FROM DENSITY TO DETECTION: DIAGNOSING LIKELIHOOD FAILURES AND RECOVERING OOD SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Likelihood-based out-of-distribution (OOD) detection assumes that high model density provides evidence of domain membership. Simple OOD images challenge this assumption across flow and autoregressive models. We examine the resulting failures at the level of individual-image decisions. Within-ID rankings reveal strong likelihood–compressibility alignment in Glow, PixelCNN++, and iGPT. Equal-entropy controls separate whole-image distribution entropy from visual complexity; training from scratch on complexity-filtered images then shows that excluded images outrank allowed images in 97.17% of Glow comparisons. Noise-only training further reveals structured high-score preferences under a known population density. These controls distinguish limitations of an oracle-density rule from errors in the density learned by a model. We develop the consequences for detection through an L–J–F decomposition: representation density, local volume change, and conditional variation along representation fibers. This framework extends the flow decomposition to general noninvertible networks and motivates learning from their component responses. Across 40 classifiers, the resulting LJF detector improves AUROC over its LF ablation by 2.01, 5.00, and 1.15 percentage points on CIFAR-10, CIFAR-100, and ImageNet. Adding geometry also improves a strong 27-score fusion in 212 of 229 model–task comparisons under source-matched calibration. The findings clarify why raw density can mislead OOD decisions and show how its underlying model responses retain useful detection information.

1 INTRODUCTION

Likelihood provides a principled objective for distribution fitting. Using it to detect out-of-distribution (OOD) images asks an additional question: does a high-scoring image belong to the intended domain? Models trained on natural images often assign higher likelihood to street-view digits or nearly uniform images (Nalisnick et al., 2019a). These observations have motivated explanations based on entropy, typicality, background statistics, image complexity, and geometry (Ren et al., 2019; Serrà et al., 2020; Caterini & Loaiza-Ganem, 2022; Kamkari et al., 2024). Each explanation supports a particular inference about the score and its usefulness for detection.

Mechanistic analyses already explain important cases. For coupling-layer flows, Kirichenko et al. (2020) identify local pixel correlations and broadly applicable image-to-latent transformations as drivers of OOD failure; architectural changes that favor semantic structure improve detection. Our starting point is the recurrence of simplicity-aligned rankings across different generators, including PixelCNN++ and iGPT. The alignment is visible among individual held-out ID images and therefore precedes any cross-dataset comparison. Related studies connect likelihood to compression, local dependence, and visual detail (Serrà et al., 2020; Umapathi et al., 2023; Osada et al., 2024; Karczewski et al., 2025). We examine the implications for OOD decisions across these model families, building on the existing model-specific explanations.

We organize the analysis around successive detection questions. An oracle-density example establishes the distinction between high density and domain membership. An exact equal-entropy comparison separates the entropy of an image distribution from the complexity of its images. A training-set intervention then removes every image in a specified simple-image region, including the

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.

effects of augmentation: excluded images still outrank allowed held-out images. Finally, fixed-pool and fresh-noise training expose model-induced score preferences under a known uniform population. Together these controls distinguish properties of the intended distribution from properties introduced by fitting the model.

This diagnosis suggests two practical operations: improving the density and calibrating its decision rule, and retaining the model's component responses for discriminative calibration. The latter extends naturally to general noninvertible networks. The coarea formula separates representation density L, local volume J, and the conditional density F of inputs sharing a representation. The third term explains what a low-dimensional output leaves unresolved and motivates an estimable residual feature score. We refer to the resulting vector as a model's *component footprint*.

Classifiers provide a demanding test of this general construction because a broad range of established OOD scores is available for comparison. Across 40 classifiers and 229 model–task cells, learned LJF improves its LF ablation across six density-estimator families, while Jacobian information also improves a strong fusion of 27 score families. The paper thus connects controlled evidence about misleading likelihoods to a detector that learns how to use the responses responsible for those scores.

2 The detection question and the quantities involved

We study OOD decisions about one image or a few images. The task specifies which domain changes should trigger rejection. Let P be the intended ID distribution and let Q describe images designated OOD by the task. For example, valid ID images can remain acceptable when their class frequencies change or when only one such image is observed. Natural-image benchmarks specify a domain distinction, while our support-restricted experiment specifies membership directly. Uniform-noise experiments serve a separate role: the true density is known, so they diagnose the model's scoring behavior.

Let m be the fitted probability model and $S(x) = \log m(x)$ its score. We use native log probabilities for discrete autoregressive models and explicitly specified dequantized log densities for flows. For detection, larger scores are initially taken to favor ID. AUROC is the probability that a randomly drawn ID image scores above an OOD image, with half credit for ties. In diagnostic experiments we also report its complement as the plain-language quantity *fraction of comparisons in which the tested image receives higher likelihood*. For a calibrated rejection rule, OOD acceptance means the fraction of OOD images the rule accepts as ID; we report actual ID retention alongside it. Main-text results are means; run and split details appear in the appendix.

Three notions of "entropy" or complexity must be kept separate. The entropy of a distribution over entire images is

$$H_{\mathrm{img}}(P) = -\sum_{x} P(x) \log P(x), \tag{1}$$

with differential entropy used only for a stated continuous representation. For one fixed image x, pixel-histogram entropy is $H_{\mathrm{pix}}(x) = -\sum_{v} h_x(v) \log h_x(v)$, where h_x is its normalized intensity histogram. This describes pixel-value variability within the fixed image and ignores spatial arrangement. JPEG length $C(x)$ measures lossy compressibility under a fixed encoder and is sensitive to spatial arrangement. We measure simplicity alignment with $c(x) = -\log C(x)$. Claims about H_{img} therefore require a specified distribution over whole images.

The distinction also clarifies the starting observation. On the same 10,000 held-out CIFAR-10 images, archived Glow, PixelCNN++, and iGPT likelihood scores have Spearman correlations of 0.901, 0.903, and 0.727 with compressibility (Appendix B). Our independently trained full-data Glow runs give a mean correlation of 0.924 (Figure 1a). Thus a common ranking preference appears within ID across both flow and autoregressive architectures, motivating the controlled tests below.

3 Revisiting explanations of high-likelihood OOD images

3.1 Would an oracle density solve the problem?

Suppose first that the model is exact: $m = p$. Its density measures local concentration under P, while an OOD decision concerns domain membership. For a comparison distribution Q with finite entropy

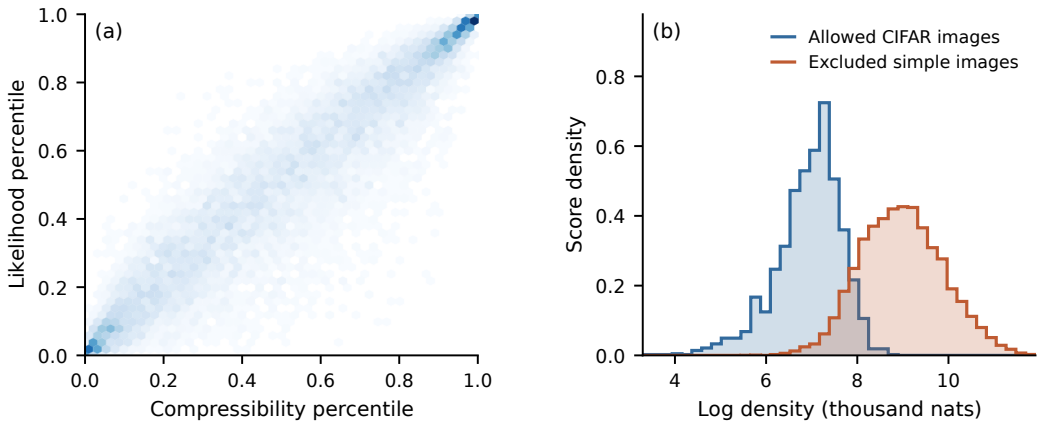


Figure 1: The same scoring preference appears within ID and across an explicit membership boundary. (a) Compressibility versus likelihood percentiles on 10,000 held-out CIFAR images; plotted likelihood percentiles average the full-data Glow runs, whose mean within-ID Spearman correlation is 0.924. (b) After training only on complexity-filtered images, excluded simple images still receive higher scores than allowed held-out images. Each histogram is normalized separately for each model and image group, then averaged across runs using common bins.

and cross-entropy under the same reference measure,

$$\mathbb{E}_Q \log p - \mathbb{E}_P \log p = H_{\text{img}}(P) - H_{\text{img}}(Q) - D_{\text{KL}}(Q\|P). \tag{2}$$

A sufficiently concentrated Q can therefore achieve higher mean log density. For example, if $P = \mathcal{N}(0, I_d)$ and $Q = \mathcal{N}(0, \sigma^2 I_d)$ with $0 < \sigma < 1$, the difference is $d(1 - \sigma^2)/2 > 0$. This illustrates the familiar distinction between high density and typical probability mass under an exact model (Nalisnick et al., 2019b; Caterini & Loaiza-Ganem, 2022; Le Lan & Dinh, 2021). It concerns distinguishing sampling distributions; whether their overlapping individual samples should be rejected requires an additional task definition.

Explaining a learned model's high OOD score requires a further check: does that score agree with its intended target? Zhang et al. (2021) explicitly question whether the overlap assumptions underlying a typical-set account describe common image OOD tasks and emphasize model misestimation. The next controls test this agreement directly, after accounting for the oracle limitation.

3.2 Does a visually simple dataset necessarily have lower entropy?

Caterini & Loaiza-Ganem (2022) discuss lower whole-image entropy as an explanation and use visual complexity as a heuristic for its direction. The relevant quantity is H_{img}, the entropy of the distribution over whole images. We define R as uniform over 10,000 distinct training images and their flips, and Q as uniform over 10,000 distinct constant-color images. Both have $H_{\text{img}} = \log 10{,}000$, although the color images have much smaller pixel-histogram entropy and compression length. Under the filtered native autoregressive models, colors nevertheless receive a higher mean log probability by 5,115.85 nats per image; AUROC is only 0.037%. For any positive discrete model,

$$\mathbb{E}_Q \log m - \mathbb{E}_R \log m = D_{\text{KL}}(R\|m) - D_{\text{KL}}(Q\|m) \quad \text{when } H_{\text{img}}(R) = H_{\text{img}}(Q). \tag{3}$$

This elementary use of the entropy–KL identity (Caterini & Loaiza-Ganem, 2022) makes the inference precise: this score gap cannot come from a difference in whole-image distribution entropy. The finite construction fixes both entropies exactly and supplies a counterexample to the implication "visually simpler, therefore lower distribution entropy."

3.3 Are simple OOD images merely in a high-density, atypical region?

A distribution can place high density on a small-volume region that is rarely sampled. This explains why high likelihood and infrequent generation need not conflict (Nalisnick et al., 2019b; Kamkari

et al., 2024). But for this account to describe an accurate model, the target must actually assign probability to that region. We test this prerequisite by *changing the training set and specifying the corresponding allowed region*, so that membership in this controlled task is explicit.

