TASTE: TASK-AWARE OUT-OF-DISTRIBUTION DETECTION VIA STEIN OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution detection methods are often either data-centric, detecting deviations from the training input distribution irrespective of their effect on a trained model, or model-centric, relying on classifier outputs without explicit reference to data geometry. We propose TASTE (Task-Aware STEin operators): a task-aware framework which links distribution shift to the input sensitivity of the model. The resulting operators admit a geometric interpretation as a projection of distribution shift onto the sensitivity field of the model, with theoretical guarantees. Beyond detecting a shift, the construction enables its localisation through a coordinate-wise decomposition, and - for image data - provides interpretable per-pixel diagnostics. Experiments on controlled Gaussian shifts, MNIST under geometric perturbations, and CIFAR-10/ImageNet-1K perturbed benchmarks demonstrate that TASTE aligns closely with task degradation while showing competitive performance compared to established baselines, with gains that depend on the shift regime.

1 INTRODUCTION

Deep models are typically trained under the assumption that training and test samples are i.i.d. from a fixed joint distribution $P(X, Y)$. At deployment, this assumption frequently fails: the test-time marginal Q_X may differ due to changes in, for example, sensors or preprocessing pipelines. Such shifts are commonly categorised as *covariate shift* ($Q_X \neq P_X$ with $P(Y \mid X)$ unchanged), *label shift* (changes in $P(Y)$), and, in streaming or non-stationary settings, broader forms of *concept/conditional* shift affecting $P(Y \mid X)$ (Quiñonero-Candela et al., 2008; Gama et al., 2014; Lu et al., 2018). A central practical question is therefore not only whether $Q_X \neq P_X$, but whether the shift *meaningfully affects* the behaviour of the model f_θ of interest.

Limitations of current OOD approaches. Out-of-distribution (OOD) detection methods usually approach this question from two complementary angles. *Data-centric* methods model the training marginal P_X and flag inputs that appear unlikely, while *model-centric* methods operate on classifier outputs (e.g. softmax confidence) or learned representations (e.g. feature-space distances); see for example Hendrycks & Gimpel (2017); Liang et al. (2018); Liu et al. (2020); Nalisnick et al. (2019). Each captures part of the picture: data-centric scores can detect distribution change but do not specify whether it matters for the task, whereas model-centric scores can reveal brittleness yet do not explicitly connect f_θ to the geometry of the input distribution (Rabanser et al., 2019; Yang et al., 2024). What is missing is a principled mechanism that links the *geometry of the data* to the *sensitivity of the model* in a way that remains usable at test time.

Our perspective: Stein operators. To bridge data geometry and model sensitivity we propose a novel way of using Stein operators. Stein's method compares a candidate distribution q to a reference distribution p via an operator $\mathcal{T}_p$ satisfying $\mathbb{E}_p[\mathcal{T}_p g] = 0$ for a rich class of test functions g (Stein, 1972; Ley et al., 2017). Departures $\mathbb{E}_q[\mathcal{T}_p g] \neq 0$ certify $q \neq p$. Stein operators underpin discrepancy measures and goodness-of-fit tests such as kernelised Stein discrepancies, and more broadly provide a computational interface between score information and distributional mismatch, see the surveys Anastasiou et al. (2023); Liu et al. (2026). Crucially, most Stein-based discrepancy tests treat g as a free element of a function class and optimise or randomise over it to obtain *distribution-level* tests. In contrast, in deployment we already have a trained model $f_\theta(x) \approx \mathbb{E}[Y \mid X = x]$, and we are specifically interested in how a distribution shift affects *this* function. This motivates a model-centric

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.

question: What can we say about distribution shift and reliability if we fix the Stein test function to be the deployed predictor f_θ itself?

TASTE: Task-Aware Stein functional. Fixing the test function $f : \mathbb{R}^d \to \mathbb{R}$ of interest instead of considering a wide class of test functions is a novel use of Stein's method. We use it to develop **TASTE**, a task-aware framework built around the Langevin Stein operator

$$\mathcal{L}_p f(x) = \Delta f(x) + \nabla \log p(x)^\top \nabla f(x), \quad (1)$$

constructed from a (potentially unnormalised) density/score model p for training data and a fixed f. The *score function* $s_p = \nabla \log p$ in (1) connects the operator to score-based views of data geometry and score-matching models (Song et al., 2021). For each sample X, the value $\mathcal{L}_p f(X)$ gives a *per-sample Stein residual*, with mean 0 under p. We study what we call the *task-aware Stein, or TASTE, functional*

$$S_f(p, q) := \mathbb{E}_q[\mathcal{L}_p f(X)] \quad (2)$$

under a test distribution q. Under mild regularity assumptions $S_f(p, q)$ admits an explicit "projection" form (Eq. (5) below), which can be interpreted as an $L^2(q)$ inner product between the sensitivity field ∇f of the model, and the shift-score field $\nabla \log(q/p)$. This identity shows the task-aware nature of the signal: for fixed f, $S_f(p, q)$ responds to the components of distribution shift that influence ∇f in expectation; see Figure 1.

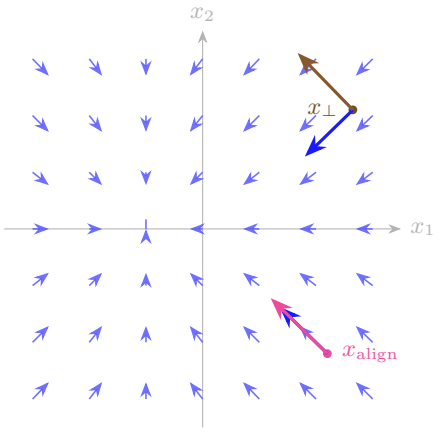


Figure 1: **Task-aware intuition behind TASTE.** Blue arrows show the score-shift field between the training density p and test density q. Without explicit knowledge of q, TASTE measures alignment between distribution shift and the input sensitivity of the model ∇f: shifts aligned with ∇f (magenta) produce a large response; orthogonal shifts (brown) are largely ignored even when large.

We focus on the Langevin Stein operator as it provides a scalar functional with a geometric interpretation. While simpler first-order Stein operators can also be constructed, they exhibit limitations such as directional blind spots and loss of linear sensitivity; see Appendix (App.) H.

Practical appeal of TASTE: plug-and-play, model-agnostic, interpretable. Our approach is *plug-and-play*: at test time it combines any model f with any score/density estimator for p and produces an OOD signal without retraining the classifier or requiring negative samples. It is *model-agnostic* in the sense that it only requires access to input-gradients of f (available via autodiff) and an estimate of $\nabla \log p$ (the quality of this score estimate influencing performance). Finally, the operator admits a natural *per-input-dimension* decomposition; for images this yields per-pixel residual maps, providing interpretable localisation of task-relevant anomalies and enabling fine-grained diagnostic visualisations.

Contributions. Our main contributions are as follows.

- **TASTE: model-centric Stein diagnostics.** We formalise a task-aware TASTE functional by applying the Langevin Stein operator to a fixed model, and derive a projection identity that links distribution shift to the sensitivity field of the model.
- **Robustness to score-model error.** We analyse how inaccuracies in learned score models affect the proposed methodology and provide a correction that removes systematic bias using training data.
- **Per-sample TASTE scores and interpretable decompositions.** We derive per-sample Stein residuals as OOD scores and show how their coordinate-wise decomposition yields per-input (and per-pixel) anomaly maps for image data.
- **Empirical validation across diverse shift regimes.** We demonstrate that the Stein signal aligns with task degradation in controlled shift models and achieves strong performance against standard OOD baselines on common benchmark suites (including corruption and perturbation benchmarks, mixed in-/out-distribution test sets, and adversarially perturbed inputs).

Scope. The goal of this paper is not to propose yet another OOD score in isolation, but to articulate a general framework in which a fixed f_θ and a density/score model for $p(x)$ are linked via a principled operator. Within this framework, TASTE functionals act as task-aware probes of distribution shift, offering a bridge between generative and discriminative views of uncertainty in deep learning.

2 Related Work

The dataset-shift literature distinguishes changes in P_X, changes in P_Y, and changes in $P(Y \mid X)$, and studies both detection and adaptation under non-stationarity (Quiñonero-Candela et al., 2008; Gama et al., 2014; Lu et al., 2018). In modern OOD detection, the dominant focus is on *detecting* atypical test inputs and quantifying separation between in- and out-distribution samples, with a growing emphasis on unified taxonomies and evaluation protocols (Yang et al., 2024; Rabanser et al., 2019). Our work complements this direction by emphasising *task relevance*: the goal is not to detect all deviations from P_X, but to detect distribution shifts that influence the deployed model.

Data-centric methods use density or score models to assess novelty under P_X, yet the likelihood can be misaligned with semantic novelty in high-dimensional image data (Nalisnick et al., 2019; Ren et al., 2019). Model-centric methods instead use confidence and calibration heuristics or representation distances, such as Maximum Softmax Probability (MSP)-style baselines, and methods that shape output energies or logits to improve separability (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020).

Recent post-hoc and hybrid methods such as ReAct (Sun et al., 2021), ViM (Wang et al., 2022), ASH (Djurisic et al., 2023), PRO (Chen et al., 2025), and SCOPED (Barkley et al., 2025) improve OOD separability through feature shaping, logit correction, perturbation sensitivity, or score-based geometry. Our Stein construction is a hybrid: it uses a learned score model to represent data geometry while coupling it to the input-sensitivity of the model, yielding a task-aware signal.

Many applications require localisation: identifying which features or pixels contribute to abnormality. In computer vision, pixel-level anomaly localisation has been studied extensively (Bergmann et al., 2020; Roth et al., 2022). Separately, feature-level shift localisation aims to identify which coordinates have shifted between train and test distributions (Kulinski et al., 2020). Our TASTE residuals provide a task-aware route to such localisation.

Stein operators have been used to construct discrepancy measures and goodness-of-fit tests such as kernelised Stein discrepancies, and computable measures of sample quality, see the surveys Anastasiou et al. (2023) and Liu et al. (2026). In contrast to the typical setting which obtains a measure of discrepancy by optimising the TASTE functional (2) over a set of test functions, TASTE fixes the test function, yielding a model-centric diagnostic tool.

Score matching and diffusion-based generative modelling are scalable ways to estimate $\nabla \log p(x)$ for high-dimensional data, and can be interpreted geometrically as learning vector fields aligned with local density gradients (Song et al., 2021). Thus, score models are a natural component in operator-based diagnostics that require access to the data geometry, including our Stein-based construction.

3 Stein Operators and Data Geometry

Here we provide the minimal Stein machinery needed for our goal to connect a ***fixed*** model f to the geometry of a reference data distribution p via a differential operator whose expectation vanishes under p; see Anastasiou et al. (2023) for details. Our setting is as follows. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a scalar test function (or predictor); for a vector-valued predictor $f : \mathbb{R}^d \to \mathbb{R}^m$ we apply all definitions componentwise. We assume that $p : \mathbb{R}^d \to (0, \infty)$ is a continuously differentiable probability density with score function $s_p(x) := \nabla \log p(x) \in \mathbb{R}^d$. The score function requires the density p to be available only up to a normalising constant. Expectations under p are denoted $\mathbb{E}_p[\cdot]$.

A so-called *Stein operator* is an operator $\mathcal{T}_p$ such that the *Stein identity*

$$\mathbb{E}_p[\mathcal{L}_p f(X)] = 0 \tag{3}$$

holds for any sufficiently regular f. Eq. (3) makes Stein operators very useful: they produce quantities that equal zero for in-distribution data (in expectation); nonzero values indicate distribution shift. Here we focus on the Langevin Stein operator from (1). A key property of $\mathcal{L}_p$ is its divergence form:

$$\mathcal{L}_p f(x) = \frac{1}{p(x)} \nabla \cdot \big(p(x)\, \nabla f(x)\big). \tag{4}$$

Under mild regularity and boundary decay conditions, (4) implies the Stein identity (3), see App. A.1. The divergence view (4) also provides a geometric interpretation. The vector field $p(x)\nabla f(x)$ is a

density-weighted *sensitivity flow* induced by the model. The Stein identity (3) states that, under the reference density p, this field cannot generate net flux in expectation. For us, ∇f captures local model sensitivity, while p encodes the data geometry; $\mathcal{L}_p f$ combines both into a single scalar diagnostic.

4 TASTE: TASK-AWARE STEIN RESIDUALS

We now *fix* f to be a pre-trained model and ask how Stein operators behave when the test distribution differs from the training distribution. Let q be another smooth density on $\mathbb{R}^d$, absolutely continuous with respect to p, and let $S_f(p,q) = \mathbb{E}_q[\mathcal{L}_p f(X)]$ be the TASTE functional of the fixed predictor f at (p,q) given in (2). When $q = p$, (3) implies $S_f(p,p) = 0$. When $q \neq p$, the functional measures how the test distribution violates the Stein identity that holds under p. The central identity

$$S_f(p,q) = -\,\mathbb{E}_q\Big[\nabla f(X)^\top \nabla \log \frac{q(X)}{p(X)}\Big] \tag{5}$$

expresses $S_f(p,q)$ in terms of the density ratio q/p; for assumptions and the proof see App. A.2. Defining the *shift score field* $u_{p\to q}(x) := \nabla \log \frac{q(x)}{p(x)}$, (5) becomes $S_f(p,q) = -\,\langle \nabla f,\, u_{p\to q}\rangle_{L^2(q)}$. This is the key "task-aware" mechanism: for a fixed predictor f, the Stein functional measures only the component of the distribution shift that aligns with the sensitivity field ∇f of the predictor. Shifts that are large in density terms but lie mostly in directions orthogonal to ∇f can be invisible to S_f; conversely, comparatively small shifts can trigger a strong response if they move mass along directions where f is sensitive.

Example 4.1. *As an example we perturb p via exponential tilts $q_\varepsilon(x) = \frac{1}{Z_\varepsilon}\, p(x)\, e^{\varepsilon h(x)}$, where $h : \mathbb{R}^d \to \mathbb{R}$ is a shift potential and ε controls the shift strength. Under regularity assumptions on h, we have $\nabla \log \frac{q_\varepsilon(x)}{p(x)} = \varepsilon\, \nabla h(x)$, and substituting into* (5) *yields*

$$S_f(p, q_\varepsilon) = -\,\varepsilon\, \mathbb{E}_{q_\varepsilon}\left[\nabla f(X)^\top \nabla h(X)\right].$$

Moreover, expanding expectations under q_ε around p gives the first-order response

$$S_f(p, q_\varepsilon) = \varepsilon\, \mathrm{Cov}_p\big(\mathcal{L}_p f(X),\, h(X)\big) + O(\varepsilon^2), \tag{6}$$

see App. A.3 for details. Thus, for small smooth shifts, the TASTE functional responds linearly in ε, with slope given by a covariance under the training distribution. In practice, this regime may provide a useful diagnostic lens even when real shifts are not exactly tilts.

4.1 ADJUSTED TASTE: CORRECTING FOR USING AN APPROXIMATE SCORE MODEL

Computing $\mathcal{L}_p f$ requires $s_p = \nabla \log p$, yet in applications p is unknown. Substituting a learned model $\tilde{p}$ with score $\tilde{s} = \nabla \log \tilde{p}$ yields the approximate operator $\mathcal{L}_{\tilde{p}} f(x) = \Delta f(x) + \tilde{s}(x)^\top \nabla f(x)$. A useful consequence of the projection identity (5) is that the error induced by $\tilde{s}$ is *directional*: it depends on how score-model error aligns with ∇f and the shift direction $l = q/p$. Taking $g(x) := (\tilde{s}(x) - s_p(x))^\top \nabla f(x)$, under mild regularity assumptions, one obtains the decomposition

$$\mathbb{E}_q[\mathcal{L}_{\tilde{p}} f] = \mathbb{E}_p[\mathcal{L}_{\tilde{p}} f] + S_f(p,q) + \langle g,\, l - 1\rangle_{L^2(p)}, \tag{7}$$

see App. A.4. This motivates the *adjusted TASTE functional*

$$\tilde{S}_f(p,q) := \mathbb{E}_q[\mathcal{L}_{\tilde{p}} f] - \mathbb{E}_p[\mathcal{L}_{\tilde{p}} f], \tag{8}$$

where $\mathbb{E}_p[\mathcal{L}_{\tilde{p}} f]$ can be estimated on held-out training data.

The adjusted TASTE functional $\tilde{S}_f(p,q)$ gives rise to several observations. First, it guarantees *zero population response under no shift*: when $q = p$, the density ratio satisfies $l \equiv 1$ and both $S_f(p,p) = 0$ and $\langle g, r - 1\rangle_{L^2(p)} = 0$, so the corrected functional vanishes even if the score model $\tilde{s}_p$ is imperfect. Second, the ideal term $S_f(p,q)$ retains the task-aware projection interpretation in (5). Score-model error contributes the additional bias $\langle g, l - 1\rangle_{L^2(p)}$, determined by the alignment of the density-ratio perturbation $l - 1$ with the directional score error $g = (\tilde{s} - s_p)^\top \nabla f$. Orthogonality to g eliminates this bias. Third, the influence of score-model mismatch is quantitatively controlled. We show in App. A.4 that the bias term satisfies

$$\big|\langle g,\, l - 1\rangle_{L^2(p)}\big| \;\le\; \sqrt{J(p\|\tilde{p})}\; \|\nabla f\|_{L^4(p)}\; \|l - 1\|_{L^4(p)}, \tag{9}$$

where $J(p\|\tilde{p})$ is the Fisher divergence minimized by denoising score matching (Song et al., 2021).

