WHEN EXACT LIKELIHOOD IS ANTI-ROBUST: DIAGNOSING AND FIXING DIFFUSION CLASSIFIERS UNDER DOMAIN SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative classifiers are widely believed to be more robust than their discriminative counterparts. We show that for diffusion classifiers that score classes by exact likelihood the opposite holds under domain shift, and that the failure is structural: it follows from how the likelihood weights noise levels. Decomposing the class-conditional negative log-likelihood $-\log p(\boldsymbol{x} \mid y)$ across noise levels via the information-theoretic diffusion identity, we find that the fraction of discriminative margin retained under shift rises steadily with noise level. Low-noise terms do not merely degrade out of distribution: classifying with them alone falls below chance, while high-noise terms retain roughly half of their in-domain margin. The exact likelihood integrates uniformly in log-SNR, which under the model's schedule places about three times the uniform-timestep weight on the lowest-noise bucket, precisely where evidence is most fragile. On the same model and the same per-noise-level denoising errors, the three scores order monotonically, exact $<$ uniform-timestep $<$ band-restricted, on all six Rotated MNIST folds and on CIFAR-10-C noise corruptions, where the exact likelihood is at chance and the uniform-timestep score used in practice reaches 36%. The diagnosis implies a one-line fix with no retraining: restrict the evidence integral to a noise band chosen without access to the test domain. On six-fold Rotated MNIST (DomainBed protocol), band restriction lifts a joint diffusion classifier from 57.9% to 92.7% mean OOD accuracy (31% to 87% on the hardest fold). In-domain validation alone selects a biased band; shifting it upward in noise by a fixed offset, as the diagnosis predicts, matches leave-one-domain-out selection while using only single-domain data. Transferred unchanged to CIFAR-10-C, the same rule recovers noise-family accuracy from 36% to 78% and lifts the repaired classifier above a matched ERM ResNet-18 on mean corruption accuracy (82% vs. 75%), though not above one trained with AugMix. Where restriction hurts, on fog, contrast, and the natural shift CIFAR-10.1, the damage sits in the high-noise band, so the per-bucket decomposition predicts which shifts the fix helps. For the shifts we study, these results reverse the robustness narrative for exact-likelihood classification and reduce the repair to a one-line change in the evidence integral.

1 INTRODUCTION

A generative classifier labels $\boldsymbol{x}$ by comparing class-conditional likelihoods $p(\boldsymbol{x} \mid y)$. With diffusion models providing tractable and increasingly exact likelihoods, a growing body of work argues that such classifiers inherit desirable robustness properties (Jaini et al., 2024; Li et al., 2023). The intuition is appealing: a model of everything about $\boldsymbol{x}$ should degrade gracefully when $\boldsymbol{x}$ changes.

We show the intuition fails in a specific, diagnosable way under *domain shift*. Our starting point is the information-theoretic diffusion (ITD) identity of Kong et al. (2023), which expresses ideal class log-likelihood differences as an integral of denoising-error differences over signal-to-noise ratio (SNR). In ϵ-prediction form the weighting is uniform in log-SNR. We use this weighting to decompose learned-denoiser scores into measurable per-noise-level terms and compare them with the uniform-timestep scores used in practical diffusion classifiers. Section 2 distinguishes the ideal identity from the estimated scores used here.

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.

Diagnosis. On six-fold Rotated MNIST under the DomainBed protocol, the per-noise-level decomposition reveals a strikingly regular structure (Figure 1): above the lowest bucket, the fraction of the in-domain discriminative margin retained out of distribution increases with noise level. The low-noise terms are not merely uninformative OOD; on most folds their evidence points the wrong way, yielding below-chance accuracy (7.7–9.4% against a 10% chance floor on 4 of 6 folds). These terms sit where the exact likelihood looks: integrating uniformly in log-SNR, it gives the lowest-noise bucket about three times the weight the uniform-timestep average does under the model's schedule (27% vs. 19% for the three lowest buckets together), and the corresponding ITD-weighted score transfers worse: on the same model and saved errors, it reaches 27% mean OOD accuracy on Rotated MNIST and chance on CIFAR-10-C noise corruptions, the uniform-timestep score 58% and 36%, and a band-restricted score 93% and 78%. The likelihood-derived weighting emphasizes a low-noise band whose class evidence is fragile under these shifts.