Construction. We retain the 5,000 CIFAR training images with largest JPEG lengths, with deterministic tie-breaking, and train every model from random initialization. The selection boundary is $\tau = 10{,}464$ bits. Because training uses horizontal flips T, the allowed image region is

$$A = \{x : C(x) \geq \tau \text{ or } C(Tx) \geq \tau\}. \tag{4}$$

All training images and their augmentations lie in A. The controlled ID task permits unseen images in A but excludes A^c: any target distribution for this task must satisfy $P_A(A^c) = 0$. This support condition defines the filtered task; the original unfiltered CIFAR distribution remains a separate reference. A test image below threshold in *both* orientations is therefore excluded even after accounting for augmentation. The test comparison contains 987 original-orientation allowed images and 9,006 definitely excluded images; seven flip-only allowed images are omitted. For flows, the same support restriction is imposed on the union of dequantization bins corresponding to A; excluded images are evaluated within the complementary bins.

Controls. A random 5,000-image training subset matches the filtered subset's class counts. Architecture, initialization seeds, optimizer, augmentation, and update budget are paired. No full-data pretraining is used, and data-dependent initialization also uses only the selected training subset. The fixed criterion therefore excludes simple images throughout training, starting from initialization. The fixed criterion defines the region tested.

Result and implication. After filtering, excluded images still outrank allowed held-out images in 97.17% of Glow comparisons. A native discrete autoregressive model gives 94.46%, while its allowed held-out bpd improves from 4.650 to 4.305 between 1,000 and 12,000 updates. Eightfold longer Glow training retains a 96.48% excluded-image win rate. The effect therefore extends beyond continuous dequantization and persists well beyond the initial fit. Figure 1b shows the separated score distributions.

The *reversed ranking* is the key finding. These images cannot be described as a high-density tail of the intended P_A: that target excludes them. They can be an unusual region of the *learned model*, but then the account already involves model mismatch. This reconciles a useful typicality test with an inaccurate density model. A two-sided test may reject the high scores, without making those scores correct estimates of the intended distribution.

3.4 Does the preference require natural-image training examples?

Filtering retains natural-image structure, so the previous result does not determine whether observing such structure is necessary for the high-score pattern. We next train Glow on noise alone. The principal finite-data condition uses a fixed pool of 5,000 independently sampled uniform-noise images, with repeated training and augmentation, as in ordinary dataset training. An additional condition draws fresh uniform noise at each update. The generating population is uniform on the pixel cube; after ideal uniform dequantization its log density is constant. Natural and constant-color images have no oracle-density advantage over any other point in that cube.

Table 1: High scores without natural-image training. Final fixed-pool and fresh-noise Glow results. Each percentage is the chance that the named image type receives higher likelihood than an independently drawn held-out noise image, averaged across training runs. A flat oracle gives 50%. These are score diagnostics, not a claim that these images lie outside the uniform target's support.

Training data	Noise test bpd	SVHN	CIFAR	Gray	Constant RGB
Fixed noise pool	11.049	97.63%	90.06%	93.65%	96.10%
Fresh noise	8.030	86.28%	66.27%	80.79%	59.82%

The fixed-pool model strongly favors SVHN and simple color images over held-out noise (Table 1). SVHN's mean log density exceeds held-out noise by $4{,}993.79$ nats. Training-noise bpd is 7.743,

versus 11.049 on held-out noise, so finite-sample overfitting and the relative preference coexist. The comparison uses independent noise from the generating population. Between 5,000 and 62,400 updates, both SVHN and held-out noise lose mean log density, but noise loses more; SVHN's growing relative advantage therefore reflects the larger decrease for noise. Thus finite-data training can produce a selective response absent from the generating population density.

Fresh-noise training shows that reusing a finite training pool is not required for every part of the pattern. As held-out noise bpd improves from 8.085 at 5,000 updates to 8.030 at 62,400, the fraction of SVHN–noise comparisons favoring SVHN rises from 27.52% to 86.28%; its final mean log-density advantage is 23.42 nats. The remaining whole-image fitting error is about 64 nats, despite the small bpd gap. This known-target example sharpens the observation that better density fitting need not improve OOD ranking (Zhang et al., 2021).

3.5 When can density still provide useful OOD information?

The preceding failures do not make likelihood universally uninformative. Consider the illustrative decomposition $S(x) = a(h(x)) + b(c(x))$, where a captures task-relevant structure and b a nuisance preference such as compressibility. Within a fixed-complexity stratum $c(x) = c_0$, the nuisance contribution is constant, and likelihood ranks images exactly as a. More generally, a task-related score margin larger than the nuisance difference preserves the desired ordering. Thus semantic changes can remain detectable after controlling the dominant preference, consistent with improved detection from flows fitted in semantic feature representations (Kirichenko et al., 2020). This is a sufficient condition for useful density ordering. Matching only the marginal distribution of complexity is weaker than fixing it in each comparison. The practical objective is to retain informative contributions while calibrating or bypassing those that dominate the wrong comparisons.

4 From misleading likelihoods to calibrated detection

The controls motivate two operations: improving the probability model and its calibrated rejection rule, and learning a detector from component responses. The first uses ID data; discriminative component calibration additionally uses labeled OOD examples.

4.1 Improving the density and calibrating both score tails

For an invertible flow f with base density b, the exact score is

$$S(x) = L(x) + J(x), \qquad L(x) = \log b(f(x)), \quad J(x) = \log|\det Df(x)|. \tag{5}$$

The Gaussian base describes encoded model samples; real ID data have pushforward $f_{\#}P$. Under change of variables, $D_{\mathrm{KL}}(P\|M) = D_{\mathrm{KL}}(f_{\#}P\|B)$ (Papamakarios et al., 2021), motivating recalibration of the encoded ID distribution.

We adapt radial recalibration (Dheur & Ben Taieb, 2025). Let t be the log squared radius of the whitened latent vector and let $a_0(t)$ be its density under the model's Gaussian base. Replacing this one-dimensional marginal by an ID-fitted normalized density $\widehat{a}(t)$ gives

$$\widetilde{m}(x) = m(x)\frac{\widehat{a}(t(x))}{a_0(t(x))}. \tag{6}$$

The corrected model remains normalized on its original continuous domain. We retain both fixed Gaussian and Gaussian-KDE estimators. We evaluate fitting and rejection separately.

For either S or $\log\widetilde{m}$, an ID-only two-sided rule uses the lower and upper empirical tail probabilities, with conservative finite-sample ranks. If $p_-(x)$ and $p_+(x)$ are those ranks, define $p_2(x) = \min\{1, 2\min(p_-(x), p_+(x))\}$ and accept when $p_2(x) > \alpha$. Under exchangeability and an independently fixed score, the rule controls marginal ID rejection at α; its OOD power depends on the score distributions (Bates et al., 2023). It can nevertheless exploit an OOD population in either score tail. This connects typicality-based ideas to individual-image testing.

4.2 JACOBIAN CANCELLATION AND COMPONENT INFORMATION

Likelihood ratios address a different question from raw likelihood: which of two specified distributions better explains the *same observed image*? Expressing both complete densities through the same invertible map yields

$$\frac{q(x)}{p(x)} = \frac{q_Z(f(x))|\det Df(x)|}{p_Z(f(x))|\det Df(x)|} = \frac{q_Z(f(x))}{p_Z(f(x))}. \tag{7}$$

The common coordinate factor cancels, consistent with the coordinate invariance of likelihood ratios (Kasieczka et al., 2023). A related strategy uses separately trained models to approximately remove shared background statistics (Ren et al., 2019).

Equation (7) retains the full alternative latent density q_Z, which can encode information discarded by the scalar ID score L. For a simple illustration, on $(0, 1)$ take $p(x) = 2x$, $q(x) = 1$, and $f(x) = x^2$. Then $p_Z = 1$, so $L = 0$, while $J = \log(2x)$ and $\log(q/p) = -J$. The common Jacobian cancels algebraically, yet the Jacobian statistic alone determines the likelihood ratio. Its information is encoded in q_Z. This illustrates the additional information a component can retain relative to a compressed scalar score. Existing joint-statistic and likelihood-free detectors already exploit informative component responses (Morningstar et al., 2021; Ahmadian & Lindsten, 2021).

4.3 GENERAL NETWORKS: THE L–J–F DECOMPOSITION

Three contributions to density under a neural representation. Let $X \in \mathbb{R}^D$ have a specified Lebesgue density p_X, and let $z = g(x) \in \mathbb{R}^r$, $r \le D$, be a neural representation. Assume g is locally Lipschitz and has row rank r almost everywhere under p_X. With $\mathcal{J}_g(x) = \sqrt{\det(Dg(x)Dg(x)^\top)}$, the coarea formula decomposes density using the representation and its fibers. Let $\rho(x \mid z)$ be the conditional density on the fiber $g^{-1}(z)$, normalized with respect to $(D - r)$-dimensional Hausdorff measure (Appendix F.1). The fiber consists of inputs sharing representation z. Thus, almost everywhere where the densities are positive,

$$\log p_X(x) = \underbrace{\log p_Z(g(x))}_{L(x)} + \underbrace{\log \mathcal{J}_g(x)}_{J(x)} + \underbrace{\log \rho(x \mid g(x))}_{F(x)}. \tag{8}$$

This extends the familiar flow decomposition to many-to-one representations using established change-of-variables geometry (Cvitkovic & Koliander, 2019). L measures representation frequency, J local transverse volume change, and F variation among inputs that the representation merges. The chosen p_X may be a population density or a normalized generative model; a classifier supplies the representation g, and the identity applies to a specified p_X.

Why the third term matters. For a bijective flow, each fiber contains one input and its conditional counting-measure density is one: $F = 0$, recovering $\log p_X = L + J$. For a many-to-one network, the conditional density along each fiber remains to be specified. Consider $g(x_1, x_2) = x_1$ and $p_X(x_1, x_2) = \varphi(x_1)\varphi(x_2)$, where φ is the standard normal density. The points $(0, 0)$ and $(0, 3)$ have identical $L = \log \varphi(0)$ and $J = 0$, while their $F = \log \varphi(x_2)$ values differ by 4.5 nats. Their density ranking is entirely carried by the discarded coordinate. This motivates retaining a statistic of information discarded by the readout alongside its density and geometry.

Estimating the components in a frozen network. We instantiate this principle with classifiers, whose broad set of existing OOD scores enables controlled comparisons. Write $h(x) \in \mathbb{R}^d$ for the feature and $z(x) = Wh(x) + b \in \mathbb{R}^r$ for a specified full-row-rank affine readout. CIFAR uses all logits; ImageNet uses a fixed 32-dimensional projection. We estimate geometry by J_ϵ, the log singular-volume statistic of $D_x z$ with fixed relative ridge 10^{-6} (Appendix F). Let Q have orthonormal rows spanning the row space of W, and let N have orthonormal columns spanning $\ker W$. From ID training features, estimate mean μ and ridged covariance Σ, and form

$$K = \Sigma Q^\top (Q\Sigma Q^\top)^{-1}, \qquad u(x) = N^\top (I - KQ)(h(x) - \mu). \tag{9}$$

u retains the feature-fiber coordinate after removing linear predictability from the readout. Our primary estimators are

$$\widehat{L}(x) = \log \sum_c \widehat{\pi}_c \mathcal{N}(z(x); \widehat{\mu}_c, \widehat{\Sigma}_Z), \qquad \widehat{F}(x) = \log \mathcal{N}(u(x); \widehat{\mu}_U, \widehat{\Sigma}_U). \tag{10}$$

The class priors are empirical; readout components share a covariance. All fits use ID training data. The unconditional residual Gaussian estimates a practical approximation to variation discarded by the readout, using the learned feature space. The exact F in Eq. (8) is conditional on the input fiber; Appendix F.1 gives the connection between these two levels. A sixth, self-consistent control instead derives the readout marginal and feature-fiber conditional from one fitted feature mixture. Its affine identity is exact for that model. Together with the primary choice, diagonal GMMs with 10/50 components, PCA–GMM10, and 50-NN scores give six fixed estimator families (Appendix F). In empirical tables, L and F denote the fitted statistics.