Diffusion score approximation. We evaluate the score model at $x_t = \alpha_t x + \sigma_t \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$, with timestep t, while evaluating predictor derivatives at the clean input x. For $t > 0$, this construction does not generally preserve the Stein identity for the clean-data distribution. We interpret $\widehat{s}_t(x; \varepsilon) = s_\theta(x_t, t)$ as a stochastic surrogate for $\nabla \log p(x)$. Its discrepancy, which can introduce both bias and variance, is treated as a score-estimation error, subject to the assumptions of our approximation analysis.

On access to the test distribution. We do not assume access to an explicit density for the test distribution q. The distribution q enters only implicitly through test samples, and all expectations with respect to q are estimated empirically by sample averages over the observed test set. Analysis regarding the choice of t is deferred to Appendix F.3.

Per-sample residuals. The adjusted TASTE functional (8) can be used at different granularity levels. Given a test input $x \sim q$, we define the adjusted *TASTE residual*

$$r_f(x) := \mathcal{L}_{\tilde{p}} f(x) - D_f,$$

where $D_f = \mathbb{E}_p[\mathcal{L}_{\tilde{p}} f(X)]$, which in practice is estimated from training data. Here are three uses of the adjusted TASTE residuals:

- Ranking: $r_f(x)$ is used for OOD ranking.
- OOD testing: Using held-out training data, let τ_α be the empirical $(1-\alpha)$-quantile of $r_f(x)$ under $X \sim p$. At test time, an input x is declared out-of-distribution if $|r_f(x)| > \tau_\alpha$. This model-agnostic decision rule approximately controls the false positive rate under the training distribution, while remaining sensitive to shifts that affect the predictor's behaviour.
- Aggregation: The adjusted TASTE residuals can be averaged over the test set or used in a batch statistic to produce a single test-set operator value $\frac{1}{|\mathcal{X}_{\text{test}}|} \sum_{x \in \mathcal{X}_{\text{test}}} r_f(x)$ for global shift detection.

Per-dimension residuals. For finer granularity, writing derivatives in $\mathcal{L}_p$ component-wise as $\mathcal{L}_p f(x) = \sum_{i=1}^d \Big(\partial_{ii} f(x) + \partial_i \log p(x)\, \partial_i f(x) \Big)$ motivates the per-dimension *TASTE residuals*

$$r_{f,i}(x) := \partial_{ii} f(x) + \partial_i \log p(x)\, \partial_i f(x), \tag{$\star$}$$

so that $r_f(x) = \sum_i r_{f,i}(x)$. Importantly, each coordinate-wise $r_{f,i}(x)$ is itself a valid Langevin Stein operator applied to the one-dimensional test function $x \mapsto \partial_i f(x)$. Under the same regularity conditions as for the full operator, each component satisfies $\mathbb{E}_p[r_{f,i}(X)] = 0$, and admits an analogous projection formula under a shifted distribution. As a result, all theoretical properties established for the aggregated TASTE functional— task-aware sensitivity to shift and robustness via baseline correction—extend directly to the per-dimension setting.

In imaging applications, the index i corresponds to a pixel (or pixel-channel), yielding an interpretable anomaly map $i \mapsto r_{f,i}(x)$. In this sense, TASTE residuals provide a *model-agnostic and interpretable* mechanism: they apply to any pre-trained predictor for which gradients are available, and their per-input[1] decomposition highlights which input coordinates contribute most strongly to the task-aware shift signal ($\star$). This setting supports use cases such as pixel-level anomaly segmentation and localisation of distributional shift within an image, which have been extensively studied in industrial inspection and medical imaging (Bergmann et al., 2020; Roth et al., 2022; Zenati et al., 2018).

For a fixed f_θ, the per-dimension decomposition computes $r_i(x) = \partial_{ii} f_\theta(x) + \hat{s}_{p,i}(x)\, \partial_i f_\theta(x)$ for each input dimension i, producing a spatial/feature map $\{r_i(x)\}_{i=1}^d$. If this map is used, the correction D_f must also be estimated as a per-dimension vector $D_f = \big(\mathbb{E}_p[r_1(X)], \ldots, \mathbb{E}_p[r_d(X)]\big)$ (in practice estimated by aggregating per-dimension raw residuals on the calibration set). This per-dimension mode enables fine-grained anomaly heatmaps but increases compute as Hutchinson approximation, which we employ for estimating the Laplacian term, cannot be used directly.

4.2 THE TASTE PIPELINE

Overall, the proposed TASTE pipeline is: (i) take any fixed predictor f_θ, (ii) take any score/density model $\tilde{p}$ for p (e.g. a diffusion/score model), (iii) compute $\mathcal{L}_{\tilde{p}} f_\theta(x)$ (exactly or approximately) over

[1]The "per-input" refers to a single coordinate of the gradient/Laplacian; each $r_{f,i}(x)$ requires the entire $x \in \mathbb{R}^n$.

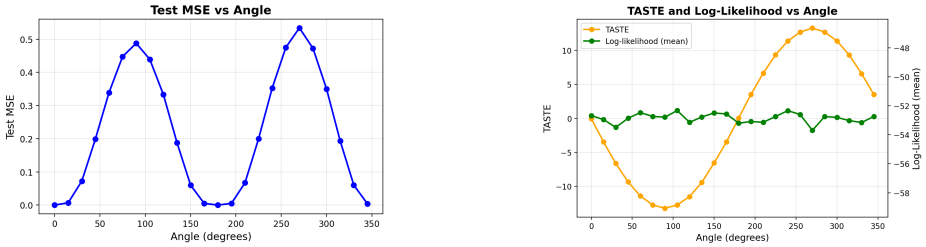


Figure 2: **Directional shift experiment.** Prediction error and the task-aware TASTE functional vary strongly with the direction of shift, peaking (in magnitude) when the shift aligns with the sensitive direction $(1, -1)$. In contrast, the density-based score remains nearly constant for all directions.

the test set, (iv) subtract a train-estimated baseline D_f, and (v) use the resulting per-sample or per-dimension residuals for task-aware OOD detection and diagnostics. Algorithm 1 in App. B gives the adjusted TASTE algorithm, with different options, in pseudocode.

Probe selection for multiclass predictors. Unless otherwise stated, our multiclass experiments use $f_c(x) = \mathrm{softmax}(z(x))_c$ as the scalar probe, with $c = \arg\max_j z_j(x)$ held fixed during differentiation. Probe selection, the use of the Laplacian, and the aggregation of residuals are implementation choices that determine the resulting detector. Appendix C.1 specifies the scoring protocol for each experiment. For ImageNet-1K, we additionally evaluate a first-order classwise variant that aggregates extreme signed responses across classes; its construction is given in Appendix C.2.

Computational efficiency. Computing the Laplacian term $\Delta f_\theta(x) = \mathrm{tr}(\nabla^2 f_\theta(x))$ exactly is computationally expensive in high dimensions. In practice we therefore rely on stochastic trace estimators, most notably Hutchinson's unbiased estimator and its modern variance-reduced variants (Hutchinson, 1989; Meyer et al., 2021).

5 Experimental Results

We present concise empirical evidence that (i) TASTE tracks task-relevant shifts, (ii) remains stable under harmless perturbations, and (iii) yields effective per-sample OOD signals in mixed test sets. Full implementation details and additional plots are deferred to App. C. Ablations on operator choice and sensitivity to quality of the learned score model and to the diffusion time step are in App. F. Research code is provided in the submitted additional zip file.

5.1 Controlled 2D shift: directional sensitivity

In this controlled two-dimensional setting, the direction of the distribution shift is varied independently of its magnitude. The training distribution is 2D standard Gaussian $p(x) = \mathcal{N}(0, I_2)$ and the test distribution is obtained by translating p by a fixed magnitude $\varepsilon = 10$ along a rotated version of the base direction $u = (1, 1)^\top/\sqrt{2}$, producing $q_\varphi = \mathcal{N}(\varepsilon R_\varphi u, I_2)$, where R_φ denotes the rotation matrix by the angle φ. The prediction task is $y = x_2 - x_1$, which is invariant to translations along $(1, 1)$ and maximally sensitive along $(1, -1)$. A small ReLU network is trained under p, and we evaluate prediction error, the TASTE functional $S_f(p, q_\varphi)$, and the log-likelihood under p as φ varies.

Figure 2 shows how the Stein signal tracks task degradation: its extrema align with those of the prediction error, while remaining invariant to shifts that leave the task unchanged. This confirms that the TASTE functional responds selectively to task-relevant components of distribution shift.

5.2 MNIST: invariance to translations, sensitivity to rotations

We next evaluate the method on MNIST (LeCun et al., 2002), a medium scale data set, under geometric perturbations. Convolutional networks are known to exhibit approximate translation invariance while remaining sensitive to rotations; we therefore compare translations (largely benign) and rotations (harmful) as test-time shifts (Hendrycks & Dietterich, 2019; Bronstein et al., 2021).

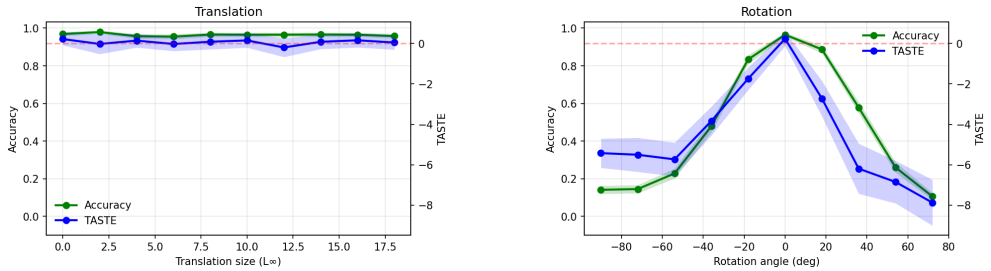


Figure 3: **MNIST under geometric perturbations.** Classification accuracy and adjusted TASTE score. Translations (left) leave both accuracy and adjusted TASTE residual unchanged, while rotations (right) induce a monotonic increase in the Stein signal aligned with performance degradation. Shaded regions show ± 2 standard deviations.

An AlexNet-style CNN (Krizhevsky et al., 2012) is trained on MNIST. So that translations do not remove semantic content, all images are zero-padded from 28×28 to 64×64 prior to training and evaluation. A denoising score-matching model provides an estimate $\hat{s}_p(x) \approx \nabla \log p(x)$. We compute the adjusted Langevin Stein residual $r_f(x)$ using Algorithm 1. Test sets are constructed using translations of increasing ℓ^∞ magnitude and rotations of increasing angle.

Figure 3 shows that the Stein residual remains near zero under translations but increases sharply under rotations, closely mirroring the drop in accuracy. Likelihood-based scores do not distinguish between these cases, highlighting the task-awareness of the proposed signal.

Consistent with the per-dimension decomposition introduced in the methodology, the Stein residual can also be evaluated coordinate-wise on image inputs. On MNIST, this yields per-pixel residual maps that localise regions most incompatible with the training distribution as processed by the classifier, see App. E; evaluations using a mixed MNIST / Fashion-MNIST OOD setup are found in App. G.

5.3 Benchmark evaluation on standard OOD data sets

We evaluate the proposed TASTE residual against widely used OOD detection baselines on CIFAR-10 (Krizhevsky, 2009) and ImageNet-1K (Russakovsky et al., 2015) covering heterogeneous shift regimes including adversarial perturbations, synthetic corruption and perturbation benchmarks, and other standard OOD datasets (Krizhevsky, 2009; Yang et al., 2024). For CIFAR-10, the classifier is a ResNet-18 (He et al., 2016); for ImageNet-1K, we use ResNet-50. In both cases, the score model is a diffusion-based estimator of $\nabla \log p(x)$ (Ho et al., 2020). Implementation and configuration details are deferred to App. C.

Benchmarks. For CIFAR-10, we group evaluation datasets into four categories: (i) *adversarial perturbations:* FGSM, PGD, and AutoAttack (Goodfellow et al., 2014; Madry et al., 2018; Croce & Hein, 2020); (ii) *corruption benchmarks:* CIFAR-10-C, averaged over all corruption types; (iii) *perturbation benchmarks:* CIFAR-10-P, averaged over all perturbation types (Hendrycks & Dietterich, 2019); and (iv) *classical OOD datasets:* SVHN, LSUN, iSUN, Textures, and Places365 (Netzer et al., 2011; Yu et al., 2016; Xu et al., 2015; Cimpoi et al., 2014; Zhou et al., 2017). For ImageNet-1K, we evaluate adversarial perturbations and the same classical OOD benchmark suite, upscaled to ImageNet resolution. All results are reported using AUROC and false positive rate at 95% true positive rate (FPR95), averaged within each category, with "overall" reporting the average across all used datasets.

Baselines. We compare against a diverse set of representative OOD detection methods spanning confidence-based, distance-based, energy-based, post-hoc, and score-curvature approaches. Implementation details and exact configurations are provided in App. C.

Results and benchmark analysis. Tables 1 and 2 report OOD detection performance on CIFAR-10 and ImageNet-1K, respectively. Consistent with prior work, no single method dominates across all regimes (Tajwar et al., 2021). On CIFAR-10, TASTE achieves the highest overall AUROC and lowest overall FPR95 under the reported aggregation. Its strengths include the highest adversarial

Table 1: OOD detection performance on CIFAR-10 across heterogeneous shift regimes.

	Adversarial		CIFAR-10-C		CIFAR-10-P		OOD-benchmarks		Overall	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP	0.3575	0.9656	0.6124	0.9109	0.5978	0.9210	0.8032	0.7878	0.5751 ± 0.0080	0.8521 ± 0.0161
ODIN	0.4895	0.9436	0.5499	0.9013	0.5478	0.9222	0.7644	0.7756	0.5776 ± 0.0084	0.9486 ± 0.0070
Mahalanobis	0.5819	0.8388	0.5880	0.9181	0.5620	0.9435	0.8188	0.7195	0.6060 ± 0.0084	0.8664 ± 0.0116
Energy	0.3550	0.9681	**0.6256**	0.8900	0.6110	0.9113	**0.8520**	0.6683	0.5871 ± 0.0081	0.8421 ± 0.0125
kNN+	0.4645	0.9591	0.6151	0.8977	0.5956	0.9223	0.8288	0.7023	0.6007 ± 0.0083	0.8526 ± 0.0123
ReAct	0.5978	0.8637	0.5403	0.9810	0.5184	0.9807	0.7078	0.9996	0.5678 ± 0.0087	0.8993 ± 0.0091
ViM	0.5495	**0.8372**	0.6045	0.9169	0.5716	0.9451	0.8174	0.7918	0.6079 ± 0.0084	0.8552 ± 0.0111
ASH	0.3614	0.9691	0.6245	0.8911	**0.6121**	0.9108	0.8498	0.6701	0.5882 ± 0.0081	0.8454 ± 0.0126
PRO	0.3478	0.9793	0.6043	0.9116	0.5838	0.9353	0.7159	0.9203	0.5560 ± 0.0082	0.9334 ± 0.0147
TASTE (ours)	**0.6144**	0.8650	0.6193	**0.8691**	0.5954	**0.8611**	0.7647	**0.5677**	**0.6301** ± 0.0082	**0.8207** ± 0.0127

OOD detection on CIFAR-10 across heterogeneous shift regimes. Results are group averages; overall entries show ±2 s.d. Bold and underlined values indicate the best and second-best results. Per-dataset results appear in Appendix D.

AUROC and the lowest FPR95 on standard OOD benchmarks. On CIFAR-10-C and CIFAR-10-P, performance remains competitive, although other methods achieve higher AUROC.

ImageNet-1K presents a substantially larger classifier output space, with 1,000 classes compared with 10 for CIFAR-10. Hence approximating the Laplacian to a reasonable level becomes infeasible; TASTE has to be run based on the score-part only, using a single predicted-class probability. However, this procedure may miss changes in other class probabilities. This motivates $\text{TASTE}_{\text{ext}}$, which computes classwise directional responses using a fast implementation and sums the largest and smallest 5% by signed value (Appendix C.2). Compared with the predicted-class configuration, $\text{TASTE}_{\text{ext}}$ improves both AUROC and FPR95 on adversarial and standard OOD benchmarks. It achieves the highest overall AUROC and lowest overall FPR95 among the compared methods under the reported aggregation. Performance nevertheless varies by regime: ViM achieves the best standard OOD results in both metrics, while SCOPED achieves the highest adversarial AUROC. $\text{TASTE}_{\text{ext}}$ achieves the lowest adversarial FPR95 and second-highest adversarial AUROC. These findings support examining multiple classwise responses on ImageNet-1K, although they do not isolate output dimensionality as the cause of the improvement.

Runtime considerations. Table 2 also reports wall-clock runtime per sample. While TASTE is slower than simple confidence-based methods such as MSP or Energy, it remains competitive with modern geometry- and diffusion-based approaches. Importantly, unlike many recent diffusion-based OOD detectors, TASTE requires only a single forward pass through the diffusion model to evaluate the score field, without iterative purification, sampling, or reconstruction procedures. This makes the method comparatively lightweight despite leveraging generative information about the data distribution.