Fix. The diagnosis implies a one-line repair: restrict the evidence integral to a noise band chosen without access to the test domain. Band restriction lifts a joint diffusion classifier from 57.9% to 92.7% mean OOD accuracy, and from 31% to 87% on the hardest extrapolation fold, with zero retraining. Crucially, the OOD-optimal band is systematically shifted *upward* in noise relative to the in-domain optimum: selecting the band by in-domain validation alone collapses to 75.0% on the hardest fold, whereas shifting the in-domain band upward matches leave-one-domain-out selection (92.7% mean for both). The $+2$ offset is selected using these Rotated MNIST OOD results and then locked for transfer to CIFAR-10-C; only the direction of the shift is predicted by the diagnosis.

Contributions.

- A per-noise-level decomposition of diffusion classifier evidence that mechanistically localizes OOD collapse in low-noise likelihood terms, including a below-chance reversal of low-noise evidence (Section 3).
- A controlled comparison of ITD-weighted, uniform-timestep, and band-restricted scores, showing how likelihood-derived weighting can reduce robustness when low-noise evidence is fragile (Section 3).
- A zero-retraining band-restriction fix with LODO band selection and a clean-selected band with a fixed upshift, including an explicit separation between offset development on Rotated MNIST and transfer to CIFAR-10-C (Sections 4, 5).

We position this paper as a diagnosis first and a fix second: the band restriction is a corollary of understanding *where* in noise space domain-transferable class evidence lives, not a new architecture or training method.

Scope. Our empirical claims concern learned-denoiser scores under rotation (Rotated MNIST) and common image corruptions (CIFAR-10-C), with CIFAR-10.1 as a natural-shift boundary check. The ITD identity motivates the weighting, but the premise that low-noise evidence is fragile is empirical and shift-dependent (Section 5.3). We do not claim anything about other generative model families, other modalities, or shifts such as spurious correlation or label shift.

2 Background

Diffusion classifiers. A generative classifier predicts $\hat{y} = \arg\max_y \log p(\boldsymbol{x} \mid y) + \log p(y)$. Practical diffusion classifiers approximate class evidence using denoising errors, commonly averaged uniformly over timesteps (Li et al., 2023). We compare this score with the log-SNR weighting motivated by the information-theoretic diffusion (ITD) identity (Kong et al., 2023), using equal class priors throughout.

The ITD identity and class comparisons. Let $\alpha = \log \gamma$, $\boldsymbol{z}_\alpha = \sqrt{\gamma}\boldsymbol{x} + \epsilon$, and $\epsilon \sim \mathcal{N}(0, I)$. For the optimal class-conditional noise predictor, define $\ell_y^*(\boldsymbol{x}, \alpha) = \mathbb{E}_\epsilon \|\epsilon - \hat{\epsilon}_y^*(\boldsymbol{z}_\alpha, \alpha)\|^2$. For continuous densities, the ITD NLL identity includes a Gaussian reference correction; its uncorrected integral is not generally finite (Kong et al., 2023, Eq. 9). For comparing classes, a common reference cancels. Under the identity's regularity and convergence conditions,

$$\log \frac{p(\boldsymbol{x} \mid y')}{p(\boldsymbol{x} \mid y)} = \frac{1}{2} \int_{-\infty}^{\infty} \left[\ell_y^*(\boldsymbol{x}, \alpha) - \ell_{y'}^*(\boldsymbol{x}, \alpha) \right] d\alpha. \tag{1}$$