From components to a decision. We call the calibrated detector LJF and use the ordered vector $\phi(x) = (\widehat{L}(x), J_\epsilon(x), \widehat{F}(x))$. Logistic regression learns the signs and weights of standardized components from ID and OOD calibration examples; LF uses $(\widehat{L}, \widehat{F})$ under the same protocol. Calibration and test images are disjoint samples from the same named ID/OOD sources. For flows the corresponding component vector is (L, J). This common design tests whether component responses improve decisions beyond their fixed density sum; the classifier implementation uses fitted component statistics. The comparisons address two levels: LJF versus LF isolates the contribution of geometry within each fixed estimator family, while adding J_ϵ to an established score fusion tests whether this signal remains useful beside existing detectors.

5 Empirical detection results

5.1 Evaluation protocol

We evaluate general-network component calibration on classifiers, using their extensive OOD baseline literature. The cohort contains 17 CIFAR-10 models with six OOD targets, 12 CIFAR-100 models with six, and 11 ImageNet models with five: 40 models and 229 model–task cells. Each calibration repetition splits cached ID and OOD examples equally into calibration and test halves; ID indices are shared across sources. We average three calibration repetitions within each cell, then cells within each benchmark. *Source-matched* evaluation uses disjoint images from the same OOD source for calibration and testing.

ALL fits the same separator to one source-selected score from each of 27 implemented families, including energy, Mahalanobis, ViM, ReAct, and GradNorm (Liu et al., 2020; Lee et al., 2018; Wang et al., 2022; Sun et al., 2021; Huang et al., 2021). Method settings and score signs are selected only on source calibration data. Component ablations and this stronger fusion establish distinct levels of utility. AUROC and FPR95 use the usual higher-is-ID direction, and all table entries are percentages. FPR95 is the OOD acceptance rate at an evaluation threshold retaining at least 95% of test ID images. Deployment-style acceptance below uses a separate calibration threshold and reports actual ID retention.

5.2 Classifier components yield consistent calibrated gains

With the prespecified tied-Gaussian proxies, learned LJF improves AUROC over learned LF by 2.01, 5.00, and 1.15 points on CIFAR-10, CIFAR-100, and ImageNet; FPR95 falls by 6.25, 14.38, and 3.89 points (Table 2). Across six prespecified proxy families, all 18 benchmark-level AUROC differences favor LJF, with 16 exploratory intervals above zero; all 18 FPR95 intervals favor LJF. The diagonal ten-component GMM, for example, gains 9.58 AUROC points on CIFAR-100 relative to its LF ablation. Appendix G reports the complete comparison.

Calibration determines how the components contribute. With a fixed $+1$ coefficient for J_ϵ, the raw summed CIFAR-100 AUROC changes from 63.56% to 63.18%, while calibrated LF and LJF attain 77.30% and 82.30%. The model's response supplies the information; the detector must learn how to use it. A self-consistent conditional feature model yields the same qualitative gain, supporting the decomposition-based interpretation across estimators.

5.3 Additional information beyond strong score fusion

Figure 2 places these methods beside named baselines and the ALL fusion; Table 2 supplies both detection metrics. ALL+J_ϵ improves matched AUROC in 212 of 229 cells. The overall increment

Table 2: Source-matched classifier detection. Entries are mean AUROC (↑) / FPR95 (↓), in percent, over 102, 72, and 55 model–target cells. Each cell first averages three calibration splits. LF/LJF use the prespecified tied-Gaussian proxies; fixed sums use their raw scores. The single baseline, its grid setting, and its sign are selected on source calibration data. ALL combines one selected score from each of 27 families. All learned rows use the same source-calibration protocol; no test-set method selection is used.

	CIFAR-10		CIFAR-100		ImageNet	
Method	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95
Fixed $L + F$	81.53	70.82	63.56	89.66	76.24	67.30
Fixed $J_\epsilon + L + F$	80.84	71.17	63.18	89.11	76.38	66.91
Learned LF (2 features)	86.86	60.60	77.30	79.37	79.42	65.13
Learned LJF (3 features)	88.87	54.35	82.30	64.99	80.57	61.24
Source-selected single baseline	90.65	45.66	82.94	66.86	88.13	47.11
ALL: 27-score fusion	93.15	33.00	87.99	49.40	92.78	29.44
ALL + J_ϵ	93.63	31.62	88.44	47.44	92.93	28.96
ALL + shape	93.38	31.55	88.49	47.67	92.92	28.96
ALL + scale and shape	93.79	30.49	88.86	46.06	93.06	28.49

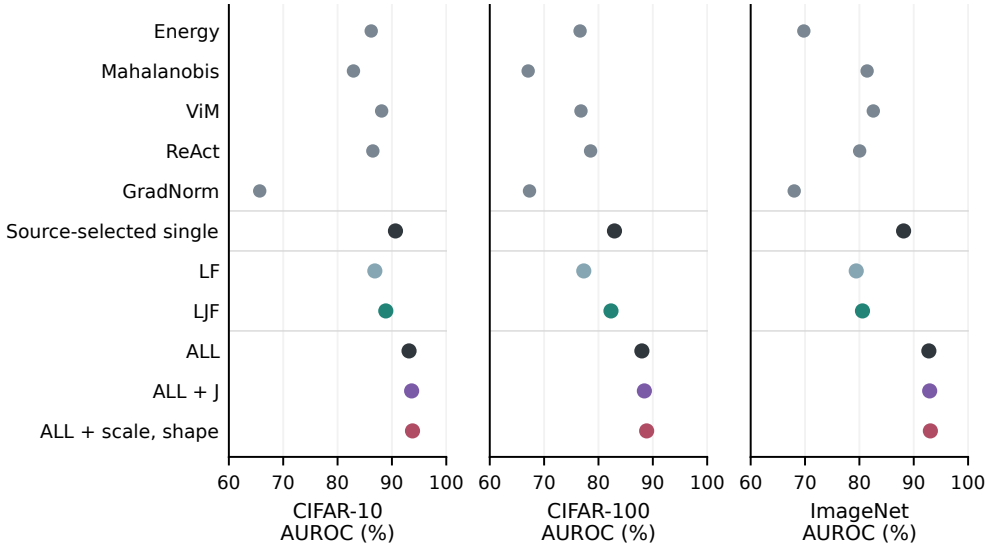


Figure 2: Component calibration alongside strong published baselines. Absolute AUROC on three ID benchmarks under the same source-matched calibration splits. Named baselines select settings and score direction using calibration data; the single-baseline row selects among all 27 families. ALL fuses their scores, and ALL+J adds geometry. Points average three splits within each model–target cell, then all cells.

is 0.393 points, with exploratory 95% interval $[0.252, 0.557]$; FPR95 decreases by 1.347 points. Every benchmark mean improves in both metrics. Duplicate-score and random-feature augmentation controls do not account for this increment under the same fitting pipeline.

Separating geometry into log-RMS scale and regularized spectral shape improves the matched result further. ALL plus both reaches AUROC 93.79%, 88.86%, and 93.06%, with FPR95 30.49%, 46.06%, and 28.49%. Its pooled AUROC is 92.06%, compared with 87.62% for the source-selected single-score baseline and 91.44% for ALL. Both summaries come from the same computed Jacobian spectrum. Separating the fixed linear combination $J_\epsilon = rA + B_\epsilon$ lets the detector weight scale and spectral shape independently.

Table 3: Generative counterpart on full-data face Glow models, tested on the same held-out SVHN images. Center dequantization; arithmetic means of two training seeds. All entries are percentages. Acceptance uses a disjoint ID calibration split at $\alpha = .05$; measured ID retention is shown. ID-only methods and supervised source calibration have different information budgets and are separated in their labels.

Detector and calibration source	AUROC ↑	OOD accepted ↓	ID retained
Raw S (ID-only)	6.56	99.90	95.06
Two-sided S (ID-only)	88.54	38.33	95.37
KDE(S) (ID-only)	88.78	38.33	95.36
DoSE(L, J) (ID-only)	87.27	44.87	94.96
LR(S), SVHN source	93.44	28.65	95.10
LR(L, J), SVHN source	95.40	17.91	95.08

5.4 Generative models: density correction and component calibration

The CelebA face-model experiments provide the exact flow counterpart. Raw likelihood largely accepts SVHN despite training on faces: AUROC is 6.56% and OOD acceptance is 99.90%. Two-sided ID-only calibration raises AUROC to 88.54% and lowers acceptance to 38.33%, retaining 95.37% ID. On the finite equal-entropy autoregressive task, the corresponding color acceptance drops from 100% to 0.19%, at 94.10% measured retention of the specified image distribution. The score distribution therefore supplies useful individual-image detection information.

With representative source OOD examples, calibrating (L, J) further lowers face-model SVHN acceptance from 28.65% for calibrated S to 17.91%, while retaining approximately 95.1% ID. ID-only KDE and DoSE baselines are included with their distinct information budget. The gain from components agrees with the classifier results: preserving how a score is assembled can help after the raw sum is misleading.

Finally, all 24 fixed radial-recalibration fits improve held-out ID log score, by 0.53–38.23 nats per image across settings. Yet full-cohort one-sided SVHN acceptance remains 99.90%, and two-sided acceptance changes from 38.33% to 39.21% or 39.26% for the two radial estimators. Density correction achieves its measured fitting improvement, while decision calibration supplies most of the rejection improvement in these tests. These are complementary operations whose success must be evaluated against their respective objectives.

6 Discussion

The controlled experiments connect distribution-level explanations to individual-image decisions. Equal-entropy comparisons, excluded-support training, and known-density noise training each isolate a different source of high model scores. Together they motivate preserving component information through the final detection step.

Simplicity and generalization. Our working interpretation is that recurring simplicity preferences can arise as a by-product of the inductive biases through which finite-capacity models generalize from natural data. Reusing predictable structure allows a model to assign density beyond observed examples; the same structure can occur especially strongly in simple OOD images. Flow-specific analyses provide concrete mechanisms for this tradeoff (Kirichenko et al., 2020), and the autoregressive results broaden its empirical setting. This is a mechanistic hypothesis informed by the observations. The contribution established here is the shared detection failure and the information retained in model responses.

From density to a decision. LJF extends the component view to noninvertible networks by making the conditional fiber term explicit. Its practical estimators support extensive comparisons on classifiers, where both component ablations and strong fused baselines show gains. The resulting detector demonstrates how useful OOD evidence can survive a misleading scalar density.