5.4 Per-pixel anomaly detection on MVTec AD

We evaluate per-pixel TASTE residuals on the MVTec Anomaly Detection dataset (Bergmann et al., 2019) to assess spatial localisation on high-resolution real-world imagery. MVTec AD provides pixel-level ground-truth annotations supporting both quantitative metrics (pixel-level AUROC, average precision, AUPRO) and visual inspection. For each test image, we compute per-pixel Stein heatmaps from the coordinate-wise operator and compare them to ground-truth defect masks. Across categories, the heatmaps align spatially with defect regions. Representative examples are shown in App. E.

Quantitative localisation results. Averaged over all 15 MVTec categories, TASTE achieves an image-level AUROC of 0.6527 and average precision of 0.817, indicating consistent image-level separability in a zero-shot regime. At the pixel level, performance reaches an AUROC of 0.6977 and an AUPRO of 0.688. Performance varies across category types. Texture categories exhibit substantially stronger localisation (pixel-level AUROC 0.8682, AUPRO 0.8745), while object categories are more challenging (AUROC 0.6124, AUPRO 0.5947).

Overall, these results demonstrate that TASTE and produce meaningful, spatially coherent localisation signals in a *zero-shot* setting: neither the classifier nor the score model is trained or fine-tuned on

Table 2: OOD detection performance on ImageNet-1K.

Method	Adversarial AUROC↑	Adversarial FPR95↓	OOD-benchmarks AUROC↑	OOD-benchmarks FPR95↓	Overall AUROC↑	Overall FPR95↓	Runtime ms/sample
MSP	0.2543	0.9551	0.8661	0.4546	0.4582 ± 0.0019	0.7883 ± 0.0039	5.95
ODIN	0.6475	0.8212	0.8356	0.4893	0.7102 ± 0.0026	0.7106 ± 0.0050	19.78
Mahalanobis	0.6785	0.7452	0.9264	<u>0.1749</u>	<u>0.7611</u> ± 0.0024	<u>0.5551</u> ± 0.0048	137.69
Energy	0.2242	0.9607	0.8971	0.3634	0.4485 ± 0.0018	0.7616 ± 0.0043	6.04
kNN+	0.6099	0.8089	0.9175	0.2501	0.7125 ± 0.0025	0.6226 ± 0.0044	95.73
PRO	0.2706	0.9519	0.3581	0.9240	0.2997 ± 0.0025	0.9426 ± 0.0024	71.45
ReAct	0.3088	0.9485	<u>0.9283</u>	0.2570	0.5153 ± 0.0020	0.7180 ± 0.0036	7.92
ASH	0.2242	0.9607	0.8979	0.3601	0.4488 ± 0.0017	0.7605 ± 0.0045	7.04
ViM	0.3763	0.9198	**0.9428**	**0.1632**	0.5651 ± 0.0019	0.6676 ± 0.0038	105.30
SCOPED	**0.7872**	<u>0.6367</u>	0.4794	0.7309	0.6846 ± 0.0026	0.6681 ± 0.0080	157.19
TASTE (ours)	0.7420	0.7122	0.6820	0.5866	0.7220 ± 0.0024	0.6703 ± 0.0057	122.32
TASTE$_{ext}$ (ours)	<u>0.7741</u>	**0.5701**	0.8485	0.4532	**0.7989** ± 0.0037	**0.5311** ± 0.0064	123.83

Results are averaged within each category. Adversarial attacks include FGSM, PGD, and AutoAttack. OOD benchmarks include SVHN, LSUN, iSUN, Textures, and Places365, all upscaled to match ImageNet-1K resolution. Runtime reports wall-clock time per sample. For readability, we only report ± 2 s.d., as approximate confidence intervals, for the "Overall" column. Best results are in bold; second-best are underlined.

MVTec AD data. While not designed to match state-of-the-art industrial detectors, the method provides a principled zero-shot localisation mechanism whose behaviour is transparent and consistent with the theoretical analysis.

6 Discussion and Conclusion

We introduced TASTE, a task-aware OOD detection framework that applies Stein operators to fixed, deployed predictors. By coupling model input sensitivity with a learned score of the training distribution, TASTE produces OOD signals that align with task degradation and admit interpretable per-sample and per-input decompositions. The method achieves competitive performance while providing diagnostic capabilities that are unavailable to purely confidence-based approaches. Unlike methods that flag distributional deviation alone, TASTE responds most strongly to shifts that interact with the predictor. This makes it well suited to post-deployment monitoring, where benign covariate changes must be distinguished from harmful task-level perturbations.

Choice of Stein operator. Different Stein operators encode different notions of compatibility between a predictor and a data distribution. The Langevin operator captures curvature and drift alignment through second-order information, while first-order variants emphasise gradient–score alignment and are computationally lighter. The choice of operator therefore determines which interaction between the predictor and the input distribution is being probed: curvature, gradient alignment, or score-weighted smoothness, echoing broader geometric perspectives on representation and invariance in deep learning (Bronstein et al., 2021).

Reversing the perspective. Rather than starting from a predefined Stein operator, one can instead specify the desired interaction between the shift score $\nabla \log(q/p)$ and a model quantity such as its gradient, curvature, or feature map, and then construct an operator that targets this interaction. This "design-by-intent" perspective suggests a route toward task-aware operators tailored to particular vulnerabilities, architectures, or modalities (Mijoule et al., 2023).

Limitations. TASTE's performance depends on score quality and probe selection, yet our analysis provides no simple practical test of whether a given score model is sufficiently accurate. Centring does not remove shift-dependent bias, and our ablations cover only limited score-model mismatches. Population guarantees do not ensure reliable pointwise detection. Differential operators add overhead relative to confidence-based baselines, potentially limiting latency-sensitive applications. Although required for the Langevin Stein identity, the Laplacian offers setting-dependent empirical benefits, while stochastic estimation adds variance. Derivative stability warrants further evaluation. Finally, the gains from TASTE$_{\text{ext}}$ do not establish output dimensionality as their cause.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and editing, discussion of mathematical arguments, visualisation, implementation of standard methods, and experimental workflows. The authors retain full responsibility for the methodology, mathematical claims, experimental results, and final content of this work.

ETHICS STATEMENT

This work uses existing public benchmark datasets and pretrained models. TASTE may support applications such as material inspection and detection of harmful distribution shifts. However, false positives may trigger unnecessary interventions, while false negatives may leave harmful shifts undetected. Deployment therefore requires application-specific validation, calibration, and careful integration with downstream decisions, particularly in safety-critical settings. Detection performance should not be interpreted as a guarantee of model safety or reliability.

REPRODUCIBILITY STATEMENT

Theoretical assumptions and proofs are provided in Appendix A, and Algorithm 1 specifies the residual computation. Appendix C describes the datasets, models, and experimental procedures, including probe selection and aggregation conventions in Appendix C.1. Detailed benchmark results and ablations appear in Appendices D and F, respectively. We release research code to reproduce the results and explore the proposed constructions. Developing a maintained Python package remains future work.

REFERENCES

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., et al. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.

Barkley, B., Culbertson, P., and Fridovich-Keil, D. SCOPED: Score-curvature out-of-distribution proximity evaluator for diffusion. *arXiv preprint arXiv:2510.01456*, 2025.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4182–4191. IEEE, 2020.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Chen, W., Yeh, R. A., Mou, S., and Gu, Y. Leveraging perturbation robustness to enhance out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4724–4733, 2025.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216. PMLR, 2020.

Djurisic, A., Bozanic, N., Ashok, A., and Liu, R. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37, March 2014. ISSN 0360-0300. doi: 10.1145/2523813.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hutchinson, M. A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics - Simulation and Computation*, 18(3):1059–1076, January 1989. ISSN 0361-0918. doi: 10.1080/03610918908812806.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kirchheim, K., Filax, M., and Ortmeier, F. Pytorch-OOD: A library for out-of-distribution detection based on pytorch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4351–4360, 2022.

Krizhevsky, A. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto*, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Kulinski, S., Bagchi, S., and Inouye, D. I. Feature shift detection: Localizing which features have shifted via conditional distribution tests. *Advances in Neural Information Processing Systems*, 33: 19523–19533, 2020.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31, 2018.

Ley, C., Reinert, G., and Swan, Y. Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Liu, Q., Mackey, L., and Oates, C. Probabilistic inference and learning with Stein's method. *arXiv preprint arXiv:2603.07467*, 2026.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Meyer, R. A., Musco, C., Musco, C., and Woodruff, D. P. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pp. 142–155. SIAM, 2021.

Mijoule, G., Raič, M., Reinert, G., and Swan, Y. Stein's density method for multivariate continuous distributions. *Electronic Journal of Probability*, 28:1–40, 2023.

Nalisnick, E., Matsukawa, A., Teh, Y., Gorur, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

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*, 2011.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N. D., Jordan, M. I., and Dietterich, T. (eds.). *Dataset Shift in Machine Learning*. Neural Information Processing Series. MIT Press, Cambridge, MA, USA, December 2008.

Rabanser, S., Günnemann, S., and Lipton, Z. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 32, 2019.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Li, F.-F. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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*, 2021.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability, volume 2: Probability Theory*, volume 6, pp. 583–603. University of California Press, 1972.

Sun, Y., Guo, C., and Li, Y. ReAct: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Tajwar, F., Kumar, A., Xie, S. M., and Liang, P. No true state-of-the-art? OOD detection methods are inconsistent across datasets. *arXiv preprint arXiv:2109.05554*, 2021.

Wang, H., Li, Z., Feng, L., and Zhang, W. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022.

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., and Xiao, J. TurkerGaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132:5635–5662, 2024.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2016.

Zenati, H., Romain, M., Foo, C.-S., Lecouat, B., and Chandrasekhar, V. Adversarially learned anomaly detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 727–736. IEEE, 2018.

Zhou, B., Lapedriza, A., Torralba, A., and Oliva, A. Places: An image database for deep scene understanding. *Journal of Vision*, 17(10):296–296, 2017.

Appendix

Contents

A.1 Stein identity for the Langevin operator 15
A.2 Proof of the projection identity (5) 15
A.3 Exponential-tilt expansions and remainder control 16
A.4 Directional score-model error decomposition 18
A.5 Score-model error decomposition and bound 19
C.1 Experimental probes and scoring protocols 20
C.2 First-order classwise aggregation for ImageNet-1K 21
C.3 Controlled 2D directional shift experiment 22
C.4 MNIST geometric perturbation experiments 22
C.5 CIFAR-10 benchmark evaluation 23
C.6 ImageNet-1K experiments 23
C.7 Compute resources and runtime. 23
C.8 Baselines and hyperparameters 24
C.9 MVTec AD localisation experiments 24
C.10 On the Laplacian term for ReLU networks. 24
C.11 Comment on licences. 25
F.1 Sensitivity to score quality. 32
F.2 Ablation: Operator decomposition 32
F.3 Sensitivity to diffusion time step 33
H.1 Directional blind spots: the 2D Gaussian illustration 37
H.2 The L^2 norm of the first-order Stein operator (no blind spots, but not a Stein operator) 38

A CORE METHODOLOGY: ASSUMPTIONS, PROOFS AND ADDITIONAL STEIN IDENTITIES

For convenience, this section of the Appendix is structured as follows.

- Appendix A.1: Assumptions and proof of the Stein identity (3) under p;
- Appendix A.2: Assumptions and proof of the projection identity (5), formulated as Proposition A.4 and used to interpret $S_f(p, q)$;
- Appendix A.3: Assumptions and proof of the controlled small-shift expansion (6) for exponential tilts;
- Appendix A.4: Assumptions and proof of the score-model error decomposition (7);
- Appendix A.5: Assumptions and proof for the score-model error stability (9).

A.1 STEIN IDENTITY FOR THE LANGEVIN OPERATOR

Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable. Recall the Langevin Stein operator acting on f by

$$\mathcal{L}_p f(x) = \Delta f(x) + (\nabla \log p(x))^\top \nabla f(x) \;=\; \frac{1}{p(x)} \nabla \cdot \big(p(x)\nabla f(x)\big).$$

Definition A.1 (Stein class for $\mathcal{L}_p$)**.** Let p be a continuously differentiable density on $\mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}$ belongs to the Stein class of p (for $\mathcal{L}_p$), denoted $f \in \mathcal{F}(p)$, if:

(S1) f is twice continuously differentiable and Δf, ∇f are locally integrable with respect to Lebesgue measure.

(S2) The vector field $p(x)\nabla f(x)$ is integrable and its flux over spheres vanishes:

$$\lim_{R\to\infty} \int_{\partial B_R} p(x)\nabla f(x) \cdot n(x)\, dS(x) = 0,$$

where $B_R \subset \mathbb{R}^d$ is the Euclidean ball of radius R in $\mathbb{R}^d$, and $n(x)$ is the outward unit normal.

(S3) $\mathcal{L}_p f$ is integrable under p.

Proposition A.2 (Stein identity)**.** *If p is continuously differentiable and $f \in \mathcal{F}(p)$, then*

$$\mathbb{E}_p[\mathcal{L}_p f(X)] = 0.$$

Proof. By (S3), $\mathbb{E}_p[\mathcal{L}_p f(X)]$ exists. Using the divergence form,

$$\mathbb{E}_p[\mathcal{L}_p f(X)] = \int_{\mathbb{R}^d} p(x)\mathcal{L}_p f(x)\, dx = \int_{\mathbb{R}^d} \nabla \cdot \big(p(x)\nabla f(x)\big)\, dx = \lim_{R\to\infty} \int_{B_R} \nabla \cdot \big(p(x)\nabla f(x)\big)\, dx,$$

where B_R is the ball of radius R, as above. By the divergence theorem,

$$\int_{B_R} \nabla \cdot \big(p\nabla f\big)\, dx = \int_{\partial B_R} p(x)\nabla f(x) \cdot n(x)\, dS(x). \tag{10}$$

Letting $R \to \infty$ and using (S2) yields $\int_{\mathbb{R}^d} \nabla \cdot (p\nabla f)\, dx = 0$, hence $\mathbb{E}_p[\mathcal{L}_p f] = 0$. □

A.2 PROOF OF THE PROJECTION IDENTITY (5)

We prove the projection identity

$$S_f(p, q) = -\,\mathbb{E}_q\Big[\nabla f(X)^\top \nabla \log \frac{q(X)}{p(X)}\Big]$$

under sufficient regularity conditions.

Assumption A.3 (Regularity for the projection identity)**.** Let p, q be continuously differentiable densities on $\mathbb{R}^d$ with q absolutely continuous w.r.t. p and let $l(x) = q(x)/p(x)$. Assume:

(P1) $f \in \mathcal{F}(p)$;

(P2) $\mathcal{L}_p f \in L^1(q)$;

(P3) l is differentiable and $p\nabla(f\,l)$ is integrable.

(P4) The boundary flux vanishes for the vector field $p\,l\,\nabla f$:

$$\lim_{R\to\infty} \int_{\partial B_R} p(x)\,l(x)\,\nabla f(x)\cdot n(x)\,dS(x) = 0.$$

(P5) The product $\nabla f^\top \nabla l$ is integrable under Lebesgue measure.

Next we give a full statement for (5), including its assumptions; for reference we formulate it as a proposition.

Proposition A.4. *[Projection identity] Let p, q and f satisfy* (P1)– (P5)*. Then*

$$S_f(p,q) = -\,\mathbb{E}_q\left[\nabla f(X)^\top \nabla \log \frac{q(X)}{p(X)}\right].$$

Proof. Let p, q and f be such that (P1) – (P4) are satisfied. From (P1), $\mathcal{L}_p f$ exists, and from (P2), $S_f(p,q)$ exists. Substituting $q = lp$ in the divergence form (4) gives

$$S_f(p,q) = \int_{\mathbb{R}^d} q(x)\,\mathcal{L}_p f(x)\,dx = \int_{\mathbb{R}^d} l(x)\,\nabla\cdot\big(p(x)\nabla f(x)\big)\,dx.$$

Integrating by parts, which is justified by (P3) and (P5), using the product rule in the divergence form gives

$$\int_{\mathbb{R}^d} l\,\nabla\cdot(p\nabla f)\,dx = \int_{\mathbb{R}^d} \nabla\cdot(l\,p\nabla f)\,dx - \int_{\mathbb{R}^d} p\,\nabla f^\top \nabla l\,dx.$$

By (P4) and (10), the first term on the right-hand side vanishes; hence

$$S_f(p,q) = -\int_{\mathbb{R}^d} p(x)\,\nabla f(x)^\top \nabla l(x)\,dx.$$

Now we use $\nabla l = l\,\nabla \log l$ to obtain

$$\begin{aligned} S_f(p,q) &= -\int_{\mathbb{R}^d} p(x)\,l(x)\,\nabla f(x)^\top \nabla \log l(x)\,dx \\ &= -\int_{\mathbb{R}^d} q(x)\,\nabla f(x)^\top \nabla \log \frac{q(x)}{p(x)}\,dx, \end{aligned}$$

which is the desired identity. □

A.3 Exponential-tilt expansions and remainder control

We derive the small-ε expansions used in (6). First, we formally state the result.

