath, which extracts a 6-dimensional statistic from the score-function trajectory of a pixel-space DM and fits a GMM in that space. A CelebA-trained model is applied zero-shot to CIFAR-10 tasks, yielding AUROC 0.910 on SVHN, 0.590 on CIFAR-100, and 0.923 on DTD. EncMin2L outperforms on both: $+0.041$ on SVHN and $+0.352$ on CIFAR-100. The CIFAR-100 gap is particularly large: pixel-space trajectories capture global distribution shift but miss the fine-grained semantic structure that representation-space log-likelihoods encode—precisely the class-conditional variance that $\eta^2$ (0.096–0.243) mea-

*Table 3.* AUROC (FPR@95%TPR) on five OOD benchmarks (CIFAR-10 as ID). **Bold**: **EncMin2L is the only method with $\geq$0.94 on all four non-saturated benchmarks.** "—": not evaluated by that work. VDM uses a ResNet-18 encoder; DiffPath is a CelebA-trained pixel-space DM applied zero-shot.

| | *Near-OOD* | | | *Far-OOD* | |
|---|---|---|---|---|---|
| **Model** | **SVHN** | **CIFAR-100** | **CIFAR-10C** | **DTD** | **CelebA** |
| *Single-encoder per-fork models* | | | | | |
| E1 normed-fork (CLIP, 512-d) | 0.977 (0.139) | 0.841 (0.479) | 0.907 (0.551) | 1.000 (0.000) | 1.000 (0.000) |
| E2 raw-fork (DINOv2, 768-d) | 0.868 (0.410) | 0.964 (0.163) | 0.569 (0.957) | 1.000 (0.001) | 1.000 (0.003) |
| E3 normed-fork (ResNet-50, 2048-d) | 0.762 (0.811) | 0.770 (0.613) | 0.981 (0.090) | 0.999 (0.004) | 1.000 (0.000) |
| *Monolithic multi-encoder models* | | | | | |
| Score Network ($H_k = 2d_k$) | 0.799 (0.859) | 0.617 (0.981) | 0.631 (0.949) | 0.961 (0.171) | 1.000 (0.000) |
| Score Network ($H_k = 4d_k$) | 0.927 (0.388) | 0.895 (0.388) | 0.973 (0.127) | 1.000 (0.000) | 0.999 (0.000) |
| **EncMin2L (ours)** | **0.951 (0.309)** | **0.942 (0.255)** | **0.963 (0.195)** | **0.999 (0.000)** | **1.000 (0.000)** |
| *Classifier-based methods (ResNet-50 on CIFAR-10)* | | | | | |
| MSP (Hendrycks & Gimpel, 2017) | 0.916 (0.544) | 0.885 (0.599) | 0.915 (0.534) | 0.899 (0.548) | 0.856 (0.705) |
| Mahalanobis (Lee et al., 2018) | 0.957 (0.238) | 0.830 (0.676) | 0.981 (0.093) | 0.957 (0.214) | 0.772 (0.779) |
| Energy (Liu et al., 2020) | 0.916 (0.409) | 0.862 (0.533) | 0.885 (0.583) | 0.849 (0.562) | 0.809 (0.708) |
| *Feature-space methods (same pretrained encoders, $k$=50)* | | | | | |
| $k$-NN (Sun et al., 2022) (CLIP repr.) | 0.930 (0.564) | 0.911 (0.439) | 0.921 (0.612) | 1.000 (0.000) | 1.000 (0.000) |
| $k$-NN (Sun et al., 2022) (DINOv2 repr.) | 0.978 (0.157) | 0.976 (0.110) | 0.867 (0.746) | 1.000 (0.001) | 1.000 (0.001) |
| $k$-NN (Sun et al., 2022) (ResNet repr.) | 0.970 (0.168) | 0.775 (0.784) | 0.936 (0.395) | 0.990 (0.051) | 0.911 (0.668) |
| *Recent diffusion-based OOD methods* | | | | | |
| EigenScore (Shoushtari et al., 2026) | 0.810 (—) | 0.880 (—) | — | 0.756 (—) | 0.873 (—) |
| VDM log $p_T$ (Järve et al., 2025) | 0.939 (0.213) | 0.885 (0.441) | — | 0.946 (0.227) | — |
| DiffPath-6D (Heng et al., 2024) | 0.910 (—) | 0.590 (—) | — | 0.923 (—) | 0.897 (—) |
| DM Dual Threshold (Kamkari et al., 2024) | 0.944 (—) | 0.560 (—) | — | — | 0.648 (—) |
| LMD (Liu et al., 2023) | 0.992 (—) | 0.607 (—) | — | — | 0.834 (—) |
| Proj. Regret (Choi et al., 2023) | 0.993 (—) | 0.775 (—) | — | 0.910 (—) | — |

sures and that is absent in raw pixel trajectories. DiffPath's zero-shot universality comes at the cost of ID-specific semantic sensitivity; EncMin2L trades universality for precise density modelling on the ID manifold.