AI USE DISCLOSURE

We used large language models (LLMs) to polish the language of the entire paper and to correct grammatical errors. We have not used generative AI tools for generating synthetic data, developing theoretical models or conceptual frameworks, formulating mathematical claims, writing proofs, proposing hypotheses, designing experiments, implementing methods, translation, dataset cleaning, data analysis, or interpreting results. All AI-assisted edits were reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Amirhossein Ahmadian and Fredrik Lindsten. Likelihood-free out-of-distribution detection with invertible generative models. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2119–2125, 2021. doi: 10.24963/ijcai.2021/292. URL https://www.ijcai.org/proceedings/2021/292.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. doi: 10.1214/22-AOS2244. URL https://arxiv.org/abs/2104.08279.

Anthony L. Caterini and Gabriel Loaiza-Ganem. Entropic issues in likelihood-based OOD detection. In *Proceedings on "I (Still) Can't Believe It's Not Better!" at NeurIPS 2021 Workshops*, volume 163 of *Proceedings of Machine Learning Research*, pp. 21–26. PMLR, 2022. URL https://proceedings.mlr.press/v163/caterini22a.html.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20s.html.

Milan Cvitkovic and Günther Koliander. Minimal achievable sufficient statistic learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1465–1474. PMLR, 2019. URL https://proceedings.mlr.press/v97/cvitkovic19a.html.

Victor Dheur and Souhaib Ben Taieb. Multivariate latent recalibration for conditional normalizing flows. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0175. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/079923016ab82f373c0a4a080b2f3ee6-Abstract-Conference.html.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/063e26c670d07bb7c4d30e6fc69fe056-Abstract.html.

Hamidreza Kamkari, Brendan Leigh Ross, Jesse C. Cresswell, Anthony L. Caterini, Rahul Krishnan, and Gabriel Loaiza-Ganem. A geometric explanation of the likelihood OOD detection paradox. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22908–22935. PMLR, 2024. URL https://proceedings.mlr.press/v235/kamkari24a.html.

Rafał Karczewski, Markus Heinonen, and Vikas Garg. Diffusion models as cartoonists: The curious case of high density regions. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/22650d2b38807c88a7226eb1fed31e43-Abstract-Conference.html.

Gregor Kasieczka, Radha Mastandrea, Vinicius Mikuni, Benjamin Nachman, Mariel Pettee, and David Shih. Anomaly detection under coordinate transformations. *Physical Review D*, 107(1): 015009, 2023. doi: 10.1103/PhysRevD.107.015009. URL https://arxiv.org/abs/2209.06225.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/d139db6a236200b21cc7f752979132d0-Abstract.html.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ecb9fe2fbb99c31f567e9823e884dbec-Abstract.html.

Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *Entropy*, 23(12):1690, 2021. doi: 10.3390/e23121690. URL https://www.mdpi.com/1099-4300/23/12/1690.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3232–3240. PMLR, 2021. URL https://proceedings.mlr.press/v130/morningstar21a.html.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019a. URL https://arxiv.org/abs/1810.09136.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. In *4th Workshop on Bayesian Deep Learning at NeurIPS*, 2019b. URL https://bayesiandeeplearning.org/2019/papers/22.pdf. Also available as arXiv:1906.02994.

Genki Osada, Tsubasa Takahashi, and Takashi Nishide. Understanding likelihood of normalizing flow and image complexity through the lens of out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21492–21500, 2024. doi: 10.1609/aaai.v38i19.30146. URL https://ojs.aaai.org/index.php/AAAI/article/view/30146.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. URL https://jmlr.org/papers/v22/19-1028.html.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e79596878b2320cac26dd792a6c51c9-Abstract.html.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1701.05517.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1909.11480.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01894d6f048493d2cacde3c579c315a3-Abstract.html.

Barath Mohan Umapathi, Kushal Chauhan, Pradeep Shenoy, and Devarajan Sridharan. Shaken, and stirred: Long-range dependencies enable robust outlier detection with PixelCNN++. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 1440–1450, 2023. doi: 10.24963/ijcai.2023/160. URL https://www.ijcai.org/proceedings/2023/160.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_ViM_Out-of-Distribution_With_Virtual-Logit_Matching_CVPR_2022_paper.html.

Andi Zhang and Damon Wischik. Falsehoods that ML researchers believe about OOD detection. In *ML Safety Workshop at NeurIPS*, 2022. URL https://arxiv.org/abs/2210.12767.

Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12427–12436. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhang21g.html.

A Probability identities and calibration

A.1 Whole-image entropy, expected scores, and overlap

For densities p, q, m with respect to the same reference measure, suppose the entropies and cross-entropies below are finite. Adding and subtracting the self-entropies gives

$$\mathbb{E}_Q \log m - \mathbb{E}_P \log m = H(P) - H(Q) + D_{\mathrm{KL}}(P\|M) - D_{\mathrm{KL}}(Q\|M). \tag{11}$$

For $m = p$, the first KL term vanishes, yielding Eq. (2). For two finite uniform distributions on equally many distinct images, the entropy difference instead vanishes, yielding Eq. (3). These are standard identities (Caterini & Loaiza-Ganem, 2022); the experiment controls which term can explain the observed gap. The expectation in these identities is taken over *log* probability. The oracle identity must not be used with finite terms when Q assigns positive mass outside the support of P: its oracle cross-entropy is then infinite.

The Gaussian illustration follows directly from $\log p(x) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\|x\|^2$. Under P, the expected squared norm is d, while under Q it is $d\sigma^2$. This establishes an expectation ordering between distributions. At the individual-image level, overlapping sources need not be perfectly distinguishable. The source-testing problem and semantic membership therefore require a stated task, as emphasized by Zhang et al. (2021); Zhang & Wischik (2022).

A.2 A finite-sample two-sided rule

Let S be a score fixed independently of n ID calibration images, and write their scores as $S_1, \ldots, S_n$. For a test image x, define

$$p_-(x) = \frac{1 + \sum_{i=1}^n \mathbf{1}\{S_i \le S(x)\}}{n+1}, \qquad p_+(x) = \frac{1 + \sum_{i=1}^n \mathbf{1}\{S_i \ge S(x)\}}{n+1}, \tag{12}$$

$$p_2(x) = \min\{1, 2\min(p_-(x), p_+(x))\}. \tag{13}$$

A test image is accepted when $p_2(x) > \alpha$. If the calibration and test ID scores are exchangeable, each tail rank is super-uniform, including with ties counted conservatively. Therefore

$$\Pr_{X\sim P}\{p_2(X) \le \alpha\} \le \Pr\{p_-(X) \le \alpha/2\} + \Pr\{p_+(X) \le \alpha/2\} \le \alpha. \tag{14}$$

This union-bound construction controls the ID error rate; see Bates et al. (2023) for the broader conformal testing framework. If a score itself is learned from ID data, those data must be separated from the rank-calibration split. The guarantee is marginal over the calibration and test draw; an observed finite test set can retain less than $1 - \alpha$. In experiments we report its measured retention. AUROC evaluates the calibrated score over all thresholds.

A.3 What component cancellation does and does not establish

Equation (7) follows from applying the same bijection to two complete distributions. It concerns two explanations of the same observed image; comparing $p(x_{\mathrm{ID}})$ with $p(x_{\mathrm{OOD}})$ is a different operation. Coordinate invariance does not imply that a scalar function of the coordinate map has zero association with the detection label. The full latent vector and the Jacobian statistic are deterministically related, but a scalar latent-density score discards most of that vector. This distinction is consistent with existing component-based methods (Morningstar et al., 2021; Ahmadian & Lindsten, 2021). We use the elementary counterexample in Section 4.2 to rule out an invalid implication, in keeping with these component-based precedents.

B Likelihood rankings across generative architectures

We evaluate archived full-data Glow, PixelCNN++, and iGPT models on the same 10,000 CIFAR-10 test identities. The score in every row is the cached full likelihood score, rather than a denoiser or embedding Jacobian. Glow scores a continuous dequantized representation; PixelCNN++ scores its discrete autoregressive image model; iGPT scores its autoregressive image-token representation.

Model-family references are Kingma & Dhariwal (2018); Salimans et al. (2017); Chen et al. (2020). The experiment uses the corresponding implementations and checkpoints in our archive.

JPEG lengths use quality 95, optimization enabled, and no chroma subsampling. All image identities are aligned before computing Spearman correlation with $-\log C(x)$. The negative-length statistic used in the original cache has the same ranks. The correlations below were independently recomputed from the stored likelihood arrays and the image-indexed JPEG table. The independently trained

Table 4: Within-ID likelihood–compressibility alignment across generator families on the same 10,000 CIFAR-10 test images. These are archived individual checkpoints, separate from the paired training runs in the controlled Glow experiments.

Model	Images	Spearman correlation
Glow	10,000	0.901240
PixelCNN++	10,000	0.902571
iGPT	10,000	0.727212

full-data Glow runs used in Figure 1 have a mean correlation of 0.924. Both cohorts show strong within-ID alignment. In a separate fixed-architecture PixelCNN++ data-amount experiment, reducing the training set from all 50,000 CIFAR images to a random stratified 500-image subset retains positive alignment: the mean correlations are 0.678 and 0.707, respectively. Each setting trains from random initialization for 12,000 updates and averages two seeds on a fixed 5,000-image evaluation set; this experiment uses its own quality-90 JPEG proxy. It supplies a data-amount control at fixed capacity. The support-exclusion test in Section 3.3 instead selects images using an explicit complexity boundary.

C Controlled generative experiments

C.1 Complexity measurement and the excluded-support experiment

The compression statistic uses JPEG quality 95, optimization enabled, and chroma subsampling disabled; lengths are in bits. The same encoder is applied to training images, held-out images, and horizontal flips. Every image is represented at 32×32 RGB resolution in these controls. A strictly monotone transformation such as $-\log C$ changes neither the ordering nor Spearman correlation. Its relation to the two entropy quantities is specified in Section 2.

The complex subset is selected from the original 50,000-image CIFAR training pool. The random-subset control has the same 5,000-image size and per-class counts. Data-dependent flow initialization sees only the assigned subset. Glow (Kingma & Dhariwal, 2018) is trained for 15,600 updates in the base comparison and 124,800 in the longer-training check. The discrete autoregressive control is trained for 12,000 updates and evaluates a native probability mass, avoiding the interpretation of a continuous point density as a discrete image probability. The allowed test set contains 987 images already satisfying $C(x) \geq \tau$; the 9,006 excluded images satisfy both $C(x) < \tau$ and $C(Tx) < \tau$. Seven additional images meeting the criterion only after flipping are omitted from this comparison. Thus the reported support comparison is conservative with respect to the specified augmentation.

For a continuous flow, let B_x be the quantization bin associated with integer image x. The corresponding restricted continuous target is supported on $\bigcup_{x \in A} B_x$. Uniformly dequantized versions of excluded images lie outside that support. Evaluated flow scores are dequantized log densities; a score at the bin center is not the bin-integrated probability, and an average log score over dequantizations is not its logarithm either. The discrete autoregressive result supplies the native-mass counterpart.

C.2 Finite equal-entropy construction

Here R is uniform over the selected 5,000 original training images together with their horizontal flips. Deduplication confirms 10,000 distinct images. The second distribution is uniform over 10,000 distinct constant-RGB images, disjoint from the first set. Consequently both entropies are exactly $\log 10{,}000$. Expected native autoregressive log probability is computed by an exhaustive uniform average over each set. This construction compares two fully specified discrete distributions; it neither

Table 5: The main generative diagnostic quantities averaged across the two training runs. A higher-score fraction includes half credit for ties. The longer Glow run uses eight times the base update budget.