Proposition A.5. *Let p be a continuously differentiable probability density on $\mathbb{R}^d$, and let $q_\varepsilon(x)$ be its exponential tilt by a measurable function h*

$$q_\varepsilon(x) = \frac{1}{Z_\varepsilon}\,p(x)\,e^{\varepsilon h(x)}.$$

Assume the following:

(i) *There exists $\delta > 0$ such that $\mathbb{E}_p\big[e^{\gamma|h(X)|}\big] < \infty$ for all $|\gamma| < \delta$.*

(ii) *The Langevin Stein operator satisfies $\mathcal{L}_p f \in L^2(p)$.*

Then, for all positive ε such that $\varepsilon < \delta/4$, the following expansions hold:

$$S_f(p,q_\varepsilon) = \varepsilon\,\mathrm{Cov}_p\big(\mathcal{L}_p f(X),\,h(X)\big) + O(\varepsilon^2), \tag{11}$$

$$\mathrm{Var}_{q_\varepsilon}[\mathcal{L}_p f(X)] = \mathrm{Var}_p[\mathcal{L}_p f(X)] + \varepsilon\,\mathrm{Cov}_p\big((\mathcal{L}_p f(X))^2,\,h(X)\big) + O(\varepsilon^2). \tag{12}$$

Proof. Our goal is to expand $\mathbb{E}_{q_\varepsilon}[g(X)]$ and $\mathrm{Var}_{q_\varepsilon}[g(X)]$ for small ε. We proceed in several steps.

Expansion of $\mathbb{E}_{q_\varepsilon}[g(X)]$. Let $g(X) = \mathcal{L}_p f(X)$, which is square-integrable under p by assumption. By the definition of the exponential tilt,

$$\mathbb{E}_{q_\varepsilon}[g(X)] = \frac{\mathbb{E}_p[g(X)e^{\varepsilon h(X)}]}{\mathbb{E}_p[e^{\varepsilon h(X)}]}. \tag{13}$$

Under the assumptions stated in the proposition, we shall prove below that

$$\mathbb{E}_p[g(X)e^{\varepsilon h(X)}] = \mathbb{E}_p[g(X)] + \varepsilon \mathbb{E}_p[g(X)h(X)] + O(\varepsilon^2), \tag{14}$$

and similarly,

$$\mathbb{E}_p[e^{-\varepsilon h(X)}] = 1 - \varepsilon\, \mathbb{E}_p[h(X)] + O(\varepsilon^2). \tag{15}$$

For now we assume that these expansions hold, in order to preserve the flow of the argument. The detailed derivation is deferred to the end of this subsection. Multiplying the numerator and denominator expansions yields, with (13), that

$$\mathbb{E}_{q_\varepsilon}[g(X)] = \mathbb{E}_p[g(X)] + \varepsilon\Big(\mathbb{E}_p[g(X)h(X)] - \mathbb{E}_p[g(X)]\,\mathbb{E}_p[h(X)]\Big) + O(\varepsilon^2),$$

which is exactly

$$\mathbb{E}_{q_\varepsilon}[g(X)] = \mathbb{E}_p[g(X)] + \varepsilon\,\mathrm{Cov}_p(g(X), h(X)) + O(\varepsilon^2).$$

Applying the same argument with $g(X)^2$ in place of $g(X)$ yields

$$\mathbb{E}_{q_\varepsilon}[g(X)^2] = \mathbb{E}_p[g(X)^2] + \varepsilon\,\mathrm{Cov}_p(g(X)^2, h(X)) + O(\varepsilon^2).$$

Specialisation to the Stein quantity $g = \mathcal{L}_p f$. Plugging in $g = \mathcal{L}_p f$ we obtain the deterministic expansions

$$\mathbb{E}_{q_\varepsilon}[\mathcal{L}_p f(X)] = \varepsilon\,\mathrm{Cov}_p(\mathcal{L}_p f(X), h(X)) + O(\varepsilon^2),$$

and

$$\mathbb{E}_{q_\varepsilon}[(\mathcal{L}_p f(X))^2] = \mathbb{E}_p[(\mathcal{L}_p f(X))^2] + \varepsilon\,\mathrm{Cov}_p((\mathcal{L}_p f(X))^2, h(X)) + O(\varepsilon^2).$$

Since the mean is $O(\varepsilon)$, its square is $O(\varepsilon^2)$, hence

$$\mathrm{Var}_{q_\varepsilon}[\mathcal{L}_p f(X)] = \mathrm{Var}_p[\mathcal{L}_p f(X)] + \varepsilon\,\mathrm{Cov}_p((\mathcal{L}_p f(X))^2, h(X)) + O(\varepsilon^2).$$

Proof of the first-order expansion of $\mathbb{E}[ge^{\varepsilon h}]$.

It now remains to prove the expressions (14) and (15) used above. We begin with the Taylor expansion of the exponential function $f(y) = e^y$ around $y = 0$. For any real scalar $y \in \mathbb{R}$, we have:

$$e^y = 1 + y + R(y),$$

where the remainder term $R(y)$ is bounded by the Lagrange error bound:

$$|R(y)| \le \frac{y^2}{2} e^{|y|}. \tag{16}$$

For any fixed vector x, let $y = \varepsilon h(x)$. Substituting this into the expression above and multiplying by $g(x)$, we obtain:

$$\begin{aligned} g(x)e^{\varepsilon h(x)} &= g(x)\Big(1 + \varepsilon h(x) + R(\varepsilon h(x))\Big) \\ &= g(x) + \varepsilon g(x)h(x) + g(x)R(\varepsilon h(x)). \end{aligned}$$

Taking the expectations:

$$\mathbb{E}_p[g(x)e^{\varepsilon h(x)}] = \mathbb{E}_p[g(x)] + \varepsilon \mathbb{E}_p[g(x)h(x)] + \mathbb{E}_p[g(x)R(\varepsilon h(x))]. \tag{17}$$

We next show that the remainder term $E_{\text{err}} = \mathbb{E}_p[g(x)R(\varepsilon h(x))]$ satisfies $|E_{\text{err}}| \le C\varepsilon^2$ for some constant C. Using the bound from (16):

$$|E_{\text{err}}| \le \mathbb{E}_p\left[|g(x)| \cdot |R(\varepsilon h(x))|\right] \le \mathbb{E}_p\left[|g(x)| \frac{\varepsilon^2 h(x)^2}{2} e^{\varepsilon |h(x)|}\right].$$

Factoring out constants:

$$|E_{\text{err}}| \leq \frac{\varepsilon^2}{2}\mathbb{E}_p\left[|g(x)|h(x)^2 e^{\varepsilon|h(x)|}\right].$$

We now apply the Cauchy-Schwarz inequality to bound the expectation:

$$\mathbb{E}_p\left[|g(x)|\left(h(x)^2 e^{\varepsilon|h(x)|}\right)\right] \leq \|g\|_{L^2(p)} \cdot \|h^2 e^{\varepsilon|h|}\|_{L^2(p)}.$$

We note that $\|g\|_{L^2(p)}$ is finite by assumption. It remains to show that the second term is bounded. The norm is given by:

$$\|h^2 e^{\varepsilon|h|}\|_{L^2(p)} = \left(\mathbb{E}_p\left[h(x)^4 e^{2\varepsilon|h(x)|}\right]\right)^{1/2}.$$

By the exponential-integrability assumption, the moment generating function of the scalar random variable $|h(X)|$ exists in a neighbourhood of zero, which in particular implies the finiteness of all polynomial moments of h under p. Moreover, exponential integrability allows control of mixed polynomial–exponential moments. Fix $\varepsilon > 0$ sufficiently small, for instance $\varepsilon < \delta/4$. Applying the Cauchy–Schwarz inequality, we obtain

$$\mathbb{E}_p\big[h(X)^4 e^{2\varepsilon|h(X)|}\big] \leq \big(\mathbb{E}_p[h(X)^8]\big)^{1/2}\big(\mathbb{E}_p[e^{4\varepsilon|h(X)|}]\big)^{1/2}.$$

Since $\varepsilon < \delta/4$, the exponential-integrability assumption on h implies $\mathbb{E}_p[e^{4\varepsilon|h(X)|}] < \infty$, and exponential integrability also ensures $\mathbb{E}_p[h(X)^8] < \infty$. Consequently, the mixed moment $\mathbb{E}_p[h(X)^4 e^{2\varepsilon|h(X)|}]$ is finite for all $\varepsilon < \delta/4$, justifying the uniform remainder bound used in (14). Expansion (15) follows similarly by using the expansion $e^{-\epsilon h(x)} = 1 - \epsilon h(x) + \tilde{R}(\epsilon h(x))$.

For completeness, we state the explicit bound. Let $K = \|g\|_{L^2(p)} \cdot \|h^2 e^{\varepsilon|h|}\|_{L^2(p)} < \infty$. Then

$$|E_{\text{err}}| \leq \frac{K}{2}\varepsilon^2.$$

Thus, $E_{\text{err}} = O(\varepsilon^2)$, completing the proof. □

A.4 Directional score-model error decomposition

Here we provide the full derivation of (7) and the tilt-specialisation (6) used in the main text. Let $\tilde{p}$ be a learned density model with score $\tilde{s}(x) = \nabla \log \tilde{p}(x)$. Define

$$\mathcal{L}_{\tilde{p}} f(x) := \Delta f(x) + \tilde{s}(x)^\top \nabla f(x),$$
$$g(x) := (\tilde{s}(x) - s_p(x))^\top \nabla f(x)$$

and the likelihood ratio

$$l(x) := \frac{q(x)}{p(x)}.$$

Proposition A.6 (Directional decomposition)**.** *Assume that p is a continuously differentiable probability density, that q is absolutely continuous w.r.t. p, that $f \in \mathcal{F}(p) \cap \mathcal{F}(q)$, and that $\mathcal{L}_{\tilde{p}} f$ and g are integrable under both p and q. Then*

$$\mathbb{E}_q[\mathcal{L}_{\tilde{p}} f] = \mathbb{E}_p[\mathcal{L}_{\tilde{p}} f] + S_f(p, q) + \mathbb{E}_p[g(X)(l(X) - 1)].$$

Proof. We begin with the decomposition

$$\mathcal{L}_{\tilde{p}} f = \mathcal{L}_p f + (\tilde{s} - \nabla \log p)^\top \nabla f,$$

which is valid for all $f \in \mathcal{F}(p) \cap \mathcal{F}(q)$. Taking expectations under q and p and subtracting gives

$$\mathbb{E}_q[\mathcal{L}_{\tilde{p}} f] - \mathbb{E}_p[\mathcal{L}_{\tilde{p}} f] = \mathbb{E}_q[\mathcal{L}_p f] + \Big(\mathbb{E}_q - \mathbb{E}_p\Big)\big[(\tilde{s} - \nabla \log p)^\top \nabla f\big]. \tag{18}$$

Here we used the Stein identity, Proposition A.2.

The first term on the right-hand side of (18) is $S_f(p, q)$ by definition. For the second term, write $q(x) = l(x)p(x)$ with $l = q/p$, and $g(x) = (\tilde{s}(x) - \nabla \log p(x))^\top \nabla f(x)$, giving

$$\Big(\mathbb{E}_q - \mathbb{E}_p\Big)[g(X)] = \mathbb{E}_p[g(X)(l(X) - 1)].$$

Thus we obtain (7).

For the exponential-tilt case $q_\varepsilon(x) \propto p(x)e^{\varepsilon h(x)}$, the expansion follows directly from the change-of-measure expansion proved in Appendix A.3, combined with the fact that $r_\varepsilon(x) = 1 + \varepsilon\,(h(x) - \mathbb{E}_p[h]) + O(\varepsilon^2)$, which can be established using the same technique as in Appendix A.3. This yields

$$\mathbb{E}_p[g(X)(r_\varepsilon(X) - 1)] = \varepsilon\,\mathrm{Cov}_p(g, h) + O(\varepsilon^2).$$

□

A.5 Score-model error decomposition and bound

Proposition A.7 (Fisher-controlled stability to score-model error)**.** *Let p be the continuously differentiable training density on $\mathbb{R}^d$, let $\tilde{p}$ be a learned density with score $\tilde{s} = \nabla \log \tilde{p}$, and let q be a test density such that $q \ll p$ with density ratio $l(x) = q(x)/p(x)$. Let $s_p = \nabla \log p$ and define*

$$g(x) := \big(\tilde{s}(x) - s_p(x)\big)^\top \nabla f(x).$$

Assume $\tilde{s} - s_p \in L^2(p)$, $\nabla f \in L^4(p)$, and $l \in L^4(p)$. Then the score-model error term satisfies

$$\big|\langle g,\, l - 1\rangle_{L^2(p)}\big| \;\le\; \sqrt{J(p\|\tilde{p})}\; \|\nabla f\|_{L^4(p)}\; \|l - 1\|_{L^4(p)},$$

where

$$J(p\|\tilde{p}) := \mathbb{E}_p\big[\|\tilde{s}(X) - s_p(X)\|^2\big]$$

is the Fisher divergence of $\tilde{p}$ from p.

Proof. By definition,

$$\langle g,\, l - 1\rangle_{L^2(p)} = \mathbb{E}_p\big[(\tilde{s}(X) - s_p(X))^\top \nabla f(X)\,(l(X) - 1)\big]\,.$$

Applying Cauchy–Schwarz in $L^2(p)$ by grouping the score-model error separately gives

$$\big|\mathbb{E}_p[g(X)(l(X) - 1)]\big| \le \Big(\mathbb{E}_p\|\tilde{s}(X) - s_p(X)\|^2\Big)^{1/2}\Big(\mathbb{E}_p\big[\|\nabla f(X)\|^2(l(X) - 1)^2\big]\Big)^{1/2}.$$

The first factor equals $\sqrt{J(p\|\tilde{p})}$ by definition. For the second factor, we apply Hölder's inequality with conjugate exponents 2 and 2:

$$\mathbb{E}_p\big[\|\nabla f(X)\|^2(l(X) - 1)^2\big] \le \big(\mathbb{E}_p\|\nabla f(X)\|^4\big)^{1/2}\big(\mathbb{E}_p|l(X) - 1|^4\big)^{1/2}.$$

Taking square roots of both sides gives

$$\begin{aligned}\Big(\mathbb{E}_p\big[\|\nabla f(X)\|^2(l(X) - 1)^2\big]\Big)^{1/2} &\le \big(\mathbb{E}_p\|\nabla f(X)\|^4\big)^{1/4}\big(\mathbb{E}_p|l(X) - 1|^4\big)^{1/4} \\ &= \|\nabla f\|_{L^4(p)}\,\|l - 1\|_{L^4(p)}.\end{aligned}$$

Combining the bounds yields

$$\big|\mathbb{E}_p[g(X)(l(X) - 1)]\big| \le \sqrt{J(p\|\tilde{p})}\; \|\nabla f\|_{L^4(p)}\; \|l - 1\|_{L^4(p)},$$

as claimed. □

Remark. The same argument applies with any Hölder pair $p, q > 1$ satisfying $1/p + 1/q = 1$, yielding bounds involving higher-order $L^{2p}(p)$ and $L^{2q}(p)$ norms of ∇f and the density ratio $l - 1$, respectively.

B TASTE: Algorithm

Algorithm 1 gives pseudocode for obtaining TASTE residuals. It includes the option to estimate the Laplacian via Hutchinson's method with K samples, where K is chosen by the user. The option whether or not to compute D reflects whether or not the density p under training is estimated. The algorithm calculates batches of adjusted Stein residuals.

Algorithm 1 Batched Adjusted TASTE Residuals

Input: test set $\mathcal{X}_{\text{test}}$, (optional) calibration set $\mathcal{X}_{\text{train}}$, predictor f_θ, score model $\hat{s}_p$, batch sizes $B_{\text{train}}, B_{\text{test}}$, options: `compute_D` (bool), `per_input` (bool), Hutchinson samples K
Helper 1: `STEIN_OP_BATCH`(X)
  Compute gradients $G \leftarrow \nabla f_\theta(X)$ for batch X
  Estimate Laplacians $L \leftarrow \Delta f_\theta(X)$ for batch X (exact or Hutchinson with K samples)
  Compute batch values $U \leftarrow L + \hat{s}_p(X)^\top G$ (vector of length $|X|$)
  Return U
Helper 2: `STEIN_OP_BATCH_ELEMENTWISE`(X)
  Compute gradients $G \leftarrow \nabla f_\theta(X)$ for batch X
  Compute elementwise second-order term $L \leftarrow \Delta_{\text{elem}} f_\theta(X)$ for batch X
  Compute elementwise contributions
    $U \leftarrow L + \hat{s}_p(X) \odot G$
  Return U (matrix of shape $|X| \times d$)
if `per_input` is true **then**
  `STEIN_OP` $\leftarrow$ `STEIN_OP_BATCH_ELEMENTWISE`
else
  `STEIN_OP` $\leftarrow$ `STEIN_OP_BATCH`
end if
if `compute_D` is true **then**
  $S \leftarrow 0,\ n \leftarrow 0$
  for each training batch $X_{\text{train},b}$ of size B_{train} **do**
    $U_b \leftarrow$ `STEIN_OP`$(X_{\text{train},b})$
    $S \leftarrow S + \sum U_b,\ n \leftarrow n + |U_b|$
  end for
  $D_f \leftarrow S/n$
else
  $D_f \leftarrow 0$
end if
Initialize list of scores Scores $\leftarrow \varnothing$
for each test batch $X_{\text{test},b}$ of size B_{test} **do**
  $U_b \leftarrow$ `STEIN_OP`$(X_{\text{test},b})$
  $R_b \leftarrow U_b - D_f$ (vector of adjusted TASTE residuals)
  Append $\{R_b\}$ to Scores
end for
Output: Scores

C Experimental setup and implementation details

Here we provide implementation details for all experiments reported in Section 5, including model architectures, training procedures, score estimation, and evaluation protocols. All experiments are conducted in a fully *unsupervised* setting: neither the classifier nor the score model is trained or fine-tuned on any OOD data. We use Adam (Kingma & Ba, 2014) for all experiments.