**Head-to-head: LL+LID dual-threshold diffusion on CIFAR-10.** Kamkari et al. (2024) flag OOD if log-likelihood $\log p_\theta(\mathbf{x})$ or local intrinsic dimension (LID) falls below a threshold, targeting the paradox that low-dimensional OOD data can receive high density yet low probability mass. Their best DM variant achieves AUROC 0.944 on SVHN and 0.560 on CIFAR-100; EncMin2L exceeds both by $+0.007$ and $+0.382$ respectively. The large CIFAR-100 gap—despite the principled LL+LID combination—confirms that pixel-space LID cannot substitute for the semantic class structure that representation-space log-likelihoods encode.

**Head-to-head: Inpainting-based diffusion on CIFAR-10.** Liu et al. (2023) propose LMD, which masks an image with a checkerboard pattern, reconstructs via DDPM inpainting, and scores OOD-ness by LPIPS perceptual distance. On SVHN, LMD reaches 0.992, surpassing EncMin2L (0.951): inpainting fails visibly on house-number images, making the perceptual gap large and discriminative. On CIFAR-100, LMD drops to 0.607 ($-0.335$ vs. EncMin2L's 0.942): CIFAR-100 shares visual structure with CIFAR-10, so DDPM reconstruction succeeds and the perceptual score collapses—the class-conditional density structure ($\eta^2 = 0.096$–$0.243$) that drives our detection is invisible to pixel reconstruction.

**Head-to-head: Projection-based diffusion on CIFAR-10.** Choi et al. (2023) propose Projection Regret (PR), scoring OOD-ness by the LPIPS distance between a test image and its diffusion-based projection, with recursive projections to cancel background statistics. PR achieves AUROC 0.993 on SVHN, exceeding EncMin2L by $+0.042$: perceptual distance is highly discriminative when global domain statistics differ sharply. On CIFAR-100, however, PR drops to 0.775 ($-0.167$ vs. EncMin2L's 0.942): CIFAR-100 shares background statistics with CIFAR-10, so the projection reconstructs OOD images convincingly and the score collapses precisely where $\eta^2$ is highest. EncMin2L sacrifices only $-0.042$ on SVHN for AUROC$\geq 0.94$ across all benchmarks— balancing semantic sensitivity for global-shift detection.

**Head-to-head: Posterior-covariance diffusion on CIFAR-10.** Shoushtari et al. (2026) propose EigenScore, which uses the eigenvalue spectrum of a diffusion denoiser's posterior covariance as the OOD signal: OOD inputs inflate

*Table 4.* Leave-one-out encoder ablation for EncMin2L reporting AUROC (FPR@95%TPR) on Near-OOD benchmarks. All configurations calibrated on the same held-out ID split. **Bold**: full combination. ↓: AUROC drop relative to full combination.

| Configuration | SVHN | CIFAR-100 | CIFAR-10C |
|---|---|---|---|
| **E1+E2+E3 (full EncMin2L)** | **0.951 (0.301)** | **0.943 (0.257)** | **0.961 (0.212)** |
| E2+E3 (no E1 / no CLIP) | 0.829 (0.494) ↓0.122 | 0.948 (0.225) | 0.961 (0.184) |
| E1+E3 (no E2 / no DINOv2) | 0.958 (0.281) | 0.830 (0.512) ↓0.113 | 0.972 (0.161) |
| E1+E2 (no E3 / no ResNet) | 0.963 (0.236) | 0.953 (0.214) | 0.834 (0.727) ↓0.127 |

the leading eigenvalues, yielding a Jacobian-free spectral detector. EigenScore achieves AUROC 0.810 on SVHN and 0.880 on CIFAR-100; EncMin2L outperforms on both by $+0.141$ and $+0.062$ respectively. EigenScore's CIFAR-100 result (0.880) is the strongest among all pixel-space diffusion methods surveyed, suggesting posterior covariance partially recovers semantic structure—yet the residual $-0.062$ gap reflects class-conditional density geometry that $\eta^2$ (0.096–0.243) quantifies and that pixel-space denoisers cannot fully access.

### 5.4. Encoder Contribution: Leave-One-Out Ablation

Table 4 reports EncMin2L with each encoder removed in turn on the three near-OOD benchmarks; DTD and CelebA are omitted as all configurations saturate at $\geq 0.999$. Each drop directly confirms the specialisation the §4.1 diagnostics predicted from ID data alone.

**Removing E1 (CLIP)** cuts SVHN AUROC by $-0.122$ while leaving CIFAR-100 and CIFAR-10C stable: CLIP's low $\eta^2$ marks it as the global-domain-shift specialist, with little contribution elsewhere.

**Removing E2 (DINOv2)** drops CIFAR-100 AUROC by $-0.113$; SVHN and CIFAR-10C improve slightly, a complementarity effect explained by E2's near-zero $\Delta\mu$—its corruption-invariant features were a drag on covariate-shift benchmarks. DINOv2's high $\eta^2 = 0.243$ identified this semantic specialisation before any OOD data were observed.