Diagnostic	Result
Full-data Glow: within-ID Spearman$(S, -\log C)$	0.924
Random-subset Glow: within-ID Spearman$(S, -\log C)$	0.923
Random-subset autoregressive: within-ID Spearman$(S, -\log C)$	0.877
Filtered Glow: excluded images score higher than allowed images	97.17%
Filtered autoregressive: excluded images score higher	94.46%
Longer filtered Glow: excluded images score higher	96.48%
Finite equal-entropy task: mean $\log m(Q) - \log m(R)$	5,115.85 nats
Finite equal-entropy task: raw AUROC	0.037%

estimates an unknown natural-image population entropy nor makes the finite training distribution the universal definition of ID.

For the acceptance experiment, independent IID samples with replacement are drawn from R for ID calibration and ID evaluation; the fitted autoregressive model is held fixed. Each split has 1,000 draws. With the two-sided rule at $\alpha = .05$, mean color acceptance is 0.19% and measured ID retention is 94.10%. The one-sided rule accepts all color images. The high color scores do not imply high total probability mass: the summed model mass of the finite color set remains extremely small. The construction therefore separates the incorrect entropy inference from the valid distinction between density and mass.

C.3 Noise-only training: protocol, trajectories, and individual runs

Training protocol. We train four Glow models: two seeds $(0, 1)$ for each of fixed-pool and fresh-noise training. All models use three levels, eight flow steps per level, and hidden width 400. Training starts without pretrained weights and uses batch size 128, Adam with learning rate 10^{-4} and weight decay 10^{-5}, a 500-update linear warmup, and gradient-norm clipping at 100. Each run completes 62,400 optimizer updates. All four runs use NVIDIA A40 GPUs, FP32 training, and disabled matrix-multiplication and cuDNN TF32. Evaluation is separate and uses FP64. The implementation uses PyTorch 2.5.1, CUDA 12.1, and normflows 1.7.3.

The shared fixed pool consists of 5,000 independent $3 \times 32 \times 32$ uint8 arrays sampled with NumPy `default_rng(20260926001)`. Fixed training cycles through shuffled copies of this pool; fresh training draws a new independent uint8 batch at every optimizer update. Both modes apply horizontal flips with probability $1/2$ and independently draw $U \sim \mathrm{Uniform}([0, 1)^{3072})$ on each occurrence, using $(B + U)/256$. This defines the ideal continuous uniform population for fresh-noise training. The actual FP32 implementation can round rare upper-bin inputs to 1; the population statements refer to the ideal generator. Conditional on its finite pool, the fixed training distribution is the augmented and dequantized empirical distribution, not the continuous uniform population. Independent noise tests measure generalization to the original generator.

Within a seed, the two modes share the random parameter state and the first 128-example ActNorm initialization batch drawn from the fixed pool. Subsequent data streams use separate recorded random generators. We retain pre-ActNorm random weights, post-ActNorm step 0, and optimizer steps $1, 10, 100, 500, 1000, 5000, 15600, 31200, 62400$. Thus each model has 11 checkpoints, with the two distinct initialization phases kept separate. Probes never initialize or update the training model. The evaluation of pre-ActNorm weights disables data-dependent initialization only on a disposable evaluation copy. No CIFAR, SVHN, or other natural image is used in these training runs.

Scores and references. Each checkpoint is evaluated using two fixed dequantizations per image identity; the two log scores are averaged *before* computing ranks and summary statistics. We use 2,000 independent noise-test images, a separate 2,000-image noise-calibration set, the 5,000-image fixed reference pool, 2,000 SVHN images, all 256 gray levels, and 2,000 fixed RGB-color images. The CIFAR probe contains 5,000 images with the frozen class weights. Full evaluation includes 13 probe families. Here $S = J + L$ is the continuous flow log density; L uses the model's multiscale latent density, including its learned parameters. Averaging log densities over dequantizations does

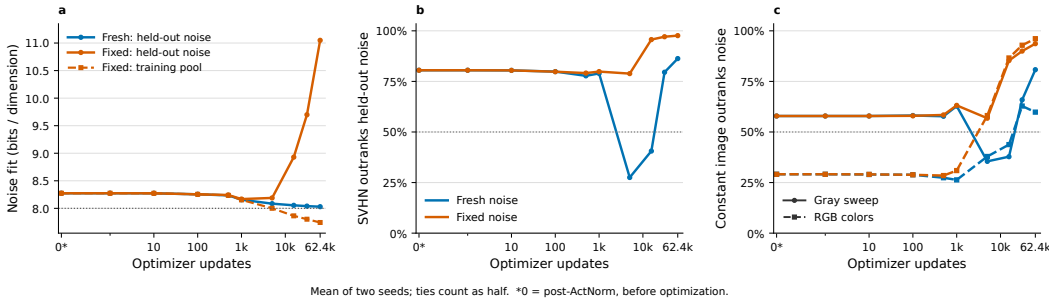


Figure 3: Noise-only training produces probe-dependent high-score preferences. (a) Checkpoint evaluations of independent noise and the fixed training pool. The dashed curve is the actual fixed-pool training reference, evaluated after the same updates as its held-out counterpart. (b) SVHN and (c) gray/RGB-constant higher-score fractions w relative to independent noise. Colors identify fresh versus fixed training; panel (c) line styles identify probe families. Every curve is the mean of two seed-specific metrics, with two dequantizations averaged per identity before ranking. Ties receive half credit. Horizontal lines mark the uniform oracle: 8 bpd in (a), 50% in (b,c). All ten post-ActNorm/optimization checkpoints appear on a symlog update axis; 0^* is after ActNorm and before optimization. True pre-ActNorm weights are a distinct phase and are not connected to this step-0 point. No smoothing or seed-derived confidence band is used.

Table 6: Selected complete-checkpoint summaries, averaged over the two seeds. Pool and test entries are bpd; remaining entries are higher-score fractions $100w$. Pool† is only a fixed reference for fresh-noise training. Step 0 is post-ActNorm, with no optimization.

Mode	Updates	Pool†	Noise test	SVHN	Gray	RGB
Fresh	0	8.274	8.274	80.51%	57.90%	29.10%
	5,000	8.085	8.085	27.52%	35.48%	37.89%
	15,600	8.055	8.055	40.57%	37.79%	43.76%
	31,200	8.040	8.040	79.54%	65.89%	62.83%
	62,400	8.030	8.030	86.28%	80.79%	59.82%
Fixed	0	8.274	8.274	80.51%	57.90%	29.10%
	5,000	8.001	8.189	78.86%	56.81%	58.12%
	15,600	7.864	8.929	95.67%	85.30%	86.58%
	31,200	7.803	9.699	97.09%	89.99%	92.88%
	62,400	7.743	11.049	97.63%	93.65%	96.10%

not produce the exact discrete image log probability. With $d = 3072$, the displayed convention is $\text{bpd} = 8 - \overline{S}/(d \log 2)$. The uniform oracle has $S = 0$ and 8 bpd.

For independent draws $Y \sim Q$ from a probe family and $X \sim N$ from the independent-noise reference, the higher-score fraction is

$$w(Q, N) = \Pr\{\overline{S}(Y) > \overline{S}(X)\} + \tfrac{1}{2} \Pr\{\overline{S}(Y) = \overline{S}(X)\} = 1 - \text{AUROC}_{N+} \,. \tag{15}$$

The score direction is never reversed after observing results. A constant uniform-oracle score gives $w = 1/2$, which does not preclude detection by other statistics. These probes lie inside the uniform support. All aggregate curves and table entries average the two seed-specific metrics arithmetically; they do not pool outputs from different models or imply an uncertainty interval from two runs. The fixed pool is an actual training reference only for fixed training. For fresh training it is a reference pool whose initialization batch was used by ActNorm, not the optimizer's training dataset.

Relative preference and absolute score changes. The fixed-pool preference grows while both probe and held-out-noise mean scores decline. From 5,000 to 62,400 updates, mean noise-test S falls from -402.58 to $-6,492.94$ nats, while SVHN falls from -228.32 to $-1,499.14$. SVHN's relative advantage therefore grows from 174.26 to $4,993.79$ nats without an increase in its own average density score. The fixed-pool train/test gap is evidence of finite-sample overfitting, not an independent identification of a memorization mechanism. Under fresh-noise training, both scores instead improve over this interval, with a larger improvement for SVHN; its mean advantage changes from -197.78 to 23.42 nats. These trajectories distinguish relative ranking from absolute score changes.

Table 7: Individual final runs at 62,400 updates. Pool/test values are checkpoint bpd, not the training mini-batch loss recorded before an update. Probe columns report $100w$ against independent noise. For fresh training the pool remains a reference, not its optimizer training set.

Mode	Seed	Pool	Noise test	SVHN	Gray	RGB
Fresh	0	8.03071	8.03074	84.540%	78.936%	57.437%
Fresh	1	8.02943	8.02949	88.025%	82.652%	62.201%
Fixed	0	7.75310	11.04188	99.942%	99.708%	97.457%
Fixed	1	7.73366	11.05664	95.322%	87.584%	94.752%

Spatial controls and scope. At the final fresh-noise checkpoint, all five histogram-matched comparisons have negative *mean* paired ΔS in both seeds. Fixed-pool point means are positive, but fixed seed 1 SVHN versus shuffle is weak: $\Delta S = 12.22$ nats, a model-conditional 95% bootstrap interval $[-16.26, 40.79]$, and 48.7% strict wins. Hence the seed-averaged positive value is not evidence of reliable preference in every seed. The CIFAR, SVHN, and texture shuffles preserve RGB triplets. Per-channel sorted-noise controls preserve channelwise value multisets but can alter RGB pairing; the cyclic-shift variant restores fixed-position RGB uniform marginals at the population-generator level, not as exact finite-sample equality. These controls delimit a probe- and training-dependent phenomenon rather than a universal preference for spatial simplicity.

All checkpoints retain optimizer, scheduler, and random-generator states. Their recorded consistency was audited, but no claim of tested cross-environment bitwise resumption is made. Results are conditional on one Glow architecture, two parameter seeds, the fixed noise-pool construction, and the prescribed finite probe families. The Gaussian-base full-support model and the bounded uniform target also differ in their support; this experiment does not isolate the independent contributions of parameterization, boundary behavior, or optimization.

D FIXED-COORDINATE AND SPECTRAL CONTROLS

These supplementary controls distinguish transporting one density through a known coordinate map from learning a new density in those coordinates. Coordinate dependence of density-based anomaly rules is established prior work (Le Lan & Dinh, 2021); our experiments concern the additional effects of representation on finite-budget Glow fitting. Local image statistics can also affect coupling-layer flows (Kirichenko et al., 2020), but the controls below do not isolate a single spectral or architectural mechanism.