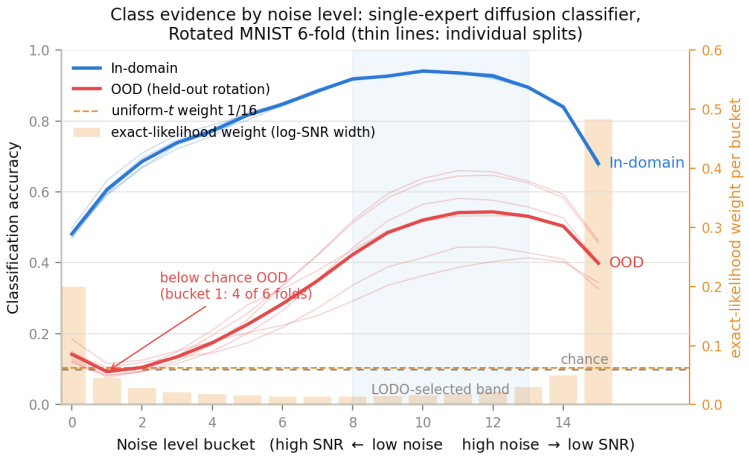


Figure 1: Per-noise-bucket classification accuracy of the per-domain expert diffusion classifiers (mean over the five experts of each fold) in domain (blue) and on the held-out rotation (red), Rotated MNIST, 6-fold LODO (thin lines: individual folds; bold: mean). The exact-likelihood weight per bucket (orange, right axis; log-SNR width share, dashed: the uniform-t $1/16$) puts 19.9% on bucket 0, three times its uniform-t share, where OOD accuracy is low and nearly class-blind, and 48% on the class-uninformative SNR$\to$0 bucket; bucket 1 falls *below* the 10% chance line on 4 of 6 folds (callout), and the LODO-selected band (shaded) lies where OOD accuracy peaks.

The discrete-probability identity gives the analogous class difference (Kong et al., 2023, Eq. 12). Thus ideal likelihood comparisons weight noise-prediction errors uniformly in log-SNR. Equivalently, $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 = \gamma^{-1}\|\epsilon - \hat{\epsilon}\|^2$ converts the SNR integral of reconstruction errors to a log-SNR integral of noise errors.

Our experiments substitute learned denoisers and use a finite schedule and numerical quadrature. We call the resulting score *ITD-weighted*, or "exact" as shorthand for its likelihood-derived weighting. This does not assert that the score is the exact log likelihood of the fitted DDPM. Changing the weighting tests the behavior of these fitted models; it does not establish the behavior of an optimal denoiser. Uniform-timestep scoring uses a different measure, related by the schedule's Jacobian $|d\alpha/dt|$ (Kingma et al., 2021).

Discretization and the three scores. We partition the schedule into $B{=}16$ equal-width buckets in t, from lowest to highest noise, and estimate bucket-mean noise errors $L_b(\boldsymbol{x}, y)$ (Appendix A). The margin assigned to the true class y over a competing class y' is $m_b = L_b(\boldsymbol{x}, y') - L_b(\boldsymbol{x}, y)$. The *ITD-weighted* score is $\sum_b w_b L_b$, with $w_b \propto \Delta\alpha_b$, the log-SNR width of each bucket. This bucketwise approximation is checked against per-timestep quadrature in Appendix D. The *uniform-t* score is $B^{-1}\sum_b L_b$; a *band* score is $|\mathcal{B}|^{-1}\sum_{b\in\mathcal{B}} L_b$. All three use the same denoising errors and differ only in their weights.

3 Diagnosis: Where Class Evidence Survives Domain Shift

Setup. We train per-domain expert diffusion models and a joint model on Rotated MNIST (angles $0°$–$75°$; DomainBed protocol, Gulrajani & Lopez-Paz, 2021), hold out one domain per fold, and measure per-bucket margins m_b on held-in (in-domain) and held-out (OOD) data.

Finding 1: retention rises with noise. Define per-bucket retention as the ratio of OOD to in-domain margin. Above the lowest bucket, retention increases with noise level in all six folds (up to a single 0.005-magnitude inversion on the $30°$ fold): buckets 1–2 retain none of their in-domain margin OOD — their retention is negative on 6 and 5 of the 6 folds respectively — bucket 3 retains at most 0.18, while high-noise buckets retain roughly half. Bucket 0's ratio (0.37–0.46) breaks the

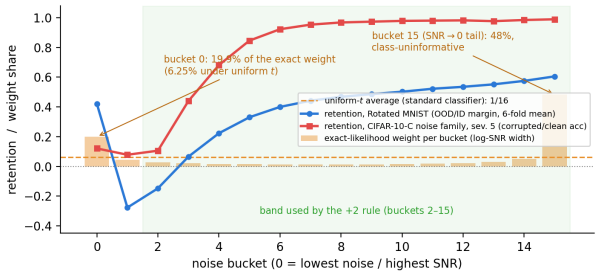


Figure 2: Where exact likelihood looks vs. where evidence survives. Orange bars: weight of each t-bucket in the exact-likelihood quadrature (its log-SNR width; $T{=}1000$ cosine schedule) against the flat $1/16$ of the uniform-t average (dashed). The lowest-noise bucket carries 19.9% of the exact weight, three times its uniform-t share; the highest-noise bucket carries 48% but its ϵ-error is class-independent. Blue: per-bucket retention on Rotated MNIST (OOD/in-domain margin ratio, 6-fold mean); red: retention of the CIFAR-10-C noise family at severity 5 (corrupted/clean accuracy).

trend, but a ratio of two near-class-independent margins does not measure transfer (Appendix C); the argument below uses its OOD accuracy instead.