C.1 Experimental probes and scoring protocols

We distinguish three operations: selecting a scalar probe, constructing its per-example residual, and aggregating the resulting residuals. These choices are specified separately because they serve different purposes and are not interchangeable.

Let $r(x)$ denote the signed per-example residual produced by the chosen implementation. In the Gaussian and MNIST shift experiments, we report the dataset-level statistic

$$\widehat{S} = \frac{1}{n} \sum_{i=1}^{n} r(x_i),$$

computed separately for each shift type and magnitude. The averaging is applied to signed residuals, without an absolute-value transformation.

Table 3: Scoring and aggregation protocols. Gaussian experiments use the scalar model output; $\text{TASTE}_{\text{ext}}$ uses all classwise softmax probabilities; the remaining experiments use the predicted-class softmax probability, holding the selected class fixed during differentiation. All scores and sums shown are signed. The `per_input` flag refers to Algorithm 1; "–" denotes the separate JVP construction in Appendix C.2.

Experiment	Level	`per_input`	Score used	Further aggregation	Reported
Gaussian shifts	Dataset	`False`	$r_f(x)$	$\frac{1}{n}\sum_{i=1}^{n} r_f(x_i)$ per shifted dataset	Shift-response plots
MNIST rotations and translations	Dataset	`False`	$r_f(x)$	$\frac{1}{n}\sum_{i=1}^{n} r_f(x_i)$ per transformation and magnitude	Shift-response plots
CIFAR-10 detection	Datapoint	`False`	$r_f(x)$	None	AUROC, FPR95
ImageNet-1K: predicted class	Datapoint	`False`	$r_f(x)$	None	AUROC, FPR95
ImageNet-1K: $\text{TASTE}_{\text{ext}}$	Datapoint	–	Classwise $d_c(x)$	$\sum_{c\in\mathcal{E}(x)} d_c(x)$: largest and smallest 5% per input	AUROC, FPR95
MVTec AD localisation	Coordinate	`True`	$r_{f,i}(x)$	Sum RGB-channel contributions per pixel	Pixel AUROC, AUPRO; maps thresholded at $q_{0.99}$

For CIFAR-10 and the original ImageNet-1K detector, we instead use the signed residual $r(x)$ directly as a per-example detection score. We compute AUROC and FPR95 from these scores, without averaging them over the evaluation dataset.

For MVTec AD, we retain the coordinate-wise residuals rather than reducing them to a single per-image value, and use these residuals in the anomaly-map construction. The ImageNet-1K classwise variant uses the separate aggregation rule in Appendix C.2.

Baseline correction and ranking metrics. A common scalar baseline correction applied to the final per-example score leaves its ordering unchanged and therefore does not affect AUROC or FPR95. This observation concerns ranking-based evaluation; the numerical value of the correction remains relevant when interpreting a dataset-level residual or applying a fixed numerical threshold.

C.2 First-order classwise aggregation for ImageNet-1K

For ImageNet-1K, approximating the Laplacian term in TASTE to a reasonable standard becomes computationally infeasible. Hence TASTE has to be run based on the score-part only. However, the predicted-class probe in TASTE examines one class probability at each input. To capture responses across the larger ImageNet-1K output space, $\text{TASTE}_{\text{ext}}$ instead computes a first-order score for every class and then aggregates a selected subset.

Let $z(x) \in \mathbb{R}^C$ denote the classifier logits and write

$$f_c(x) = \sigma_c(z(x)), \qquad \sigma_c(z) = \frac{\exp(z_c)}{\sum_{j=1}^{C}\exp(z_j)},$$

so that f_c is the class-c softmax probe used throughout the paper; let $\hat{s}_p(x) \in \mathbb{R}^d$ denote the score estimate.

The first-order TASTE score for class c is therefore

$$d_c(x) = \nabla f_c(x)^\top \hat{s}_p(x).$$

To compute these scores efficiently for a set of classes, we first differentiate the logits along the same direction. Let $J_z(x) \in \mathbb{R}^{C\times d}$ be the input Jacobian of the logits, with $[J_z(x)]_{ji} = \partial z_j(x)/\partial x_i$. A

single Jacobian–vector product (JVP) gives

$$u(x) = J_z(x)\hat{s}_p(x), \qquad u_j(x) = \nabla z_j(x)^\top \hat{s}_p(x).$$

Thus, $u_j(x)$ describes the directional change in logit j, and all C such changes are obtained together.

We next convert the logit changes into probability changes. The softmax derivative is

$$\frac{\partial \sigma_c(z)}{\partial z_j} = \sigma_c(z)\big(\delta_{cj} - \sigma_j(z)\big),$$

where $\delta_{cj} = 1$ when $c = j$ and is zero otherwise. Applying the chain rule at $z = z(x)$ yields

$$\begin{aligned} d_c(x) &= \sum_{j=1}^{C} \frac{\partial \sigma_c}{\partial z_j}(z(x))\, \nabla z_j(x)^\top \hat{s}_p(x) \\ &= \sum_{j=1}^{C} f_c(x)\big(\delta_{cj} - f_j(x)\big) u_j(x) \\ &= f_c(x) \left(u_c(x) - \sum_{j=1}^{C} f_j(x) u_j(x) \right). \end{aligned}$$

The final expression requires only the probabilities $f_j(x)$ and the JVP output $u(x)$. Consequently, all classwise scores can be computed without constructing a full Jacobian or performing a separate backward pass for each class.

Write $d(x) = (d_1(x), \ldots, d_C(x))^\top$. Since the class probabilities sum to one,

$$\sum_{c=1}^{C} d_c(x) = \left(\nabla \sum_{c=1}^{C} f_c(x) \right)^\top \hat{s}_p(x) = 0.$$

Summing all classwise scores would therefore produce an identically zero statistic.

Instead, for ImageNet-1K, where $C = 1000$, we sort the entries of $d(x)$ and retain the largest 50 and smallest 50 entries. These correspond to the upper and lower 5% of classes. Let

$$\mathcal{E}(x) = \mathrm{Top}_{50}(d(x)) \cup \mathrm{Bottom}_{50}(d(x)),$$

where Top_{50} and Bottom_{50} return the corresponding class indices. The final per-example OOD $\mathrm{TASTE}_{\mathrm{ext}}$ statistic is the signed sum

$$T_{\mathrm{ext}}(x) = \sum_{c \in \mathcal{E}(x)} d_c(x).$$

C.3 Controlled 2D directional shift experiment

Data generation. Training samples are drawn from the standard 2-dimensional Gaussian $p(x) = \mathcal{N}(0, I_2)$. Test distributions are obtained by translating p by a fixed magnitude $\varepsilon = 10$ along directions $\varepsilon R_\varphi u$, where $u = (1, 1)^\top / \sqrt{2}$ and $\varphi \in [0, 2\pi)$. Each test set contains 1000 samples.

Model. The model is a fully connected ReLU network with a single hidden layer of size 64. The model is trained on samples from p using learning rate 0.001, for 100 epochs.

TASTE calculation. Since p is known analytically, the score $s_p(x) = -x$ is used directly. Adjusted Stein residuals $r_f(x)$ are computed using Algorithm 1 and averaged across each test set to obtain an estimate of $S_f(p, q_\varphi)$.

C.4 MNIST geometric perturbation experiments

Dataset and preprocessing. MNIST images are zero-padded from 28×28 to 64×64 prior to training and evaluation. This padding ensures that translations do not remove semantic content, and it allows invariance effects to be isolated from boundary artefacts.

Perturbations. Test sets are generated using:

- Translations with ℓ^∞ magnitude $t \in \{0, 2, 4, \dots\}$. For each datapoint, the translation direction is chosen uniformly from a ℓ^∞ ball of desired radius.
- Rotations with angle $\alpha \in \{-90, -72, \dots, 72\}$ degrees.

Classifier. We use an AlexNet-style convolutional classifier adapted to single-channel inputs, with five 3×3 convolutional layers, ReLU activations, and channel widths 64, 192, 384, 256, and 256. Spatial downsampling is performed using average pooling, yielding feature maps of sizes 32×32 and 16×16, followed by adaptive average pooling to a 1×1 resolution. The classifier head is a three-layer fully connected component with sizes $(512 \to 128 \to 10)$.

Score model. We use a lightweight convolutional score network to approximate the input-space score field $\hat{s}_p(x) \approx \nabla_x \log p(x)$. The model consists of four 3×3 convolutional layers with ReLU activations and channel widths 32, 64, 64, and 1, preserving the 64×64 spatial resolution throughout. The model is trained using Denoising Score Matching employing standard practices.

Stein residuals. Adjusted Stein residuals $r_f(x)$ are computed using Algorithm 1, with the correction term D_f estimated on held-out MNIST training samples. The Laplacian is estimated using Hutchinson's estimator with the number of samples K set to 5.

C.5 CIFAR-10 BENCHMARK EVALUATION

Classifier. The in-distribution classifier is a ResNet-18. We take model weights pre-trained on ImageNet and fine-tune them on CIFAR-10, with learning rate 0.0001 for 5 epochs.

Score model. The score function $\hat{s}_p(x)$ is estimated using a pretrained diffusion model. Specifically, we obtain the weights from the publicly released `google/ddpm-cifar10-32` checkpoint. The model is trained to predict the additive noise ε in the diffusion process, from which the score is obtained via the standard conversion $\hat{s}_p(x_t) = -\varepsilon_\theta(x_t, t)/\sigma_t$.

While the diffusion model is trained over 1000 noise levels, we empirically find that evaluating the score at a relatively low time step ($t = 50$ out of 1000) yields a useful balance between proximity to the clean data score and numerical stability. Lower noise levels (smaller t) tend to produce noisier or unstable score estimates, whereas higher noise levels carry less information about the data distribution. Prior work on score-based models suggests that intermediate noise levels often provide the most robust signal for downstream tasks (Song et al., 2021; Ho et al., 2020), even though they are not typically interpreted as clean-data scores per se. For empirical analysis see Appendix F.3.

C.6 IMAGENET-1K EXPERIMENTS

Classifier. For ImageNet-1K experiments we use a ResNet-50 backbone pretrained on ImageNet-1K available via torchvision package (`ResNet50_Weights.IMAGENET1K_V1`: https://download.pytorch.org/models/resnet50-0676ba61.pth).

Score model. The score function $\hat{s}_p(x)$ is estimated using a pretrained unconditional diffusion model trained on 256×256 ImageNet images. Specifically, we use the publicly released OpenAI diffusion checkpoint from the guided diffusion framework: https://github.com/openai/guided-diffusion. As in the CIFAR-10 experiments, the diffusion model predicts the additive noise ε in the forward diffusion process, and the score estimate is obtained via the standard relation

$$\hat{s}_p(x_t) = -\varepsilon_\theta(x_t, t)/\sigma_t.$$

All ImageNet experiments use the same diffusion timestep configuration as the CIFAR-10 setup.

Operator and probe choice. Both ImageNet-1K variants omit the Laplacian term for computational reasons, a choice alsosupported by the evidence in Table 9. The TASTE row uses the predicted-class softmax probe, whereas $\text{TASTE}_{\text{ext}}$ uses the classwise aggregation in Appendix C.2.

C.7 COMPUTE RESOURCES AND RUNTIME.

All experiments were conducted on a single NVIDIA H100 GPU. The full evaluation suite on ImageNet-1K corresponds to approximately 240k evaluations per detector: 40k samples for ID

train+calibration, 10k ID test samples, roughly 100k adversarial samples (across 10 attacks), and approximately 100k samples across classic OOD benchmarks.

Depending on the detector, this results in evaluation costs ranging from approximately 0.4 GPU-hours (MSP/Energy) to over 10 GPU-hours (SCOPED) per full sweep. TASTE requires approximately 8 GPU-hours per full evaluation sweep. Summed across all reported detectors, the total evaluation cost is approximately 50 GPU-hours.

These estimates refer to evaluation only; the classifier and score-model training are not accounted for, as we used pretrained checkpoints. The full research project additionally required more compute for preliminary experiments, smaller-scale experiment, and ablations, but their total computational cost is substantially smaller than that of the reported ImageNet-1K experimental sweep.

C.8 Baselines and hyperparameters

We evaluate against a broad range of established and recent OOD detection methods, covering confidence-based, distance-based, energy-based, representation-based, and post-hoc approaches.

Confidence-based baselines include Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2017) and ODIN (Liang et al., 2018). Distance- and representation-based methods include Mahalanobis distance (Lee et al., 2018) and kNN-based detectors (Sun et al., 2022). Energy-based scoring is implemented following (Liu et al., 2020).

We additionally include several recent post-hoc and hybrid methods that improve OOD separability through feature shaping, logit correction, perturbation sensitivity, or score-based geometry, including ReAct (Sun et al., 2021), ViM (Wang et al., 2022), ASH (Djurisic et al., 2023), PRO (Chen et al., 2025), and SCOPED (Barkley et al., 2025).

All baseline methods are implemented using publicly available codebases or standard reference implementations, primarily via the `pytorch-ood` library (Kirchheim et al., 2022). Hyperparameters are selected according to the respective original papers or default settings provided in the library.

C.9 MVTec AD localisation experiments

Dataset. We use the MVTec Anomaly Detection dataset with the standard train/test split. Only `train/good` images are used for calibration.

Models.

- Classifier: ImageNet-pretrained ResNet-50, with weights downloaded through torchvision.
- Score model: ImageNet-pretrained diffusion model - obtained from the publicly available Hugging Face checkpoint *google/ddpm-ema-celebahq-256*. Similarly to C.5, the score is obtained via the standard conversion $\hat{s}_p(x_t) = -\varepsilon_\theta(x_t, t)/\sigma_t$ with time step $t = 50$ (out of 1000).

No model is fine-tuned on MVTec AD.

Per-pixel Stein heatmaps. Per-pixel residuals are computed via the coordinate-wise Stein operator. Note that this gives a per-input score, and so there are three per pixel - one for each input channel. We take their sum to obtain the per-pixel score. Heatmaps are resized to 256×256 for inference and upsampled to ground-truth resolution for evaluation.

Evaluation. We report pixel-level AUROC, average precision, and AUPRO. Thresholds for visualisation are obtained via upper-tail calibration on `train/good` images at quantile $\alpha = 0.01$.

Additional qualitative results are shown in Appendix E.

C.10 On the Laplacian term for ReLU networks.

For architectures with a piecewise-affine backbone (e.g. ResNets with ReLU and pooling) followed by a smooth softmax output, classical second derivatives of the backbone vanish almost everywhere. Let $z(x) \in \mathbb{R}^K$ denote the logits and $f_k(x) = \sigma_k(z(x))$ the softmax probabilities. Since $\nabla_x^2 z_i(x) = 0$

a.e., all non-zero curvature of f_k arises from the softmax via the chain rule, yielding

$$\Delta f_k(x) = \sum_{i,j} \frac{\partial^2 \sigma_k}{\partial z_i \partial z_j} \langle \nabla_x z_i(x), \nabla_x z_j(x) \rangle.$$

Thus, for such networks the Laplacian can be computed using only first-order input gradients of the logits and the closed-form softmax Hessian, without second-order backpropagation through ReLU or pooling layers. If the model has no smooth output nonlinearity (e.g. regression), the Laplacian vanishes a.e. and may be omitted. (For large K, the sum can even be restricted to the top-k logits for efficiency).

This construction is used in the MVTecAD localisation experiments C.9 to calculate the per-input second derivative explicitly.

C.11 Comment on licences.

All datasets, pretrained models, and benchmark checkpoints used in this work are publicly available and used in accordance with their respective licences and terms of use. CIFAR-10 and MNIST are used under their standard academic usage conditions, while ImageNet-1K is used under the ImageNet non-commercial research access policy. Publicly released pretrained classifiers and diffusion checkpoints are used under their corresponding open-source or research licences.

D Granular Experimental Results

Benchmark datasets and shift regimes. We evaluate task-aware OOD detection across a broad set of datasets designed to capture qualitatively different forms of distribution shift, ranging from semantic dataset shift to low-level corruptions, structured geometric perturbations, and adversarial examples, see also Tajwar et al. (2021).

Classical OOD benchmarks. Following standard practice, we use SVHN, LSUN, iSUN, Describable Textures, and Places365 as semantic OOD datasets with respect to CIFAR-10. These datasets differ substantially from CIFAR-10 in terms of image content, statistics, and semantics, and are commonly used to benchmark the ability of OOD methods to detect inputs drawn from entirely different data-generating processes. SVHN consists of real-world digit images with markedly different backgrounds and color statistics; LSUN and iSUN contain large-scale scene images; Textures focuses on fine-grained texture patterns; and Places365 covers a wide range of scene categories. Together, these datasets probe classical semantic OOD detection, where the test distribution differs from the training distribution at a high semantic level.