Two fixed byte maps and the transported comparator. For an image $k \in \{0, \dots, 255\}^{3072}$, the additive map is $T_{\text{mask}}(k)_j = (k_j + M_j) \bmod 256$, using one fixed mask drawn with NumPy seed 20260925201. The second map is $T_{\text{sbox}}(k)_j = \pi_j(k_j)$, where each coordinate has one fixed permutation of the 256 byte values, generated in coordinate order using seed 20260925301. Neither map is resampled per image or selected from model scores. Horizontal flips precede the map. For continuous dequantization we keep the same within-bin variable: $T((k+u)/256) = (T(k)+u)/256$. Both byte maps preserve all image-atom probabilities and each distribution's own Shannon entropy. Their ideal within-bin extensions preserve volume and differential entropy where defined. This does not make CIFAR and SVHN equal-entropy distributions. Extending the map inside each unit cube while retaining integer coordinates gives a full-space bijection, so the transported model $m_T(y) = m(T^{-1}y)$ has exactly the corresponding original scores almost everywhere. A newly trained model on $T(\text{CIFAR})$ need not recover this transported density: the maps change local geometry and convolutional learnability.

Operational spectral matching, not equality of distributions. Before the new S-box model outcomes, fixed diagnostic classifiers compared CIFAR and SVHN using five feature blocks: radial non-DC normalized power, full two-dimensional non-DC normalized power, normalized power including DC, log absolute power, and channel means/standard deviations/energies. Each block used both fixed logistic and shallow boosted-tree probes, with 2,000 fitting and 2,000 held-out images per domain. All ten S-box probe AUROCs were 48.38–51.27%, and all their conditional image-bootstrap 95% intervals lay within the prespecified 45–55% band. The same fixed probes gave 94.27–99.93%

Table 8: Fixed-coordinate retraining at 62,400 updates. AUROC compares held-out ID with correspondingly mapped SVHN, with higher log density fixed as ID. The training/test columns are mean continuous-density bpd, $8 - S/(3072 \log 2)$; they are not discrete image code lengths. Both predefined training seeds are retained.

Coordinates	Seed	Train bpd	Test bpd	SVHN bpd	AUROC (%)
Original	0	3.627	4.065	2.635	3.812
Original	1	3.613	4.114	2.649	3.740
Additive mask	0	6.800	7.885	7.093	13.483
Additive mask	1	6.787	7.797	6.992	13.878
S-box	0	7.729	10.912	10.919	50.468
S-box	1	7.738	10.921	10.951	52.156

on raw images and 63.65–90.08% after the additive map. These AUROCs measure separation by the specified spectral classifiers, not by model density. The result applies to the specified summaries and fitted probes; power features omit phase, and an unrestricted observer can invert either map. The diagnostic cohort was initially unweighted. A fixed-class-proportion reweighting sensitivity retained the S-box near-chance finding; neither version proves equality of image distributions or spectral indistinguishability to all possible classifiers.

Retraining and held-out detection. The original, masked, and S-box Glow models use the same 5,000 training identities, two training seeds, paired pre-map random streams, and the fixed 62,400-update endpoint. Each representation uses its own correctly transformed ID reference. CIFAR evaluation has 5,000 held-out identities, separate from 5,000 calibration identities; CIFAR scores use training-class-proportion weights, two dequantizations, and both horizontal orientations. SVHN has 2,000 identities with two dequantizations. These repeated scores are pooled within a model and are not independent images.

The arithmetic mean AUROC over the two seeds changes from 3.776% to 13.680% and 51.312%, respectively. This is not a successful detector repair. The S-box train–test gap is about 3.183 bpd; its training scores separate from SVHN with diagnostic AUROC 100% in both seeds, whereas unseen ID and SVHN nearly overlap. Independent one-sided ID-calibration thresholds still accept 95.000% and 94.125% of SVHN while retaining 95.189% and 95.022% of held-out ID. The result is consistent with severe generalization failure under this representation and budget, without uniquely identifying a memorization mechanism. It cannot attribute the AUROC change solely to spectrum, nor establish that simplicity preference is necessary for successful generalization.

Untransformed images in the new coordinates. We also directly input untransformed CIFAR and SVHN into each transformed-data model. Untransformed CIFAR is then a coordinate-shift probe, not its ordinary ID distribution. With true ID $T(\mathrm{CIFAR})$ as the positive reference, native-SVHN AUROCs are $61.77/73.48\%$ for the two mask seeds but $0.23/4.93\%$ for the S-box seeds. Both models rank native SVHN above native CIFAR (native-CIFAR-positive AUCs $29.08/24.18\%$ and $21.96/21.04\%$), but only S-box assigns both native families higher mean log density than its true held-out ID. The mask is therefore a necessary contrasting result; these probes do not support a claim that every scrambling makes native simple images outrank ID.

A distinct frozen-model Fourier control. Separately, one previously selected Glow checkpoint was evaluated under six fixed Fourier phase operators: two phase seeds and strengths $1/64, 1/32, 1/16$. The same conjugate-symmetric phase multiplier is applied to all RGB channels and images, fixing DC/Nyquist frequencies. Each operator preserves image-wise periodic cross-spectra, Euclidean distances, and differential entropy where finite. These operators are not the S-box training intervention. On the fixed 1,000-image-per-domain analysis split, applying the operator to both CIFAR and SVHN leaves raw-score AUROC at 4.19–4.86% (clean: 4.94%), while the fixed reversed Jacobian score retains 95.94–96.27% AUROC. Thus the preserved summaries do not uniquely determine the frozen model's component scores, and the Jacobian remains informative in this comparison. No clipping was used; some outputs leave the pixel cube, so support change remains a confound. The model density was not transported with the images, and the checkpoint had been selected using CIFAR test bpd. This supplementary intervention neither violates density transport identities nor establishes an oracle-density failure.

E FLOW CORRECTION AND COMPONENT CALIBRATION

Face-domain protocol. We use aligned CelebA images (Liu et al., 2015), converted to RGB, center-cropped from 178×218 to 178×178 (box $(0, 20, 178, 198)$), and bilinearly resized with Pillow to 32×32 uint8 images. The core cohort contains the 200 most frequent identities (numeric-ID tie break), split within identity using seed 2026092513 into 60%/10%/15%/remainder training/validation/calibration/test portions. Exact decoded-image duplicates are removed in that role order, yielding 3,616/600/805/1,016 images. The full cohort uses official training/test partitions and a shuffled, halved official validation partition; after duplicate removal, counts are 162,685/9,922/9,917/19,942. Each cohort uses its own held-out reference.

Glow has three levels, eight blocks per level, hidden width 400, and 7,545,216 parameters. Both seeds (0,1) train from scratch with Adam (learning rate 10^{-4}, weight decay 10^{-5}), 500-update warmup, batch size 128, gradient-norm clipping at 100, horizontal flips, and uniform dequantization $(x + U)/256$, without AMP or TF32. Core/full runs use 15,600/62,400 updates. Selection maximizes validation mean log density at fixed milestones: $(250, 1000, 4000, 15600)$ for core and $(1000, 4000, 15600, 31200, 62400)$ for full; all four selected the final milestone. The main-table FP64 evaluation uses bin centers $(x + 0.5)/256$, not integrated discrete mass. Separate calibration sets define rank tests at $\alpha = 0.05$; realized test retention is reported.

E.1 THE RADIAL DENSITY CORRECTION

The image dimension is $D = 3 \cdot 32 \cdot 32 = 3072$. The model's Gaussian base comprises three scale blocks of dimensions 768, 768, and 1536, with their fitted diagonal means and scales. Whiten the concatenated latent vector using those parameters. If r^2 is its squared radius and $t = \log r^2$, the model base has $r^2 \sim \chi^2_D$. For $k = D/2$, the log density of t is

$$\log a_0(t) = kt - \tfrac{1}{2}e^t - k \log 2 - \log \Gamma(k). \tag{16}$$

Equivalently, if $L = C - r^2/2$ is the computed Gaussian log density, then $r^2 = 2(C - L)$, with C including the learned scale normalizers. Thus L is evaluated directly from the fitted Gaussian base.

The fitted marginal $\widehat{a}$ is either Gaussian or Gaussian KDE on the ID fitting split. Replacing the base radial marginal while keeping its conditional direction distribution gives Eq. (6); normalization follows by integrating first conditional on t. The model is normalized on its full continuous domain. This does not turn center-evaluated scores into normalized discrete image masses or condition the model on the bounded pixel cube.

Two face-data cohorts, two training seeds, three fixed dequantization conventions, and two radial estimators yield 24 fits. Each uses disjoint ID fitting, threshold-calibration, and test splits. The full cohort has 4,958 fitting, 4,959 calibration, and 19,942 test images; the smaller cohort has 402, 403, and 1,016. SVHN evaluation uses 1,033 held-out images. The main table reports the full cohort and center dequantization, averaging the two training runs. Improvements in mean held-out log density range from 0.53 to 38.23 nats across all settings. We separately evaluate each corrected score as a detector.

E.2 ID-ONLY AND SOURCE-SUPERVISED RULES

An ID-only KDE score fits the distribution of S using ID fitting examples. The DoSE-style component rule fits marginal densities of J and L and combines their log values (Morningstar et al., 2021). Supervised LR denotes logistic regression: it is fitted on either S or (L, J) with labeled ID and source-OOD calibration examples. SVHN-source evaluation uses disjoint SVHN calibration and test samples. A separate ID split calibrates acceptance. The two-sided rule directly accepts $p_2 > \alpha$ without fitting another threshold to p_2; the other rules use their prescribed ID-calibrated score thresholds. Actual test-ID retention is reported. This comparison identifies the additional value of retaining the components at a fixed calibration-information budget.

F CLASSIFIER IMPLEMENTATION AND EVALUATION

F.1 THE FIBER FACTOR AND THE SCOPE OF THE OPERATIONAL ESTIMATORS

Coarea identity. Let p_X be absolutely continuous on an open subset of $\mathbb{R}^D$ and let $g : \mathbb{R}^D \to \mathbb{R}^r$ be locally Lipschitz. Suppose Dg has row rank r at p_X-almost every point. Applying coarea to $p_X/\mathcal{J}_g$ shows that the pushforward is absolutely continuous and that, for almost every z with $p_Z(z) > 0$,

$$\int_{g^{-1}(z)} \rho(x \mid z)\, d\mathcal{H}^{D-r}(x) = 1. \tag{17}$$

Disintegration then identifies this normalized function as a conditional density on the fiber. Multiplication gives $p_X(x) = p_Z(g(x))\mathcal{J}_g(x)\rho(x \mid g(x))$ almost everywhere, and taking logarithms gives Eq. (8) on its positive locus. For $D = r$, $\mathcal{H}^0$ is counting measure. Bijectivity makes the conditional mass one. A noninjective square map can have several inverse branches and a nonzero log conditional branch mass; matching dimensions alone therefore does not imply $F = 0$.

The rank hypothesis is substantive. A full-row-rank affine head W does not establish full row rank of WD_xh. ReLU networks are locally Lipschitz and differentiable almost everywhere, but may still lose rank on regions with positive probability. A positive ridge makes a numerical score finite; it does not prove the hypothesis or supply the missing singular component of a probability distribution. A distribution on a lower-dimensional data manifold also requires a corresponding intrinsic reference measure, beyond the ambient Lebesgue formulation used here.