Finding 2: low-noise evidence can reverse sign. On 4 of 6 folds, classifying with bucket-1 evidence alone gives 7.7–9.4% OOD accuracy, below the 10% chance reference. This is a descriptive result across folds, not a significance test formed by pooling only the below-chance folds. Together with the negative margin retention, it shows that some low-noise terms favor wrong classes under shift.

Finding 3: band and single-expert results. The joint classifier with a selected band reaches 82.9–97.0% across folds, while the best single bucket of the per-domain experts, averaged over experts, reaches 41.4–66.0% (Figure 1). Because these classifiers have different training domains, this comparison alone does not isolate the benefit of aggregating multiple buckets. The same-model score comparisons below provide the controlled evidence for changing the weighting.

The anti-robustness argument. Equation 1 weights class-conditional error differences uniformly in log-SNR, motivating a bucket weight proportional to its log-SNR width. Under the $T{=}1000$ cosine schedule used throughout, log-SNR stretches at the clean end: bucket 0 spans $\alpha \in [4.4, 10.1]$ and carries 19.9% of the exact weight — 38.5% of the weight outside the class-uninformative SNR$\to$0 tail — against $1/16 = 6.25\%$ under the uniform-t average (Figure 2). The ITD-derived weighting therefore up-weights, by about a factor of three, a bucket that is nearly class-blind out of distribution: bucket 0 classifies at 14% OOD against 48% in domain (Figure 1; 8.0% on the CIFAR-10-C noise family averaged over severities, below chance and the worst single bucket), and the next two buckets reverse sign (6-fold mean retention -0.28 and -0.15). What the weighting controls is leverage over the decision, not where the quadrature mass sits: the SNR$\to$0 bucket carries 48% of the mass but a nearly class-independent ϵ-error, whereas under the exact weights bucket 0 supplies 34% of the in-domain and 37% of the OOD class margin $\sum_b w_b m_b$, against 5% under uniform-t. The consequence is directly testable on the same saved errors, and it holds: exact quadrature $<$ uniform-t average $<$ band, at 27.2%, 57.9%, and 92.7% mean on Rotated MNIST and 9.8%, 36.3%, and 77.8% on the CIFAR-10-C noise family (Tables 1 and 3). Two controls pin the cause to bucket 0 (Appendix B): dropping it from the exact score (weights renormalized) restores 50.1% on Rotated MNIST and 40.0% on the CIFAR-10-C noise family, whereas dropping the 48%-weight high-noise bucket does not repair it (24.5% and 9.6%); and although the weight shares depend on where the schedule truncates log-SNR (bucket 0 carries 20–34% across three truncations), the exact score's OOD accuracy does not (24.9–27.2% and 9.6–9.8%). These controls isolate sensitivity to score weighting for fixed learned models. Uniform-t scoring down-weights the low-noise endpoint relative to ITD weighting; it is a different weighting scheme, not a numerical refinement of the same integral. The observed ordering therefore does not imply that increasing quadrature accuracy reduces robustness. The likelihood-derived weighting is theoretical, whereas the fragility of low-noise evidence is an empirical, shift-dependent premise (Section 5.3).

Table 1: Mean OOD accuracy over six LODO folds of Rotated MNIST. The three joint-classifier rows differ only in how the same saved per-bucket errors are weighted; no model is retrained.

Method	Mean OOD acc (%)
ERM CNN (discriminative reference)	97.1
Joint diffusion classifier, exact likelihood (ITD quadrature)	27.2
Joint diffusion classifier, uniform-t average	57.9
Joint + noise band (LODO), **ours**	**92.7**
Best single expert + band	89.6
Posterior-composed experts + band	79.9

4 FIX: BAND-RESTRICTED EVIDENCE

We replace full-schedule scoring with a band-restricted sum,

$$\hat{y} = \arg\max_y -\sum_{b\in\mathcal{B}} L_b(\boldsymbol{x}, y), \qquad \mathcal{B} \subseteq \{0, \dots, B-1\}, \tag{2}$$

using uniform weights within the band and no retraining. This changes both the support and, relative to the ITD-weighted score, the weights. We distinguish selection of the band from development of the offset rule.