CIFAR-10-C (common corruptions). To evaluate robustness under realistic covariate shift, we use CIFAR-10-C, which applies 19 common image corruptions—such as Gaussian noise, blur, fog, snow, contrast changes, and compression artifacts—each at five severity levels. These corruptions alter low-level image statistics while largely preserving high-level semantic content. CIFAR-10-C is widely used to assess distribution shift arising from sensor noise, environmental conditions, or data acquisition artifacts, and provides a systematic test of how OOD methods respond to gradual, non-adversarial changes in the input distribution.

CIFAR-10-P (structured perturbations). We additionally evaluate on CIFAR-10-P, which applies temporally consistent and structured geometric perturbations such as translations, rotations, scaling, shearing, and motion blur. Unlike CIFAR-10-C, these perturbations often leave pixel-level statistics relatively unchanged while inducing changes that directly interact with the classifier's invariances and equivariances. CIFAR-10-P therefore provides a particularly relevant testbed for *task-aware* OOD detection, as some perturbations (e.g. translations) are largely benign for convolutional networks, whereas others (e.g. rotations) can substantially degrade task performance despite minimal changes in overall image statistics.

Adversarial attacks. To probe behaviour under worst-case, task-adversarial distribution shifts, we include adversarial examples generated using FGSM, PGD, and AutoAttack under both ℓ^∞ and ℓ^2 constraints, with multiple perturbation budgets. These attacks are specifically constructed to induce

classifier failure while keeping perturbations visually subtle and often imperceptible. Adversarial examples therefore represent a regime in which confidence-based and likelihood-based OOD methods are known to struggle, as inputs may remain close to the training distribution in pixel space while causing severe task-level errors. Including adversarial attacks allows us to assess whether task-aware Stein residuals respond to shifts that directly affect the predictor, even when traditional OOD signals fail.

Summary. Across these datasets, we cover four complementary shift regimes:

1. semantic dataset shift (classical OOD benchmarks),
2. low-level covariate shift (CIFAR-10-C),
3. structured, invariance-breaking perturbations (CIFAR-10-P), and
4. worst-case, task-adversarial perturbations (adversarial attacks).

This comprehensive evaluation enables a detailed assessment of how different OOD detection methods respond to shifts that are benign, harmful, or adversarial with respect to the underlying prediction task.

Granular results. Each table in this section reports a per-dataset breakdown for a single OOD detection method (see the baselines in Section 5.3 for details) and corresponds to one column of the aggregated results presented in the main text. AUROC is the area under the receiver operating characteristic curve; FPR95 is the false positive rate at 95% true positive rate.

These fine-grained results expose substantial variability in OOD detection performance across datasets, corruption types, and shift regimes. Such variability highlights the sensitivity of existing methods to the specific nature of distribution shift and provides additional empirical support for the central observation of Tajwar et al. (2021): namely, that no single OOD detection method consistently dominates across all benchmarks, and that comparative performance is strongly dataset-dependent. Tables 4 - 7 show per-dataset results on CIFAR-10 for a subset of baseline methods. As in the results presented in the main text, TASTE uses predicted class probability as the probe.

Table 4: OOD detection results: adversarial

OOD Dataset	metric	MSP	Energy	ODIN	Mahalanobis	kNN+	TASTE
adversarial_autoattack_linf_2_255	AUROC	0.5019	0.4896	0.5393	**0.5445**	0.5276	0.4786
	FPR95	0.9611	0.9643	**0.93**	0.9371	0.9513	0.9571
adversarial_autoattack_linf_4_255	AUROC	0.2841	0.289	0.4851	0.505	0.3839	**0.6543**
	FPR95	0.9681	0.9711	0.9532	0.9213	0.9711	**0.893**
adversarial_autoattack_linf_8_255	AUROC	0.1557	0.1675	0.3319	0.7207	0.4341	**0.8609**
	FPR95	0.9688	0.9713	0.9718	**0.513**	0.9688	0.602
adversarial_fgsm_linf_4_255	AUROC	0.6254	0.6039	0.5673	0.6172	**0.6266**	0.4866
	FPR95	0.923	0.9147	0.9306	0.9071	**0.8999**	0.9508
adversarial_fgsm_linf_8_255	AUROC	0.6075	0.6026	0.5151	0.6341	**0.6388**	0.4949
	FPR95	0.922	0.8993	0.9715	0.8969	**0.8847**	0.948
adversarial_pgd_l2_0.5_steps=50	AUROC	0.2656	0.2741	0.5036	0.4714	0.3748	**0.5727**
	FPR95	0.998	0.9985	0.9418	0.9507	0.9918	**0.9348**
adversarial_pgd_l2_1.0_steps=50	AUROC	0.05975	0.07737	0.3817	0.634	0.311	**0.8612**
	FPR95	1	1	0.9631	0.6251	0.9981	**0.5297**
adversarial_pgd_linf_2_255_steps=50	AUROC	0.5837	0.5521	0.5402	**0.5937**	0.5812	0.433
	FPR95	0.9646	0.9689	**0.9205**	0.9273	0.9484	0.9637
adversarial_pgd_linf_4_255_steps=50	AUROC	0.3609	0.3548	**0.5535**	0.5159	0.4419	0.5017
	FPR95	0.9908	0.9929	**0.9151**	0.9442	0.981	0.9486
adversarial_pgd_linf_8_255_steps=50	AUROC	0.1304	0.1389	0.4769	0.5827	0.3248	**0.7578**
	FPR95	0.9991	0.9997	0.9381	**0.7652**	0.996	0.7966

Table 5: OOD detection results: CIFAR-10-C

OOD Dataset	metric	MSP	Energy	ODIN	Mahalanobis	kNN+	TASTE
cifar10c_brightness	AUROC	0.5321	0.5389	0.4928	0.5356	**0.543**	0.5373
	FPR95	0.9454	0.9354	0.9562	0.9452	0.939	**0.9335**
cifar10c_contrast	AUROC	0.6945	0.7206	**0.802**	0.5841	0.6653	0.6846
	FPR95	0.86	0.781	**0.5962**	0.929	0.8666	0.7849
cifar10c_defocus_blur	AUROC	0.5801	0.5763	**0.5864**	0.5667	0.5847	0.5842
	FPR95	0.9308	0.9212	**0.8546**	0.9318	0.9186	0.8969
cifar10c_elastic_transform	AUROC	0.6027	**0.6091**	0.5777	0.5549	0.5956	0.6088
	FPR95	0.9272	0.9236	0.8942	0.9504	0.9278	**0.8505**
cifar10c_fog	AUROC	0.6217	0.6364	**0.6875**	0.5347	0.6108	0.6465
	FPR95	0.9018	0.8914	**0.7756**	0.9596	0.9116	0.8526
cifar10c_frost	AUROC	0.6086	**0.6278**	0.528	0.5703	0.6049	0.627
	FPR95	0.9166	0.8964	0.957	0.9458	0.9148	**0.8693**
cifar10c_gaussian_blur	AUROC	0.617	0.6142	**0.6462**	0.6018	0.6192	0.6206
	FPR95	0.9222	0.9128	**0.8006**	0.9184	0.91	0.8583
cifar10c_gaussian_noise	AUROC	0.6341	**0.6509**	0.4403	0.6333	0.646	0.6332
	FPR95	0.8998	0.881	0.9884	0.8932	**0.869**	0.8778
cifar10c_glass_blur	AUROC	0.6751	**0.7104**	0.5464	0.6415	0.6791	0.6725
	FPR95	0.882	0.8446	0.9454	0.8968	0.8704	**0.7801**
cifar10c_impulse_noise	AUROC	0.702	**0.7305**	0.4776	0.7233	0.7255	0.6812
	FPR95	0.865	0.8148	0.985	**0.7908**	0.7958	0.7985
cifar10c_jpeg_compression	AUROC	0.5505	**0.5577**	0.5125	0.5168	0.5397	0.5521
	FPR95	0.936	0.927	0.9338	0.9516	0.9344	**0.9042**
cifar10c_motion_blur	AUROC	0.6302	0.6409	0.6239	0.6099	**0.6453**	0.6365
	FPR95	0.9102	0.9062	0.8426	0.9094	0.8946	**0.8343**
cifar10c_pixelate	AUROC	0.5407	0.5479	0.5094	0.5214	0.5376	**0.5486**
	FPR95	0.9368	**0.9178**	0.9372	0.9408	0.9268	0.9193
cifar10c_saturate	AUROC	0.5668	0.5892	0.5264	0.5833	**0.5923**	0.5599
	FPR95	0.9298	**0.9018**	0.9292	0.9038	0.9024	0.9253
cifar10c_shot_noise	AUROC	0.5986	**0.6101**	0.437	0.5952	0.6074	0.6015
	FPR95	0.9104	0.8944	0.9848	0.9096	**0.8914**	0.9004
cifar10c_snow	AUROC	0.6104	0.6159	0.4865	0.5842	0.6032	**0.6169**
	FPR95	0.9206	0.9046	0.972	0.9346	0.9134	**0.8872**
cifar10c_spatter	AUROC	0.621	**0.6448**	0.4804	0.6007	0.626	0.6176
	FPR95	0.905	**0.8742**	0.9726	0.9044	0.887	0.892
cifar10c_speckle_noise	AUROC	0.6064	**0.6209**	0.444	0.6026	0.6138	0.6119
	FPR95	0.9118	0.8902	0.9842	0.912	**0.8878**	0.9009
cifar10c_zoom_blur	AUROC	0.6434	0.645	0.6435	0.6118	**0.6466**	0.6444
	FPR95	0.895	0.8912	**0.8146**	0.9182	0.8952	0.8275

Table 6: OOD detection results: CIFAR-10-P

OOD Dataset	metric	MSP	Energy	ODIN	Mahalanobis	kNN+	TASTE
cifar10p_brightness	AUROC	0.5497	0.5549	0.499	0.5488	**0.5583**	0.5557
	FPR95	0.9395	0.9357	0.9585	0.9415	0.934	**0.9205**
cifar10p_gaussian_noise	AUROC	0.5069	**0.5107**	0.4618	0.4965	0.5076	**0.5107**
	FPR95	0.9473	**0.9454**	0.9629	0.9559	0.9505	0.9456
cifar10p_motion_blur	AUROC	0.6294	0.6368	0.6134	0.6183	**0.6446**	0.6342
	FPR95	0.9197	0.9203	0.8495	0.9152	0.9095	**0.8123**
cifar10p_rotate	AUROC	0.7012	**0.7351**	0.5982	0.6109	0.6842	0.6822
	FPR95	0.8787	0.8463	0.9183	0.9477	0.8917	**0.731**
cifar10p_scale	AUROC	0.6425	0.636	0.6375	0.6235	**0.6482**	0.6448
	FPR95	0.9102	0.9122	0.8345	0.9203	0.9056	**0.8149**
cifar10p_shear	AUROC	0.6641	**0.6996**	0.5723	0.5878	0.6514	0.6496
	FPR95	0.8966	0.8619	0.921	0.938	0.8969	**0.8146**
cifar10p_shot_noise	AUROC	0.5136	**0.5184**	0.4522	0.4989	0.5137	0.5178
	FPR95	0.9444	**0.9422**	0.9708	0.9559	0.946	0.9459
cifar10p_snow	AUROC	0.5989	**0.6132**	0.475	0.5697	0.5946	0.6041
	FPR95	0.9179	0.905	0.9735	0.9357	0.9116	**0.8996**
cifar10p_tilt	AUROC	0.5564	**0.5699**	0.5331	0.5059	0.5428	0.5572
	FPR95	0.9378	0.9359	0.9293	0.9629	0.947	**0.8954**
cifar10p_translate	AUROC	0.6523	**0.6884**	0.6304	0.5762	0.6388	0.6318
	FPR95	0.9019	0.8788	0.9218	0.959	0.9148	**0.7862**
cifar10p_zoom_blur	AUROC	0.5605	0.5583	0.5526	0.5457	0.567	**0.5676**
	FPR95	0.9374	0.9411	0.9044	0.946	0.9379	**0.8937**

Table 7: OOD detection results: OOD benchmarks

OOD Dataset	metric	MSP	Energy	ODIN	Mahalanobis	kNN+	TASTE
iSUN	AUROC	0.7926	**0.8561**	0.682	0.8379	0.8203	0.7408
	FPR95	0.7848	0.6455	0.9396	0.6829	0.6995	**0.5901**
LSUN	AUROC	0.8208	**0.8999**	0.7047	0.88	0.8677	0.7666
	FPR95	0.7391	**0.5161**	0.9421	0.5912	0.5945	0.5364
Places365	AUROC	0.8005	**0.854**	0.7206	0.7943	0.8335	0.7746
	FPR95	0.7878	0.6461	0.86	0.7714	0.6894	**0.5609**
SVHN	AUROC	0.8044	0.816	**0.9216**	0.7705	0.8122	0.749
	FPR95	0.8258	0.8118	**0.4123**	0.8402	0.7944	0.5432
Textures	AUROC	0.7976	**0.8339**	0.7931	0.8113	0.8101	0.7445
	FPR95	0.8018	0.7222	0.7239	0.7119	0.7358	**0.592**

E PER-PIXEL ANOMALY HEATMAPS

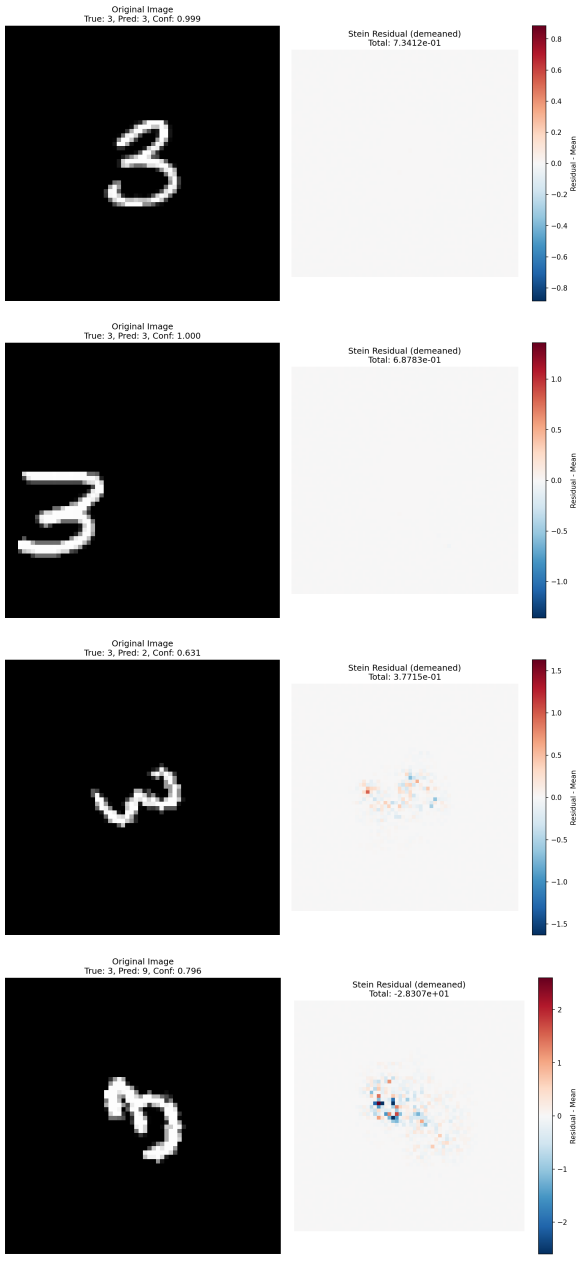


Figure 4: **Demonstration of per-pixel anomaly heatmaps for the MNIST-based task.** Input digits are shown on the left, while the heatmaps are on the right. Note that both original and translated digits do not generate significant anomaly signal.