Why a feature conditional does not supply the input conditional. If, additionally, $D \geq d \geq r$ and both h and $g = Wh + b$ satisfy the regularity assumptions, apply coarea first to h and then to W:

$$\log p_X = \log p_H(h) + J_h + F_{X|H}, \tag{18}$$

$$\log p_H(h) = \log p_Z(z) + J_W + F_{H|Z}, \qquad J_W = \tfrac{1}{2}\log\det(WW^\top). \tag{19}$$

Comparing with the direct decomposition of g gives

$$F_{X|Z} = F_{H|Z} + F_{X|H} + J_h + J_W - J_g. \tag{20}$$

The geometric correction is needed: normal Jacobians of dimension-reducing maps do not generally obey the ordinary square-determinant chain rule. Our classifier scores do not estimate every term on the right. Consequently, even an exact fitted feature conditional does not identify the input-space $F_{X|Z}$.

Primary residual coordinate and exact fitting convention. Write an SVD $W = U_W \operatorname{diag}(s)Q$ with $QQ^\top = I_r$, $N^\top N = I_{d-r}$, and $QN = 0$. The geometry covariance is the unbiased training covariance with the ridge

$$\Sigma = \frac{1}{n-1}\sum_i (h_i - \mu)(h_i - \mu)^\top + 10^{-6}\frac{\operatorname{tr}(\Sigma_0)}{d} I_d, \tag{21}$$

where Σ_0 denotes the covariance before adding the ridge. The construction in Eq. (9) has $QK = I_r$. On a fixed affine fiber, $h \mapsto u$ is an isometry: differences satisfy $h - h' = N(u - u')$. The map $h \mapsto (z, u)$ is invertible with absolute determinant $\mathcal{J}_W = \sqrt{\det(WW^\top)}$; in orthogonal (Q, N) coordinates, its residual step is a triangular shear with unit determinant. Thus independent normalized fits $\widehat{q}_Z$ and $\widehat{q}_U$ can define a normalized feature model $\widehat{q}_H(h) = \mathcal{J}_W \widehat{q}_Z(z(h))\widehat{q}_U(u(h))$. This fact concerns that particular product feature model and does not make it equal to a separately fitted feature mixture or a population distribution.

For the primary readout fit, $\widehat{\pi}_c = n_c/n$ and $\widehat{\mu}_c = n_c^{-1}\sum_{i:y_i=c} z_i$. With $S_Z = n^{-1}\sum_i (z_i - \widehat{\mu}_{y_i})(z_i - \widehat{\mu}_{y_i})^\top$, $\widehat{\Sigma}_Z = S_Z + 10^{-6}\operatorname{tr}(S_Z)I_r/r$. The residual Gaussian uses its unconditional sample mean and the maximum-likelihood covariance with the same relative ridge in dimension $d - r$. These fitted covariances use denominator n, whereas the geometry covariance above uses $n - 1$. All readouts are reconstructed in FP64 from cached features and the actual checkpoint head; ImageNet uses $W = P^\top W_0$, $b = P^\top b_0$ for the frozen projection P.

Self-consistent conditional control. Fit $\widehat{p}_H(h) = \sum_c \pi_c \mathcal{N}(h; m_c, \Sigma_H)$ once. Its readout marginal has means $\nu_c = W m_c + b$ and covariance $A = W \Sigma_H W^\top$. With $v = N^\top h$, define

$$\gamma_c(z) = \frac{\pi_c \mathcal{N}(z; \nu_c, A)}{\sum_j \pi_j \mathcal{N}(z; \nu_j, A)}, \tag{22}$$

$$a_c(z) = N^\top m_c + N^\top \Sigma_H W^\top A^{-1}(z - \nu_c), \tag{23}$$

$$B = N^\top [\Sigma_H - \Sigma_H W^\top A^{-1} W \Sigma_H] N. \tag{24}$$

Then the feature-fiber conditional, measured in the orthonormal coordinate v, is

$$\rho_H(h \mid z) = \sum_c \gamma_c(z) \mathcal{N}(v; a_c(z), B). \tag{25}$$

The control sets $L = \log \widehat{p}_Z$ and $F = \log \rho_H$ from these same parameters. Algebra gives $\log \widehat{p}_H(h) = L(h) + J_W + F(h)$. The implementation evaluates the equivalent conditional after Cholesky whitening and a complete QR decomposition, avoiding cancellation from tiny head singular values. It independently evaluates the conditional rather than defining it by subtracting $\log \widehat{p}_H - L - J_W$. The verified maximum identity residual across 269 caches is 2.70×10^{-10} nats. In the detector, J_ϵ remains the input-to-readout statistic; it is not the constant feature-head term J_W.

Other four fitting choices. The diagonal GMMs use 10 or 50 components, random state 0, covariance ridge 10^{-6}, and at most 200 EM iterations. PCA–GMM10 retains $\min(64, \text{dimension}, \lfloor n/10 \rfloor)$ principal coordinates, then fits a full-covariance 10-component mixture (ridge 10^{-4}, random state 0, at most 200 iterations). This scores the retained coordinates. The 50-NN score is $-k \log(r_{50} + 10^{-12})$ in coordinate dimension k, using a fixed random-state-0 bank of at most 20,000 training points. Its additive density normalization constant is omitted. These four families fit the readout and residual slots independently and do not use OOD calibration labels. Along with the primary independent Gaussian choice and the joint conditional control, all six choices are retained.

An operational decomposition-based detector. Let $h(x) \in \mathbb{R}^d$ denote a frozen classifier's penultimate feature, and let $z(x) = Wh(x) + b \in \mathbb{R}^r$ be its specified affine readout. We use all 10 or 100 logits for CIFAR-10/100; on ImageNet, z is a fixed 32-dimensional projection of the original 1,000 logits. The baseline scores retain their own original readouts. Our detector combines three statistics: an input-to-readout geometric statistic J_ϵ, a fitted readout-density score L, and a fitted feature-residual score F. They provide the operational LJF component vector motivated by the general density decomposition.

Geometry and its parameterization. For the r cached singular values $\sigma_i(x)$ of $J_z(x) = \partial z / \partial x$ in raw $[0, 1]$ input coordinates, define

$$a(x) = r^{-1} \sum_{i=1}^{r} \sigma_i(x)^2, \tag{26}$$

$$J_\epsilon(x) = \tfrac{1}{2} \sum_{i=1}^{r} \log \max\{\sigma_i(x)^2 + \epsilon a(x), \delta^2\}, \qquad \epsilon = 10^{-6}, \quad \delta = 10^{-30}. \tag{27}$$

The ridge was fixed using numerical sensitivity checks before inspection of the stable detection results. It stabilizes a spectral statistic; it does not establish local full rank or recover an exact coarea factor at rank loss. We separate log-RMS scale from regularized shape:

$$A(x) = \tfrac{1}{2} \log \left[r^{-1} \sum_i \max\{\sigma_i(x), \delta\}^2 \right], \qquad B_\epsilon(x) = J_\epsilon(x) - rA(x). \tag{28}$$

We compare adding J_ϵ, A, B_ϵ, or the pair (A, B_ϵ) to the same fusion baseline. The pair contains a linear expression for J_ϵ; separating its terms preserves distinctions lost by their scalar combination and permits independently fitted weights. To avoid differentiating through an ill-conditioned head directly, we factor $W = U \operatorname{diag}(s) Q$, differentiate the orthonormal readout $Qh(x)$, and reconstruct the original-readout Jacobian in FP64 before its SVD. Backbone differentiation is FP32; this is not an entirely FP64 backward pass.

Fitted density proxies. We fit all proxy estimators on cached ID training features, independently of OOD calibration labels. Let μ and Σ be the training feature mean and covariance, with relative covariance ridge 10^{-6}, and let the columns of N span $\ker W$ orthonormally. Using the orthonormal row basis Q above,

$$K = \Sigma Q^\top (Q\Sigma Q^\top)^{-1}, \qquad u(x) = N^\top (I - KQ)(h(x) - \mu). \tag{29}$$

We set $L(x) = \log \widehat{q}_Z(z(x))$ and $F(x) = \log \widehat{q}_U(u(x))$. The primary tied-Gaussian choice uses an empirical-class-prior mixture with class means and a shared covariance for $\widehat{q}_Z$, and one unconditional Gaussian for $\widehat{q}_U$. The residual construction removes linear predictability under the fitted covariance; it does not make $\widehat{q}_U$ the true conditional fiber law. We retain all prespecified alternatives: diagonal GMMs with 10 or 50 components, a 10-component full-covariance GMM on at most 64 PCA coordinates, and a 50-NN score $-d_u \log(r_{50} + 10^{-12})$ (with the appropriate coordinate dimension for each slot and a maximum 20,000-point training bank). The PCA score is a projected-coordinate density and the NN score omits an additive normalization constant. Neither is silently treated as an ambient input likelihood.

A self-consistent conditional control. As a sixth family, we fit one tied-Gaussian mixture $\widehat{p}_H$ in feature space and compute its affine-readout marginal and feature-fiber conditional from the same model. Independent whitening/QR conditional formulas verify $\log \widehat{p}_H = \log J_W + L + F$, with $J_W = \sqrt{\det(WW^\top)}$. Across all 269 model–dataset caches, the largest affine-identity residual is 2.70×10^{-10} nats. This certifies internal consistency of the fitted feature model, not correctness of its population fit or identification of the input-space conditional density.

Learning the detection rule. The three-feature detector is

$$D_{\mathrm{LJF}}(x) = \beta_0 + \beta^\top \,\mathrm{standardize}_{\mathrm{source}}[L(x), J_\epsilon(x), F(x)]^\top. \tag{30}$$

LF uses the same procedure on (L, F). We fit binary logistic regression with ID as the positive label, $C = 1$, L2 regularization, no class weighting, and L-BFGS with at most 5,000 iterations. The pooled ID/source-OOD calibration mean and population standard deviation (plus 10^{-12}) define each standardizer. Coefficient signs are learned. Fixed raw sums and ID-only standardized positive sums are separate controls, each compared with its corresponding two-term rule.

A strong fusion baseline. ALL applies the same logistic calibration to one selected score from each of 27 implemented baseline families. A family's grid setting and sign maximize AUROC on source calibration data only; the best-single baseline is also selected there, never on test data. ALL+J_ϵ refits the identical pipeline with one additional feature. Shape, scale, their two-dimensional pair, duplicate-MSP, and five independently generated label-independent Gaussian features are separate augmentation controls. The 27 families are MSP, MLS, Energy, EnergyT, Entropy, probability margin, logit margin, temperature scaling, GEN, KLM, MDS, RMDS, KNN, NNGuide, ViM, Residual, NuSA, ReAct, ASH-P/B/S, SCALE, DICE, SHE, fDBD, GradNorm, and ODIN. All methods use the fixed implementation suite and source-calibration selection protocol.

Source-matched evaluation. For calibration seeds $0, 1, 2$, each ID and OOD cache is randomly split 50/50; ID indices are shared across OOD sources within a seed. Source-matched evaluation uses disjoint calibration/test images from the same OOD family. The cohort has 17 CIFAR-10 classifiers with six OOD targets, 12 CIFAR-100 classifiers with six, and 11 ImageNet classifiers with five: 40 classifiers and 229 model–target cells. Results first average the three calibration splits within each cell. Exploratory intervals use 1,000 crossed resamples of models and targets within each benchmark; overall weights are fixed at $102 : 72 : 55$. Calibration splits share each fixed base model; each benchmark contains five or six target clusters. FPR95 is computed from the held-out ROC curve.