LODO selection. With multiple training domains, select the contiguous band maximizing leave-one-domain-out validation accuracy across training domains, without using the final test domain. We search all 136 contiguous bands of B=16 buckets.

Clean-selected band with a fixed upshift. First select a band using labeled in-domain data, then shift both endpoints toward higher noise by k buckets, clipping at bucket 15. The diagnosis motivates the direction; k=2 is chosen from the Rotated MNIST six-fold OOD sweep and matches its LODO mean. It is therefore an in-sample offset choice on that benchmark, not a validation-free result. For CIFAR-10-C we lock k=2 and select the initial band using clean CIFAR-10 images and labels, with no corruption labels used by this rule. The clean test split serves as calibration data (Appendix A); clean accuracy after band selection is therefore not an independent test estimate. Corrupted-validation band selection is reported separately as a comparison.

5 EXPERIMENTS

5.1 ROTATED MNIST (DOMAINBED, 6-FOLD LODO)

Table 1 gives the headline result: band restriction recovers 34.8 points of mean OOD accuracy for the identical joint model. The largest per-fold gains are at 60° and 75°: 41%→97% and 31%→87%. Interpolation folds come within 1.4–2.2 points of the ERM reference; the extrapolation folds (0°, 75°) remain 8.9–10.1 points below it — the fix removes the collapse, it does not make the generative classifier dominant (see Section 6 for framing). Selected bands are stable across folds (buckets 8–13 of 16).

5.2 SEVERING IN-DOMAIN TUNING FROM OOD-SAFE SELECTION

A natural reduction of our fix is "this is just timestep tuning on in-domain validation." The severing experiment separates the two.

In-domain selection chooses wide bands that include low buckets (e.g. 5–15 on the 75° fold) and collapses where shift is largest; LODO nearly matches the oracle; and the upshift rule matches LODO in the mean (Figure 3 shows this fold by fold). The offset +2 is the one that maximizes the six-fold mean (92.7%; +1: 92.3%, +3: 92.5%, Appendix B), so on Rotated MNIST it is an in-sample choice. A cross-fold offset sensitivity check selects k using the other five folds and applies it to the remaining fold, giving 91.9%, within 0.8 points of the global k=2. Every such selection shifts

Table 2: Band-selection rules on Rotated MNIST (mean over 6 folds and the hardest extrapolation fold). In-domain validation actively rewards fragile low-noise buckets; the $+2$ upshift closes the mean gap. Its offset is selected on these Rotated MNIST OOD folds and then fixed for CIFAR-10-C.

Band selection	Mean (%)	75° fold (%)
In-domain validation	90.7	75.0
In-domain $+2$ upshift (offset tuned here)	92.7	84.4
LODO	92.7	86.8
Oracle (test-selected)	93.4	86.9

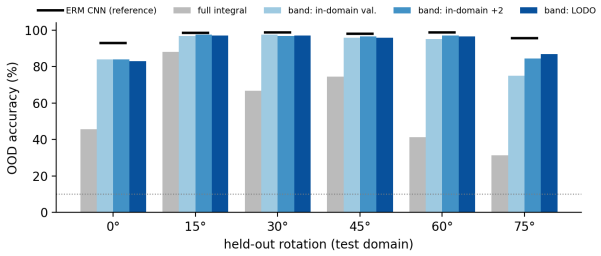


Figure 3: Rotated MNIST, per fold: OOD accuracy of the joint diffusion classifier with the uniform-t average and with each band-selection rule, against the ERM reference (black bars). Every band rule repairs the collapse on every fold; the rules separate mainly on the 75° extrapolation fold, where in-domain validation keeps fragile low buckets and the upshift recovers most of the gap to LODO. Per-fold values in Appendix C.

toward higher noise. This check is not presented as a fully nested retraining protocol; CIFAR-10-C below tests transfer of the locked $k{=}2$. The OOD-optimal band is consistently shifted up from the in-domain optimum, as the diagnosis predicts.