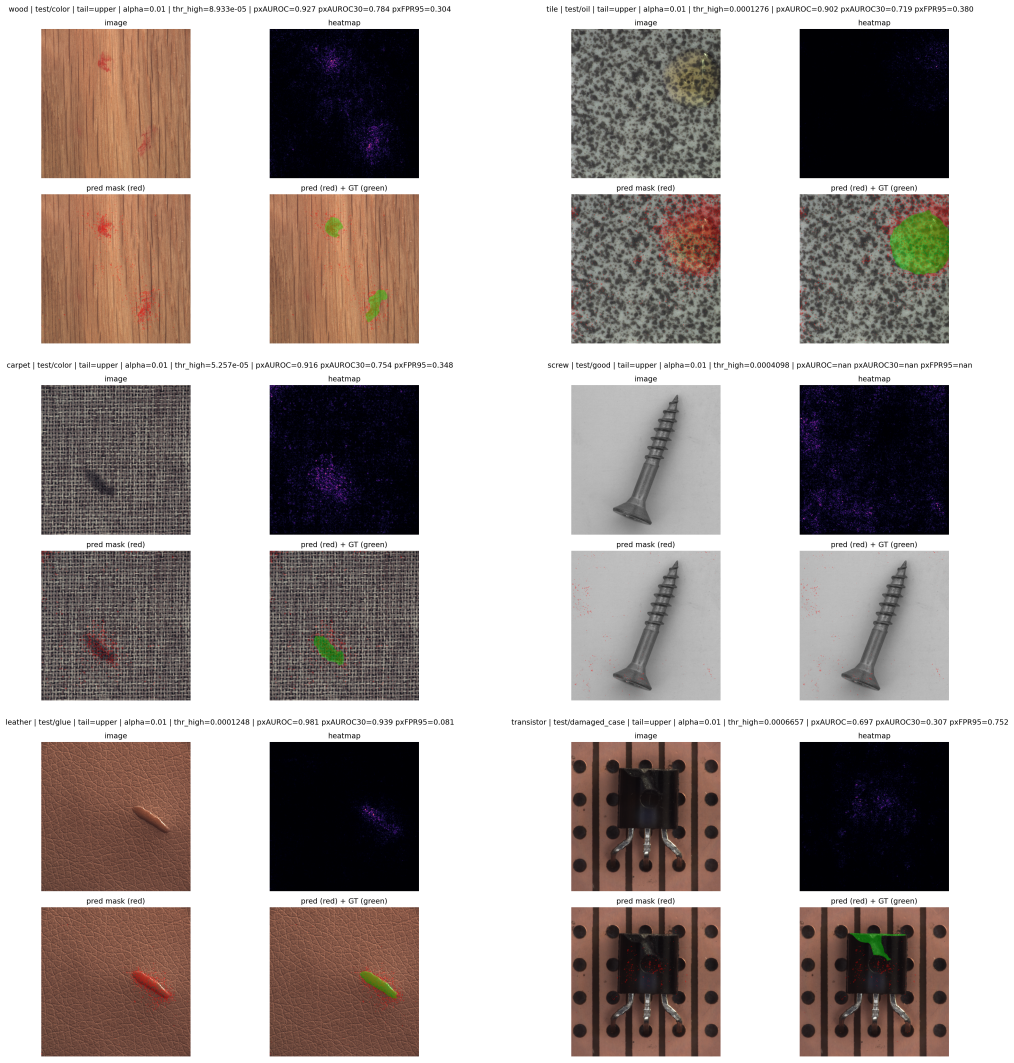


Figure 5: **Additional qualitative results.** Each panel is a randomly chosen example for one of the MVTec AD categories. Within each panel there are four subplots: the original image (with corruption), detector's heatmap, prediction overlay at $\alpha = 0.01$, prediction and ground truth (GT) overlay.

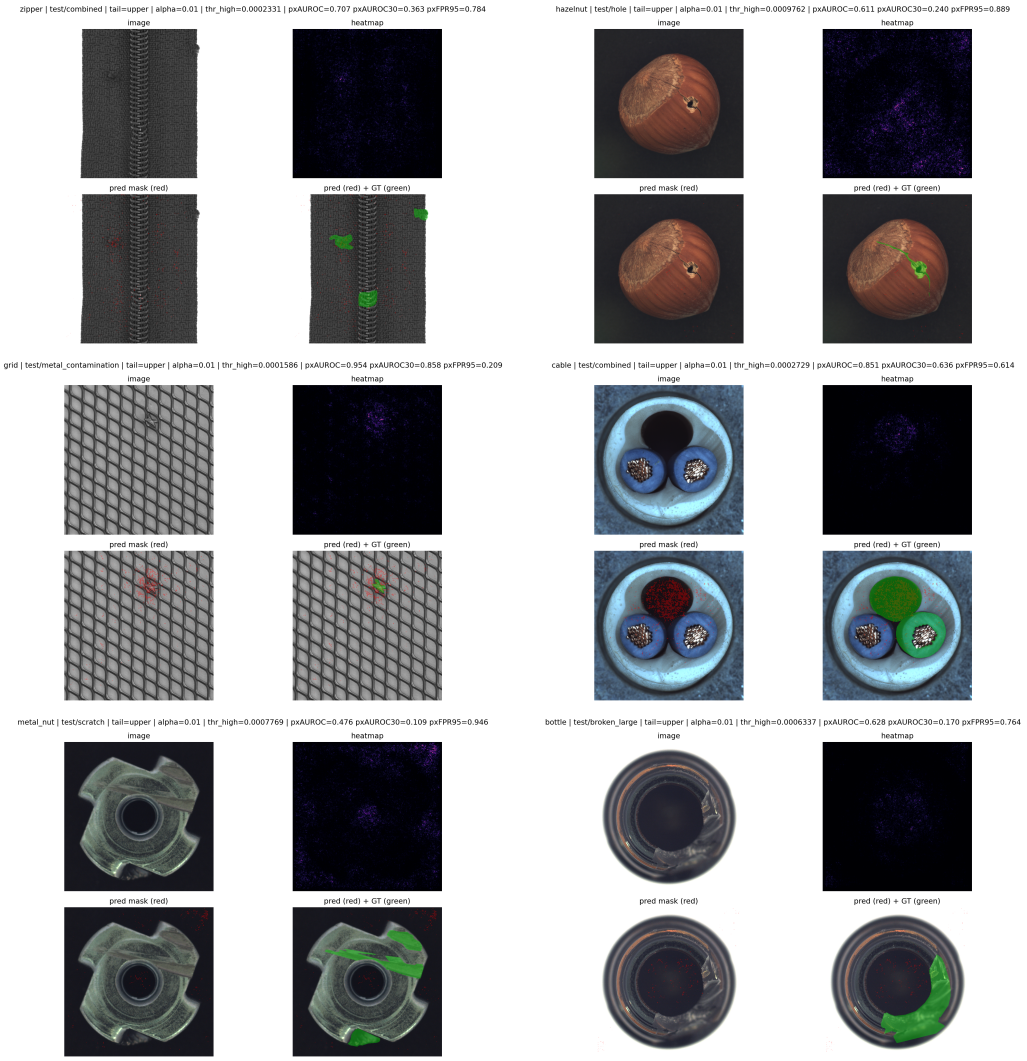


Figure 6: **Additional qualitative results.** Each panel is a randomly chosen example for one of the MVTec AD categories. Within each panel there are four subplots: the original image (with corruption), detector's heatmap, prediction overlay at $\alpha = 0.01$, prediction and ground truth (GT) overlay.

Observations. Figures 5 and 6 show that across MVTec AD categories, per-pixel Stein residuals can produce spatially coherent and visually meaningful localisation signals, particularly on texture classes, where defects manifest as systematic deviations in local appearance. Quantitative metrics and qualitative inspection indicate substantially weaker performance on object categories, where anomalies are often small, highly specific, and semantically defined (e.g. missing components or structural deformations). We hypothesise that such object-level defects are less accessible to a zero-shot pipeline based on a generic ImageNet-pretrained classifier and score model, which have not been exposed to object-specific normality constraints. Importantly, no component of our pipeline—neither the classifier nor the score model—has been trained or adapted on MVTec AD data, making this evaluation fully *zero-shot*.

F Ablation Study

Next we present an ablation study of individual operator components and a sensitivity analysis of the diffusion time step in the score model. Most ablation experiments are carried out on the full

CIFAR-10 OOD benchmark suite used in the main results, as it probes detection performance across heterogeneous shift regimes.

F.1 Sensitivity to score quality.

TASTE relies on an estimate of the score function $\nabla \log p(x)$, raising the question of robustness to imperfect score models. To evaluate this, we degrade the quality of the score estimate by applying a low resolution diffusion model to downscaled input and then upscaling to the original size.

Table 8 shows that this mismatch leads to a consistent drop in AUROC and an increase in FPR95 across datasets. However, the degradation is moderate: the low-resolution variant remains competitive while providing an approximately $4\times$ speedup. Resolution mismatch reduces average performance in this experiment, although individual metrics need not degrade. This suggests that TASTE does not require highly precise score estimates to remain effective, and can trade accuracy for efficiency in practical deployments.

F.2 Ablation: Operator decomposition

We study the effect of progressively simplifying the Langevin Stein operator by removing individual components. Specifically, we compare:

- **Full operator (TASTE):**
$$\mathcal{L}_p f(x) = \Delta f(x) + s_p(x)^\top \nabla f(x),$$
- **No-Laplacian:**
$$\mathcal{L}_p^{\text{no-lap}} f(x) = s_p(x)^\top \nabla f(x),$$
- **Score-only:** using only the score L^2 norm, without interaction with ∇f.

Table 9 reports performance across adversarial attacks, corruption benchmarks (CIFAR-10-C), structured perturbations (CIFAR-10-P), and standard OOD datasets. Across all benchmark categories, the full operator and the no-Laplacian variant exhibit nearly identical performance, while the score-only variant performs substantially worse. In particular, the AUROC and FPR95 of TASTE and TASTE-no-Laplacian differ only marginally, whereas additionally removing the ∇f interaction leads to a pronounced drop in performance. This behaviour is explained by the observation that in this experiment the full score is dominated by

$$|s(x)^\top \nabla f_c(x)| \;\approx\; \|s(x)\|\,\|\nabla f_c(x)\|\,|\cos\theta_c|,$$

where θ_c is the angle between the score direction and the class-specific gradient. Empirically, we observe that while the gradient and the Laplacian are of comparable magnitude, the output of the score model is often much larger in scale, causing the dot-product term of the operator to dominate, even if the correlation between the score and the gradient is small.

Role of the Laplacian term. Under the stated regularity assumptions, the Laplacian provides the curvature correction required for the Langevin Stein identity. Its theoretical role does not, however, imply improved detection performance in every setting.

For a fixed probe f, let $\widetilde{S}_f^{\text{full}}$ and $\widetilde{S}_f^{(1)}$ denote the population-centred functionals with and without the Laplacian, using the same score estimate. Then

$$\widetilde{S}_f^{\text{full}}(p,q) - \widetilde{S}_f^{(1)}(p,q) = (\mathbb{E}_q - \mathbb{E}_p)[\Delta f].$$

Thus, omitting the Laplacian changes the dataset-level response according to the shift in expected probe curvature. This identity characterises the dataset-level response; the effect on per-example detection is assessed empirically.

Our CIFAR-10 ablation (Table 9) shows little detection benefit from the Laplacian under the evaluated implementation, supporting the cheaper first-order variant in this setting. The MNIST rotation experiment (Figure 7) demonstrates that the term can matter, but does not establish a general advantage on larger benchmarks. Its practical contribution depends on the probe, shift, and accuracy of curvature estimation.

Table 8: Effect of score model degradation on OOD detection performance on ImageNet-1K. Pre-softmax logit of the predicted class is used as the probe.

	iSUN		LSUN		Places365		SVHN		Textures		Average	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
TASTE	0.7603	0.9720	0.7820	0.9380	0.5927	0.9410	0.9477	0.1200	0.6396	0.8890	**0.7445**	**0.7720**
TASTE (low res)	0.7120	0.9460	0.7044	0.9480	0.5330	0.9510	0.9206	0.2050	0.5806	0.9490	**0.6901**	**0.8190**

Table 9: Operator construction ablation study - performance on CIFAR-10 across heterogeneous shift regimes. Results are averaged within each category. The dataset composition is identical to that used in the CIFAR-10 evaluation suite presented in Section 5.3.

	Adversarial		CIFAR-10-C		CIFAR-10-P		OOD-benchmarks		Overall	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
TASTE	0.6144	0.865	0.6193	0.8691	0.5954	0.8611	0.7647	0.5677	**0.6301**	**0.8207**
TASTE-no-Laplacian	0.6264	0.8711	0.6098	0.8708	0.5902	0.8657	0.7597	0.5751	0.6253	0.8367
Score-only	0.5337	0.9437	0.6364	0.787	0.5789	0.8756	0.4452	0.8623	0.5783	0.8518

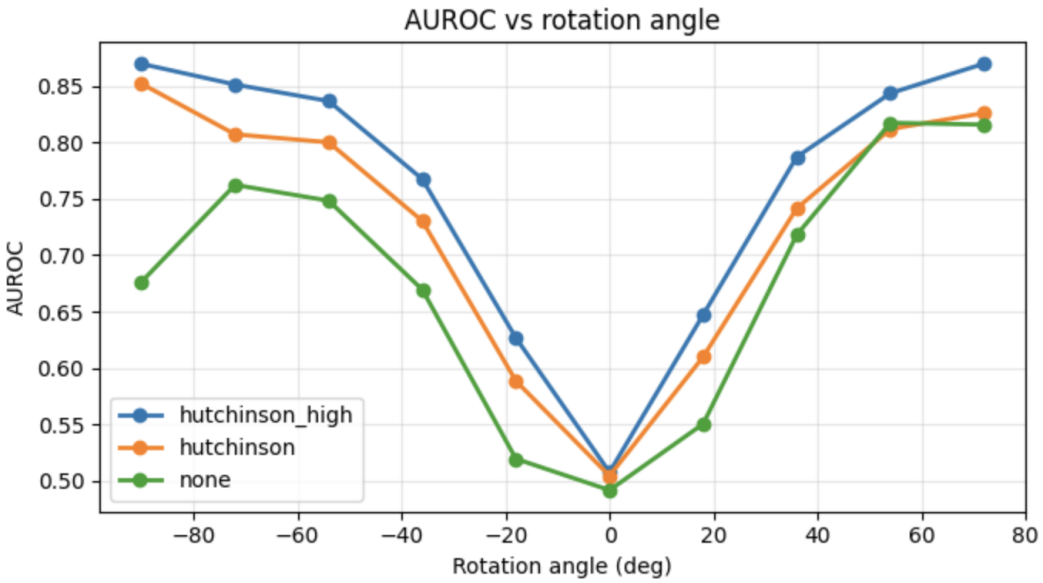
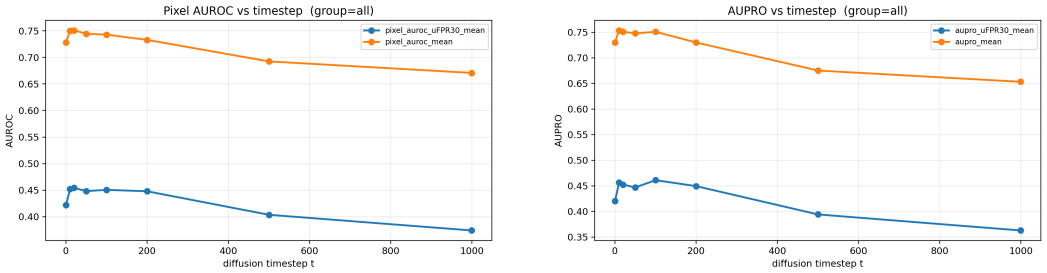


Figure 7: **The effect of Laplacian estimation on performance of TASTE.** Evaluation on MNIST under rotation-induced shifts using an AlexNet-style CNN shows monotone improvement as the quality of Laplacian estimation increases.

F.3 Sensitivity to diffusion time step

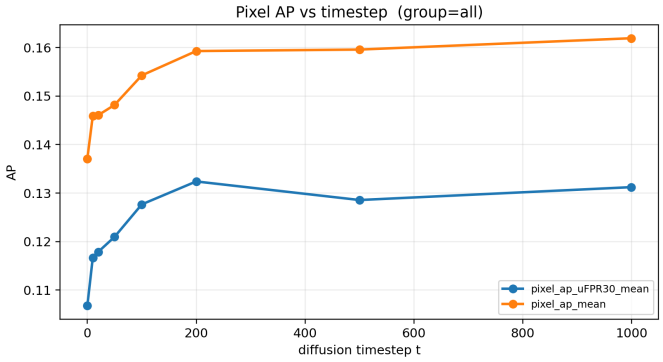
This sensitivity analysis is carried out on a per-pixel level, using the MVTec AD dataset from the main part of the paper. The per-pixel Stein residual depends on the quality and scale of the score estimate $\hat{s}_p(x) \approx \nabla_x \log p_t(x)$ produced by the diffusion model at a chosen time step t. Since the diffusion score interpolates between fine-grained data geometry (small t) and heavily smoothed distributions (large t), the time step implicitly controls the granularity of the resulting anomaly signal. We therefore study how per-pixel OOD performance varies as a function of t.

Setup. Using the MVTec AD dataset, we evaluate pixel-level anomaly detection metrics across a range of diffusion time steps while keeping all other components fixed: the classifier, Stein operator, residual aggregation, and evaluation protocol. For each time step, we compute per-pixel Stein heatmaps and report pixel-level AUROC, AUPRO, and average precision (AP), averaged across categories.



(a) Pixel AUROC

(b) AUPRO



(c) Pixel AP

Figure 8: **Effect of diffusion time step on per-pixel anomaly detection.** As the diffusion time step increases, ranking-based metrics (AUROC (Panel a), AUPRO (Panel b)) decrease, while average precision (Panel c) improves, indicating a shift from global separability to more confident, sparse anomaly responses. Orange: standard area under curve statistics, blue: additionally restricting the False Positive Rate to below 30% and thus promoting early separation.

Observed behaviour. Figure 8 shows a consistent and interpretable trade-off. As the diffusion time step increases, both pixel-level AUROC and AUPRO decrease, indicating reduced global separability and weaker region-level coherence of the anomaly scores. In contrast, average precision increases monotonically with time step. This divergence reflects a shift in the score distribution: higher time steps smoothen the score field and suppress moderate responses, while amplifying extreme deviations from the data manifold. As a result, fewer pixels receive large scores, but those that do are more likely to correspond to true defect regions.