**Removing E3 (ResNet-50)** causes the largest single drop: CIFAR-10C falls by $-0.127$, while SVHN and CIFAR-100 improve slightly, confirming that E3's large $\Delta\mu$ makes it indispensable for covariate shift.

## 6. Discussion

**Encoder Generalization Advantage.** The diagnostic pipeline applies to any pretrained encoder: compute $\eta^2$ and $\Delta\mu$ on ID data, screen for complementary pairs via Spearman $\rho$, then train per-encoder diffusion models only for encoders that add coverage. A natural screening criterion is $\rho < 0.5$ between the candidate encoder and all currently selected encoders; the remaining encoders each bring genuinely new information. This makes the framework a practical design-time tool for multi-encoder OOD detector construction without ever needing OOD examples.

**Recalibration-free Detection Advantage.** The joint threshold $\tau$ is calibrated at FPR $\alpha = 0.05$ on the ID validation set. Because the combined score $s(z)$ is approximately uniform on ID data (Proposition 1), changing $\alpha$ amounts to a direct rescaling of $\tau$: the empirical FPR at any other $\alpha'$ can be read directly from the ID validation CDF without recalibration. This makes the system straightforwardly adjustable for deployment scenarios with different false-positive budgets (e.g. conservative $\alpha = 0.01$ for safety-critical applications vs. permissive $\alpha = 0.10$ for screening).

**Limitations.** The synthetic corruption probe ($\Delta\mu$) uses a fixed two-stage perturbation (Gaussian blur + additive noise) to approximate covariate shift. While this correctly ranks encoder sensitivity in our experiments, it may not generalize to all covariate shift types (e.g. weather effects, sensor noise). Users should run the probe with domain-relevant corruptions when the expected shift type is known. The $\min(\cdot)$-gate's formal hypothesis testing guarantee (Tippett's test) requires the per-encoder $p$-values to be independent under $H_0$. In practice, Spearman correlations of 0.17–0.39 between encoder pairs indicate moderate dependence. Near-OOD cases where OOD distribution partially overlaps with ID distribution (e.g. CIFAR-10 "truck" vs. CIFAR-100 "truck" classes) remain challenging for all density-based methods.

## 7. Conclusion

We presented a framework that statistically identifies the strengths and weaknesses of pretrained encoders from ID data alone, and introduces EncMin2L—an encoder-agnostic two-level $\min(\cdot)$-gate for combining and calibrating per-encoder diffusion-based likelihood OOD detectors—achieving $\geq 0.94$ AUROC across all four shift types simultaneously and outperforming the state-of-the-art diffusion OOD detectors at 2.3× lower parameter cost. $\eta^2$ and $\Delta\mu$ rank CIFAR-100 and CIFAR-10C AUROCs with Spearman $\rho = 0.771$ and $\rho = 1.0$ respectively before any OOD examples are observed, and leave-one-out ablations confirm that each encoder contributes precisely where its diagnostic profile predicted.

## References


Choi, S., Lee, H., Lee, H., and Lee, M. Projection regret: Reducing background bias for novelty detection via diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Ding, Y., Aleksandraus, A., Ahmadian, A., Unger, J., Lindsten, F., and Eilertsen, G. Revisiting likelihood-based out-of-distribution detection by modeling representations. In *Scandinavian Conference on Image Analysis*, 2025.

Graham, M. S., Pinaya, W. H. L., Tudosiu, P.-D., Nachev, P., Ourselin, S., and Cardoso, M. J. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

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

Heng, A., Thiery, A. H., and Soh, H. Out-of-distribution detection with a single unconditional diffusion model. In *Advances in Neural Information Processing Systems*, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Järve, J., Haavel, K. K., and Kull, M. Probability density from latent diffusion models for out-of-distribution detection. *arXiv preprint arXiv:2508.15737*, 2025.

Kamkari, H., Ross, B. L., Cresswell, J. C., Caterini, A. L., Krishnan, R. G., and Loaiza-Ganem, G. A geometric explanation of the likelihood OOD detection paradox. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-100 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

Liu, W., Wang, X., Owens, J. D., and Li, Y. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

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

Liu, Z., Zhou, J. P., Wang, Y., and Weinberger, K. Q. Unsupervised out-of-distribution detection with diffusion inpainting. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Livernoche, V., Jain, V., Hezaveh, Y., and Ravanbakhsh, S. On diffusion modeling for anomaly detection. In *International Conference on Learning Representations*, volume 2024, pp. 25836–25866, 2024.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y., et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Shalev, G., Adi, Y., and Keshet, J. Out-of-distribution detection using multiple semantic label representations. In *Advances in Neural Information Processing Systems*, 2018.

Shoushtari, S., Wang, Y., Shi, X., Asif, S., and Kamilov, U. S. Eigenscore: OOD detection using posterior covariance in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Dq64kthckN.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 2022.

Westfall, P. H. and Young, S. S. *Resampling-Based Multiple Testing: Examples and Methods for $p$-Value Adjustment*. John Wiley & Sons, 1993.

Xiao, Z., Yan, Q., and Amit, Y. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *Advances in Neural Information Processing Systems*, 2020.