5.3 CIFAR-10-C

Setup. We train a class-conditional DDPM (13.3M parameters) and an ERM ResNet-18 on clean CIFAR-10 and evaluate on CIFAR-10-C (Hendrycks & Dietterich, 2019): 15 standard test corruptions at 5 severities, with the benchmark's 4 hold-out corruptions (gaussian_blur, saturate, spatter, speckle_noise) reserved for the band-selection ablation below and excluded from all reported means. Seed 0 uses the full 10,000-image split per setting; seeds 1–2 use 2,048-image subsets (Appendix A). Table 3 reports seed-0 scores with cross-run standard deviations. Clean accuracy is 95.1% (ERM), 94.0% (uniform-t average), and 91.9% (exact score): on clean data neither score is the problem.

The diagnosis generalizes — with a mechanism-revealing split. Appendix Figure 6 shows the Rotated MNIST retention structure reproduced on natural images: corruption damage concentrates in low-noise buckets and scales with severity, while mid-to-high buckets are nearly untouched. The split by corruption type sharpens the mechanism. For corruptions in the *noise family* — those resembling the diffusion forward process — the uniform-t average collapses to chance at the highest severity (10.5% on impulse noise, 12.3% on Gaussian noise) while buckets 9 and above retain at least 94% of their clean accuracy (at least 96% from bucket 6 on for Gaussian and shot noise; Figure 6b): the corruption directly poisons the low-noise denoising errors that the exact weighting emphasizes. For blur, weather, and digital corruptions the low buckets are degraded far less, and the uniform-t average is competitive with or better than ERM (83.7% vs. 80.0%). Fog and contrast are the mirror image (Figure 6c): the lowest-noise bucket retains 86–88% of its clean accuracy while the mid-band buckets retain only 20–32%, because these corruptions alter global, low-frequency image statistics rather than the pixel-scale detail that low-noise denoising reads — and, consistently, the exact score *beats* the uniform-t average on exactly these two corruptions (contrast at severities 2–5, fog at 3–5; contrast at severity 5: 69.2% vs. 41.6%) and nowhere else. Which band a shift destroys is thus a property of the shift, and the failure of exact-likelihood classification is not "corruption hurts

Table 3: CIFAR-10-C mean accuracy (%) over the 15 standard corruptions × 5 severities (the full N=10,000 images per setting, seed-0 models; ± is the SD across three training seeds, with bands re-selected per seed; seeds 1–2 use 2,048-image subsets and different evaluation noise, so these are seed-0 values with cross-run SD, not three-seed means), split into the noise family (gaussian/shot/impulse) and the remaining 12 corruptions. The AugMix row is the same ResNet-18 trained with AugMix augmentation (single seed), included as the stronger discriminative reference. Bands are selected without access to any of the 15 test corruptions: on clean data, on clean data with the fixed +2 upshift, or on the benchmark's 4 hold-out validation corruptions — the last of which leaks corruption-family information through speckle noise (see the selection ablation below) and is reported for comparison only.

Method	Band	Mean	Noise fam.	Non-noise
ERM ResNet-18	—	75.0 ±0.6	54.8 ±1.6	80.0 ±0.3
ERM ResNet-18 + AugMix (single seed)	—	85.3	77.8	87.1
Diffusion classifier, exact likelihood (ITD quadrature)	0–15	55.9 ±0.5	9.8 ±0.6	67.4 ±0.5
Diffusion classifier, uniform-t average	0–15	74.2 ±0.3	36.3 ±0.3	83.7 ±0.4
band: clean-selected	0–15	74.2 ±0.5	36.3 ±0.2	83.7 ±0.5
band: clean-selected +2 (fixed offset)	2–15	81.8 ±0.3	77.8 ±0.3	82.8 ±0.4
band: hold-out-selected (leaks noise fam.)	3–15	82.6 ±0.3	85.6 ±0.1	81.8 ±0.4
band: oracle (upper bound)	3–15	82.6 ±0.3	85.6 ±0.1	81.8 ±0.4