Interpretation. These results highlight that different metrics probe distinct aspects of per-pixel anomaly detection. AUROC and AUPRO assess global ranking quality and spatial consistency, respectively, and therefore favour lower diffusion noise where fine-grained structure is preserved. Average precision, by contrast, emphasises the reliability of the highest-scoring pixels and is sensitive to the purity of the score tail. From this perspective, increasing the diffusion time step trades global separability for more confident, sparse anomaly responses. This behaviour is consistent with the geometry of diffusion score fields and suggests that the time step can be selected based on the desired operating regime: lower time steps for comprehensive localisation and higher time steps for high-confidence defect highlighting.

Practical Implications. This sensitivity analysis demonstrates that the proposed Stein residual does not behave arbitrarily with respect to the score model, but exhibits predictable and structured changes as the underlying notion of data geometry is varied. In all cases, the method remains fully zero-shot, requiring no retraining or fine-tuning on the target dataset.

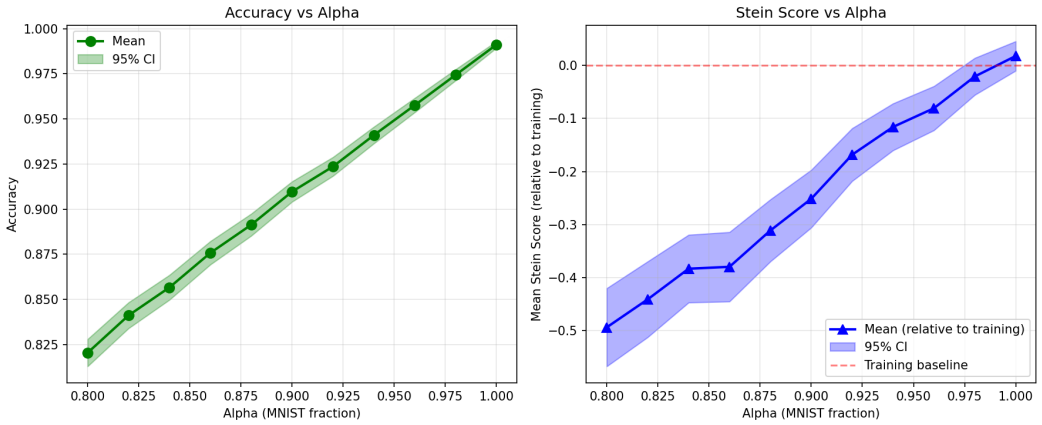


Figure 9: **Model performance and Stein score as a function of test set corruption level.** The methodology picks up on the presence of out-of-distribution input data for even very small corruption levels. Shaded regions: ± 2 s.d., as approximate 95% confidence intervals.

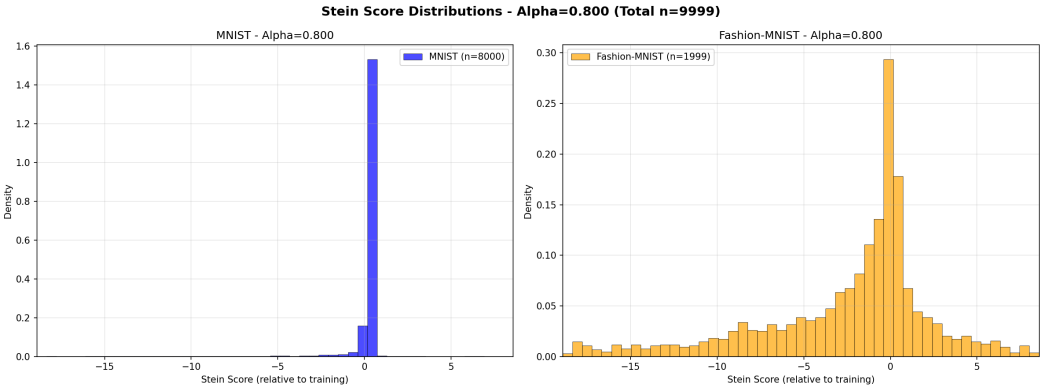


Figure 10: **Distribution of the per-sample TASTE residuals.** The left panel shows evaluation on clean (MNIST) samples, while the right presents the scores generated from corrupted data. The considerable distribution shift of the per-sample TASTE residuals allows for per-input anomaly detection.

G Per-sample Stein residuals in mixed MNIST / Fashion-MNIST OOD

We now assess the discriminative power of *per-sample* Stein residuals in a heterogeneous OOD scenario based on mixed in- and out-of-distribution test sets, a setting that is known to be challenging for OOD detection and has been used to expose failure modes of likelihood-based and confidence-based approaches (Nalisnick et al., 2019; Rabanser et al., 2019).

Setup. The classifier f_θ and score model $\hat{s}_p$ are trained exactly as in the previous subsection. Test sets are constructed by mixing MNIST (in-distribution) and Fashion-MNIST (OOD) according to a corruption level $\beta \in \{0.80, 0.82, \ldots, 1.00\}$, where the test set contains $\beta \cdot 10{,}000$ MNIST images and $(1-\beta) \cdot 10000$ Fashion-MNIST images.

Aggregate behaviour. For each α we compute the classifier accuracy and the mean Stein signal obtained by aggregating the TASTE residuals. As shown in Figure 9, both quantities vary monotonically with the corruption level: accuracy declines as Fashion-MNIST content increases, while the Stein score increases in magnitude.

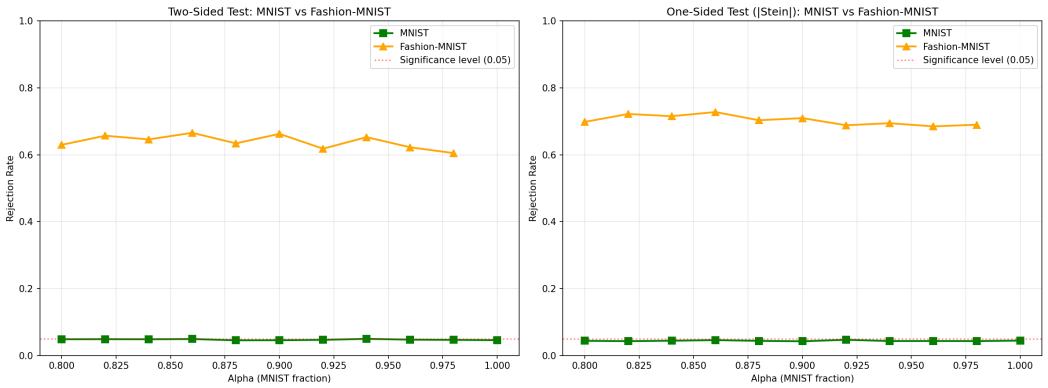


Figure 11: **Stein score-based test power comparison as a function of test set corruption level α.** The left panel presents results for two-sided tests, while a one-sided (arguably more powerful) test is presented on the right. Also, see **Appendix C** for power analysis based on a first-order operator

Per-sample separation. For a fixed corruption level (e.g. $\beta = 0.8$), the distribution of $r_f(x)$ is sharply concentrated near zero for MNIST and displays heavy tails for Fashion-MNIST (Figure 10). This separation enables a simple calibrated test: using held-out MNIST data, estimate the 95th percentile $\tau_{0.95}$ of $|r_f(x)|$ and declare a test point OOD if $|r_f(x)| > \tau_{0.95}$. The resulting test achieves high power across all corruption levels (Figure 11) despite requiring no supervision on Fashion-MNIST.

H First-Order Stein Operator Diagnostics

For a function $f \in \mathcal{F}(p)$, instead of the Langevin operator $\mathcal{L}_p f(x) = \Delta f(x) + \nabla \log p(x)^\top \nabla f(x) \in \mathbb{R}$, consider a first-order Stein operator defined as

$$\mathcal{A}_p f(x) = \nabla f(x) + f(x) s_p(x) \in \mathbb{R}^d.$$

To obtain a scalar diagnostic one typically projects this vector onto a fixed direction v, which we denote by

$$\mathcal{A}_p^v f(X) = v^\top \nabla f(x) + f(x) v^\top s_p(x) \tag{19}$$

It is easy to show that $\mathbb{E}_p[\mathcal{A}_p^v f(X)] = 0$ under mild assumptions on f. Indeed, let $X \sim p$ and assume $f \in \mathcal{F}(p)$ is such that the following integration by parts is valid (e.g., fp vanishes at infinity). Then

$$\mathbb{E}_p\left[\mathcal{A}_p^v f(X)\right] = \mathbb{E}_p\left[v^\top \nabla f(X)\right] + \mathbb{E}_p\left[f(X)\, v^\top s_p(X)\right]. \tag{20}$$

Since $s_p = \nabla \log p$, we have $f\, s_p = \nabla(fp)/p - \nabla f$, and thus

$$\begin{aligned}\mathbb{E}_p\left[f(X)\, v^\top s_p(X)\right] &= \int v^\top f(x) \nabla \log p(x)\, p(x)\, dx \\ &= \int v^\top \nabla(f(x)p(x))\, dx - \int v^\top \nabla f(x)\, p(x)\, dx.\end{aligned}$$

Under the assumed boundary conditions, the first integral vanishes, yielding

$$\mathbb{E}_p\left[f(X)\, v^\top s_p(X)\right] = -\mathbb{E}_p\left[v^\top \nabla f(X)\right].$$

Substituting back into (20), the two terms cancel and we obtain

$$\mathbb{E}_p[\mathcal{A}_p^v f(X)] = 0.$$

We now consider q, another smooth density on $\mathbb{R}^d$, absolutely continuous with respect to p. Mirroring the argument from the main text, we obtain

$$\mathbb{E}_q[\mathcal{A}_p^v f(X)] = -\,v^\top \mathbb{E}_q\left[f(X)\nabla \log \frac{q(X)}{p(X)}\right].$$

Indeed, taking expectation under q yields

$$\mathbb{E}_q[\mathcal{A}_p^v f(X)] = v^\top \mathbb{E}_q[\nabla f(X)] + v^\top \mathbb{E}_q[f(X)\nabla \log p(X)].$$

Using integration by parts under q and the identity

$$\mathbb{E}_q[\nabla f(X)] = -\,\mathbb{E}_q[f(X)\nabla \log q(X)]\,,$$

which holds for f in the Stein class, we obtain

$$\mathbb{E}_q[\mathcal{A}_p^v f(X)] = -\,v^\top \mathbb{E}_q\big[f(X)\big(\nabla \log q(X) - \nabla \log p(X)\big)\big]\,.$$

Rearranging terms gives

$$\mathbb{E}_q[\mathcal{A}_p^v f(X)] = -\,v^\top \mathbb{E}_q\left[f(X)\nabla \log \frac{q(X)}{p(X)}\right],$$

as claimed.

Thus this functional detects only *value-based alignment*: between the output $f(X)$ and the score mismatch $\nabla \log(q/p)(X)$. In contrast to the Langevin operator, which probes geometric distortions via $\nabla f(X)$ and $\Delta f(X)$, the first-order functional reacts primarily to changes in the distribution of the scalar outputs $f(X)$. This might lead to unexpected behaviour.

H.1 Directional blind spots: the 2D Gaussian illustration

To illustrate the potential blind spots, consider the two-dimensional Gaussian mean-shift model introduced in Section 5.1 of the main text

$$p(x) = \mathcal{N}(0, I_2), \qquad q(x) = \mathcal{N}(\varepsilon u, I_2), \qquad u = (\cos\theta, \sin\theta).$$

For the task $f(x) = x_2 - x_1$, the scalar first-order functional admits the exact expression

$$\mathbb{E}_q[\mathcal{A}_p^v f(X)] = -\varepsilon^2\big(v_1\cos\theta + v_2\sin\theta\big)\,(\sin\theta - \cos\theta).$$

For $v = (1, 1)$ this expression simplifies to

$$\mathbb{E}_q[\mathcal{A}_p^v f(X)] = \varepsilon^2\cos(2\theta)$$

which corresponds to taking the component-wise average of the operator.

Thus the response is *quadratic* in the shift magnitude and vanishes for all $\theta = \pi/4, 3\pi/4, \ldots$, creating entire undetectable families of distributional shifts (see Figure 12). By contrast, the Langevin Stein operator responds linearly in ε and exhibits no such degeneracies.

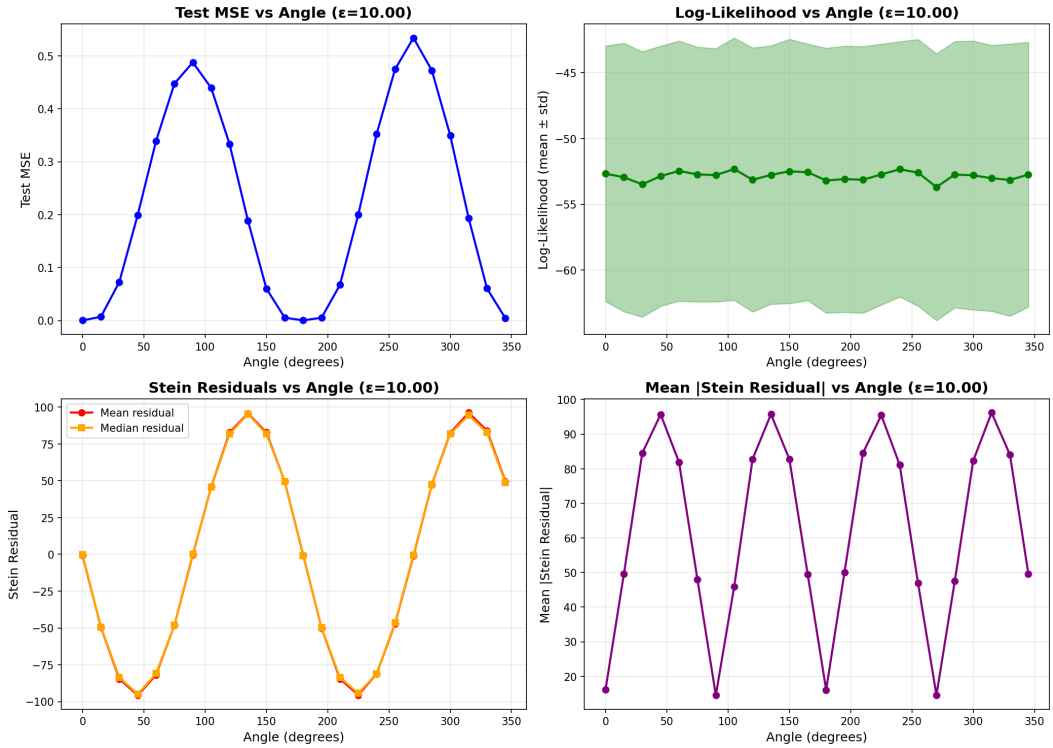


Figure 12: **Directional evaluation of model sensitivity and OOD scoring under the first-order Stein operator** (19)**.** This figure should be compared with **Figure 2**. The signal generated by the (analog of the) TASTE functional for the first-order Stein operator no longer aligns with performance.

H.2 The L^2 norm of the first-order Stein operator (no blind spots, but not a Stein operator)

Another way to obtain a scalar quantity - and perhaps a more natural one - is to consider the squared L^2 norm of the operator under the test distribution q:

$$\mathcal{T}_f(p,q) := \mathbb{E}_q\Big[\,\big\|\mathcal{A}_p f(X)\big\|_2^2\Big] = \mathbb{E}_q\big[\|\nabla f(X) + f(X)s_p(X)\|_2^2\big]\,. \tag{21}$$

No blind spots. Unlike the linear functional $\mathbb{E}_q[\mathcal{A}_p^v f]$, which can vanish for large families of shifts (e.g. whole cones of directions in the 2D Gaussian example), the quadratic quantity (21) is much harder to drive to zero. Thus the L^2 norm of the full first-order field does not suffer from the same directional "blind spot" issues as the projected linear functional.

Nonlinearity and the loss of the Stein property. A key drawback is that $\|\mathcal{A}_p f\|_2^2$ is *not* linear in f or in the operator. Therefore,

$$\mathbb{E}_p\big[\|\mathcal{A}_p f(X)\|_2^2\big] \neq 0$$

in general. In fact, for any nontrivial f in the Stein class of p,

$$\mathbb{E}_p\big[\|\mathcal{A}_p f(X)\|_2^2\big] = \mathrm{Var}_p(\mathcal{A}_p f(X)) > 0.$$

Thus $\mathcal{T}_f(p,q)$ is *not* a Stein operator: it does not vanish under the null hypothesis that $q = p$ and therefore cannot be used directly for discrepancy testing. To convert it into a usable shift-sensitive statistic, we apply the same principle used for correcting misspecified score models: subtract the empirical baseline under the training distribution. Under $q = p$, this corrected statistic concentrates near zero, although this is more of a heuristic and does not have the same theoretical properties as a proper Stein operator. Under $q \neq p$, we see that it typically becomes positive because the magnitude of the first-order field changes whenever the distributional shift interacts with either f or s_p.

Informally, unlike the linear first-order TASTE functional, the squared-norm statistic captures *both magnitude and directional* changes in $\nabla f + f s_p$, thus eliminating blind spots.