G Extended classifier results

The following tables report all six density-proxy families and both detection metrics under the source-matched calibration protocol. A positive AUROC difference and a negative FPR95 difference favor the geometry-augmented method. Bootstrap intervals describe variation across the available model and target clusters; they are exploratory rather than simultaneous significance guarantees over all comparisons.

Table 9: The complete 40-model classifier cohort (17 CIFAR-10, 12 CIFAR-100, and 11 ImageNet models). Identifiers encode the architecture and width/depth variant. TV V1/V2 denotes the corresponding torchvision `IMAGENET1K_V1/V2` weights; CIFAR hub denotes pretrained weights from `chenyaofo/pytorch-cifar-models`. The final column is the exact checkpoint filename suffix, not a newly computed digest. For CIFAR hub and torchvision checkpoints, the basename is the model identifier followed by a hyphen and this suffix, except that EfficientNet adds `_rwightman` before the hyphen. The local adversarial ResNet-20 uses `adv_resnet20_s0.pt`; it is a batch-normalized ResNet-20. These are 40 selected checkpoints, not 40 independent datasets or training-seed replications.

ID dataset	Model identifier	Weight source	Filename suffix
CIFAR-10	`adv_resnet20`	Local (seed 0)	`s0.pt`
	`cifar10_mobilenetv2_x0_5`	CIFAR hub	`ca14ced9.pt`
	`cifar10_mobilenetv2_x1_0`	CIFAR hub	`fe6a5b48.pt`
	`cifar10_mobilenetv2_x1_4`	CIFAR hub	`3bbbd6e2.pt`
	`cifar10_repvgg_a0`	CIFAR hub	`ef08a50e.pt`
	`cifar10_repvgg_a1`	CIFAR hub	`38d2431b.pt`
	`cifar10_repvgg_a2`	CIFAR hub	`09488915.pt`
	`cifar10_resnet20`	CIFAR hub	`4118986f.pt`
	`cifar10_resnet32`	CIFAR hub	`ef93fc4d.pt`
	`cifar10_resnet44`	CIFAR hub	`2a3cabcb.pt`
	`cifar10_resnet56`	CIFAR hub	`187c023a.pt`
	`cifar10_shufflenetv2_x1_0`	CIFAR hub	`98807be3.pt`
	`cifar10_shufflenetv2_x2_0`	CIFAR hub	`ec31611c.pt`
	`cifar10_vgg11_bn`	CIFAR hub	`eaeebf42.pt`
	`cifar10_vgg13_bn`	CIFAR hub	`c01e4a43.pt`
	`cifar10_vgg16_bn`	CIFAR hub	`6ee7ea24.pt`
	`cifar10_vgg19_bn`	CIFAR hub	`57191229.pt`
CIFAR-100	`cifar100_vgg11_bn`	CIFAR hub	`57d0759e.pt`
	`cifar100_vgg13_bn`	CIFAR hub	`5ebe5778.pt`
	`cifar100_vgg16_bn`	CIFAR hub	`7d8c4031.pt`
	`cifar100_vgg19_bn`	CIFAR hub	`b98f7bd7.pt`
	`cifar100_mobilenetv2_x0_5`	CIFAR hub	`9f915757.pt`
	`cifar100_mobilenetv2_x1_0`	CIFAR hub	`1311f9ff.pt`
	`cifar100_mobilenetv2_x1_4`	CIFAR hub	`8a269f5e.pt`
	`cifar100_repvgg_a0`	CIFAR hub	`2df1edd0.pt`
	`cifar100_repvgg_a1`	CIFAR hub	`c06b21a7.pt`
	`cifar100_repvgg_a2`	CIFAR hub	`8e71b1f8.pt`
	`cifar100_shufflenetv2_x1_0`	CIFAR hub	`9ae22beb.pt`
	`cifar100_shufflenetv2_x2_0`	CIFAR hub	`e7e584cd.pt`
ImageNet	`resnet50`	TV V2	`11ad3fa6.pth`
	`resnet101`	TV V2	`cd907fc2.pth`
	`resnet152`	TV V2	`f82ba261.pth`
	`wide_resnet50_2`	TV V2	`9ba9bcbe.pth`
	`resnext50_32x4d`	TV V2	`1a0047aa.pth`
	`densenet121`	TV V1	`a639ec97.pth`
	`densenet201`	TV V1	`c1103571.pth`
	`efficientnet_b0`	TV V1	`7f5810bc.pth`
	`efficientnet_b3`	TV V1	`b3899882.pth`
	`convnext_base`	TV V1	`6075fbad.pth`
	`mobilenet_v3_large`	TV V2	`5c1a4163.pth`

Table 10: Adding geometry to the density-proxy detector: learned LJF minus learned LF, AUROC percentage points with saved exploratory 95% intervals. All six prespecified proxy families are retained. Positive AUROC differences and negative FPR95 differences favor LJF. These compare against the two-feature LF detector, not against ALL. Intervals use 1,000 model–target crossed-bootstrap draws within each benchmark after averaging calibration splits.

Proxy family	CIFAR-10	CIFAR-100	ImageNet
Tied Gaussian	+2.01 [+1.08, +3.18]	+5.00 [+0.72, +13.30]	+1.15 [-0.22, +4.12]
Diagonal GMM (10)	+2.57 [+1.50, +3.88]	+9.58 [+4.74, +17.35]	+2.03 [+0.53, +4.68]
Diagonal GMM (50)	+2.25 [+1.20, +3.42]	+6.90 [+2.43, +14.73]	+2.18 [+0.58, +4.64]
PCA64 + GMM (10)	+2.05 [+1.26, +2.92]	+5.49 [+1.70, +13.01]	+2.98 [+0.78, +6.93]
50-NN score	+2.01 [+0.83, +3.27]	+5.35 [+0.60, +13.74]	+2.27 [+0.89, +4.24]
Joint conditional Gaussian	+2.06 [+1.13, +3.24]	+4.88 [+0.71, +13.07]	+1.16 [-0.04, +3.44]

Table 11: Adding geometry to the density-proxy detector: learned LJF minus learned LF, FPR95 percentage points with saved exploratory 95% intervals. All six prespecified proxy families are retained. Positive AUROC differences and negative FPR95 differences favor LJF. These compare against the two-feature LF detector, not against ALL. Intervals use 1,000 model–target crossed-bootstrap draws within each benchmark after averaging calibration splits.

Proxy family	CIFAR-10	CIFAR-100	ImageNet
Tied Gaussian	-6.25 [-8.80, -3.08]	-14.38 [-38.80, -1.31]	-3.89 [-10.08, -0.08]
Diagonal GMM (10)	-9.40 [-13.38, -5.27]	-16.20 [-38.22, -3.48]	-5.95 [-12.24, -0.55]
Diagonal GMM (50)	-8.86 [-13.74, -3.81]	-15.25 [-38.53, -2.47]	-6.68 [-13.57, -0.63]
PCA64 + GMM (10)	-8.47 [-11.67, -4.75]	-14.82 [-38.72, -1.98]	-8.25 [-16.95, -0.97]
50-NN score	-7.56 [-14.39, -2.61]	-16.18 [-43.66, -0.59]	-7.07 [-14.38, -0.85]
Joint conditional Gaussian	-6.32 [-8.88, -3.12]	-14.51 [-39.15, -1.41]	-4.53 [-10.44, -0.36]

Table 12: Complete source-matched classifier results by OOD target. Each entry is a percentage: AUROC (↑) and FPR95 (↓). We first average the three calibration splits for each classifier–target cell, then average classifiers; n counts classifiers, not calibration splits. LJF uses the primary tied-Gaussian proxies. ALL fuses 27 source-selected score families, and ALL+J_ϵ adds one geometric statistic using the same calibration protocol. All 17 prespecified benchmark–target pairs are shown; no target is selected by its observed improvement. FPR95 uses the held-out-ID ROC threshold and is distinct from independently calibrated deployment acceptance.

			Learned LJF		ALL		ALL+J_ϵ	
ID benchmark	OOD target	n	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95
CIFAR-10	CIFAR-100	17	86.17	61.99	88.26	55.71	88.56	56.25
	Tiny ImageNet	17	87.28	55.94	88.49	51.60	89.22	50.86
	SVHN	17	90.44	58.56	97.68	12.35	97.91	10.95
	Textures	17	90.51	50.10	94.44	30.70	94.68	29.86
	MNIST	17	89.50	52.16	98.23	9.00	98.99	4.58
	Places365	17	89.33	47.37	91.79	38.64	92.41	37.23
CIFAR-100	CIFAR-10	12	75.20	84.10	79.53	80.68	79.67	80.50
	Tiny ImageNet	12	78.49	78.92	80.34	75.83	80.58	75.51
	SVHN	12	82.92	68.00	95.52	23.51	96.85	16.18
	Textures	12	82.31	65.66	90.21	43.41	90.49	42.39
	MNIST	12	95.89	17.67	99.17	3.35	99.69	1.15
	Places365	12	78.99	75.58	83.15	69.65	83.39	68.89
ImageNet	SSB-hard	11	65.73	88.52	81.92	60.59	82.14	60.36
	NINCO	11	72.74	84.17	90.41	44.83	90.42	44.80
	iNaturalist	11	89.78	41.33	98.56	6.12	98.64	5.78
	Textures	11	91.54	31.47	97.71	10.74	98.12	9.04
	OpenImage-O	11	83.05	60.70	95.31	24.91	95.34	24.80

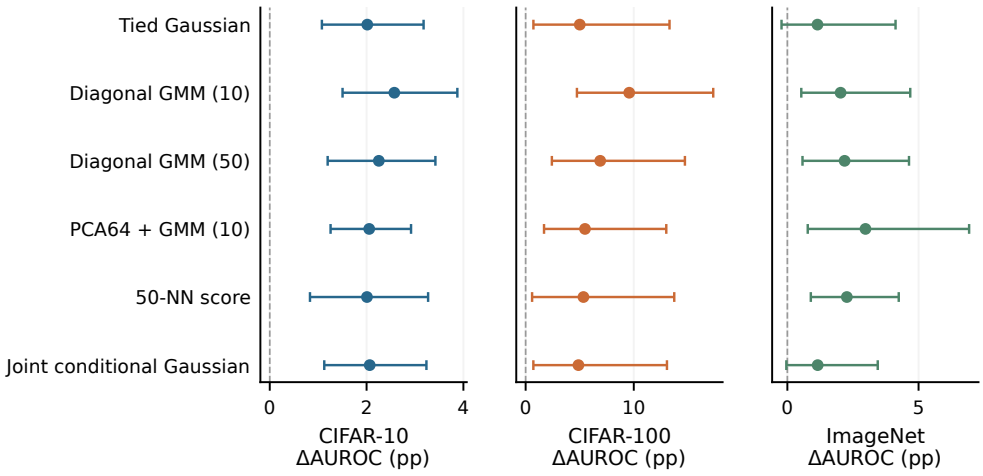


Figure 4: Adding J across six fixed density-estimator families. Points show mean AUROC(LJF)−AUROC(LF); bars show exploratory 95% crossed model–target bootstrap intervals after averaging calibration splits. All 18 means are positive; 16 intervals exclude zero. Panel scales differ.