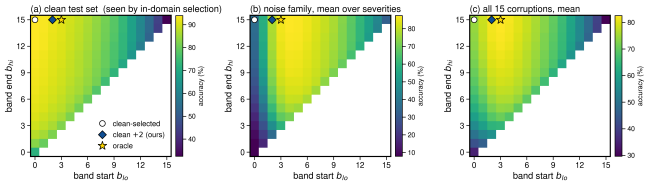


Figure 4: Accuracy of every contiguous band $[b_{lo}, b_{hi}]$ of the 16 buckets on CIFAR-10-C (N=10,000 per setting). (a) On clean data the landscape is flat across the top-left — including the low buckets costs nothing in domain, so in-domain selection picks the full band (circle). (b) On the noise family the same low buckets are catastrophic: every band starting at 0 or 1 sits in the dark column, and the optimum starts at bucket 4. (c) Averaged over all 15 corruptions the optimum (star) is band 3–15; the clean band shifted by +2 (diamond) lands next to it. The shape of the landscape, not its argmax, is what the upshift rule exploits.

generative models": it is specifically the concentration of likelihood weight on the noise scales that the shift corrupts. Read on a natural shift, CIFAR-10.1, where low-noise evidence is well preserved, the profile is consistent with the observed cost of removing it: the locked band falls below the uniform-t score (83.4% vs. 86.8%; Appendix E).

The fix generalizes — and the in-domain trap reappears. Table 3 mirrors the severing analysis of Section 5.2 exactly, and Figure 4 shows why. On clean data (panel a) the accuracy landscape over bands is flat wherever the band reaches the mid-noise buckets: including buckets 0–2 costs nothing in domain, so band selection on clean data picks the *full* band (0–15) — in-domain selection again fails to exclude the fragile low buckets, and inherits the uniform-t score's collapse (36.3% on the noise family). On the corrupted data (panels b–c) the same low buckets form a dark column: any band that starts at 0 or 1 is poisoned, and the optimum starts at bucket 4 on the noise family and at bucket 3 over all 15 corruptions. Applying the same +2 upshift used on Rotated MNIST recovers the bulk of the collapse (77.8% on the noise family) at a 1.0-point mean cost on non-noise corruptions — a cost that is not uniform and is consistent with the decomposition: on the two corruptions whose damage sits in the high-noise band (Figure 6c) the restriction discards the buckets that still work (contrast at severity 4 drops 74.9% → 54.2%; fog at severity 5, 58.0% → 41.4%), so the band trades worst-case non-noise accuracy for noise-family recovery rather than dominating outright. With this fixed-offset rule, the band-restricted diffusion classifier exceeds the ERM ResNet-18 trained on clean CIFAR-10 in mean corruption accuracy (81.8% vs. 75.0%, stable across seeds). The advantage is carried by the noise family; on the ten non-noise corruptions other than fog and contrast the band is

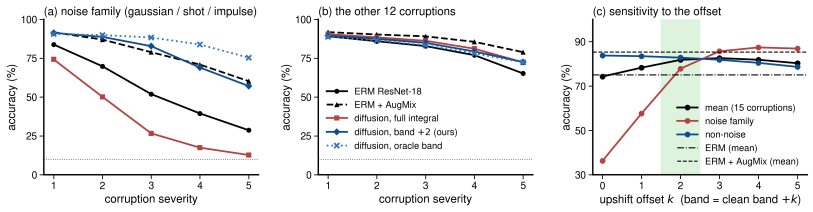


Figure 5: CIFAR-10-C by severity (N=10,000 per setting, seed 0). (a) Noise family: the uniform-t average collapses toward chance (dotted) as severity grows; the $+2$ band (blue) removes the collapse and follows the AugMix reference. (b) The other twelve corruptions: no collapse, and the band costs a few points. (c) Sensitivity of the upshift rule to the offset k: every tested $k \in \{2, 3, 4, 5\}$ beats the matched ERM (dash-dot) on the mean; the noise/non-noise trade-off moves smoothly with k.

Table 4: Which validation data can select the band? Band chosen by maximizing accuracy on the listed corruptions (all severities), and the resulting accuracy on the 15 test corruptions. Only validation corruptions that themselves damage the low-noise buckets find the repair.

Selection data	Band	Noise fam.	Mean (15)
All 4 hold-outs (blur, saturate, spatter, speckle)	3–15	85.6	82.6
3 non-noise hold-outs (speckle removed)	1–15	57.6	78.3
speckle noise alone	3–15	85.6	82.6
gaussian blur alone	1–15	57.6	78.3
saturate alone	0–15	36.3	74.2
spatter alone	2–15	77.8	81.8
oracle (15 test corruptions)	3–15	85.6	82.6

within a point of the uniform-t score (85.6% vs. 84.7%; ERM 79.2%). The advantage grows with severity, from $+1.8$ points at severity 1 to $+11.5$ at severity 5 (flat at $+7.8$ between severities 3 and 4) (Appendix C). The ERM reference is not the strongest discriminative baseline: the same ResNet-18 architecture trained with AugMix (Hendrycks et al., 2020) reaches 85.3% mean accuracy, ties the band-restricted classifier on the noise family (77.8% vs. 77.8%; the band wins on Gaussian and shot noise, 86.7% and 87.4% vs. 72.9% and 81.1%, AugMix on impulse noise, 79.2% vs. 59.3%) and leads on non-noise corruptions (87.1% vs. 82.8%). Augmentation is an orthogonal axis — it changes the training distribution, whereas the band restriction changes only how a fixed likelihood model is read — and the claim here is that the diagnosis repairs a structural collapse, not that a repaired generative classifier is the strongest CIFAR-10-C classifier available. The two are also not exclusive: nothing prevents training the diffusion model with augmentation as well. Figure 5 lays the comparison out by severity: on the noise family (panel a) the uniform-t score falls from 74% to 13% while the band-restricted classifier tracks AugMix and stays above the matched ERM at every severity; on the other twelve corruptions (panel b) all classifiers degrade together and the band costs at most 1.8 points relative to the uniform-t average (and nothing at severity 5).

Which validation corruptions can select the band? Selecting the band on the benchmark's 4 hold-out corruptions recovers even more (85.6% noise family, 82.6% mean, band 3–15 = oracle) — but Table 4 shows this rule *leaks*: its success is driven by speckle noise, the one hold-out corruption from the noise family. A validation corruption identifies the fragile buckets to the extent that it damages the same low-noise scales the test shift damages; corrupted-validation selection is therefore shift-family tuning, and the upshift of the clean-selected band, which sees no corruption, is the paper's rule.

5.4 Sensitivity to the upshift offset

The $+2$ offset is selected on Rotated MNIST and transferred to CIFAR-10-C unchanged; the theory predicts its direction, not its magnitude. The rule is nonetheless insensitive to the exact value (Figure 5c): offsets 0 through $+5$ give 74.2%, 78.3%, 81.8%, 82.6%, 81.8%, and 80.3% mean accuracy,

every tested $k \in \{2, 3, 4, 5\}$ above the ERM reference by 5.3–7.6 points, with the noise/non-noise trade-off moving smoothly with the offset — a graceful knob rather than a brittle hyperparameter.

6 Related work

Generative classifiers and robustness. Jaini et al. (2024) and Chen et al. (2024b;a) argue that diffusion-based generative classifiers are robust: human-like shape bias, OOD behavior, and empirical or certified adversarial robustness (an earlier score-based study found little robustness to common corruptions, Zimmermann et al., 2021). Our results qualify this narrative along the domain-shift axis: on the noise-family shifts we study, ITD weighting is less robust than uniform-t scoring of the same learned errors. The latter reduces the relative influence of the fragile low-noise terms. Our CIFAR-10-C results locate the boundary of that narrative: on blur, weather, and digital corruptions the uniform-t diffusion classifier does beat ERM (83.7% vs. 80.0%), but on the noise family — the shifts that most resemble the forward process — it collapses to 36.3% while ERM retains 54.8%. The robustness is real but band-specific, and the per-bucket decomposition helps explain which of the evaluated shifts break it. That a denoiser run at moderate noise is robust to additive noise is also the premise of denoised smoothing (Salman et al., 2020) and diffusion purification (Nie et al., 2022); our point is the converse: the likelihood integral is dominated by the low-noise terms this robustness does not protect.

Information-theoretic denoising. Yu et al. (2026) develop MMG to estimate mutual information by integrating the MMSE gap between unconditional and conditional denoisers over SNR. The present work studies how noise-level weighting of class-conditional denoising errors affects classification under domain shift.

Diffusion classifiers and timestep choice. Li et al. (2023) and Clark & Jaini (2023) observe that intermediate timesteps are most accurate for in-domain classification, and Wang & Chen (2025) optimize the noise draw at fixed timesteps — in-domain heuristics with no OOD diagnosis or likelihood-level explanation. Jeong et al. (2025) establish an empirical relationship between domain gap and timestep sensitivity for large diffusion classifiers and show that timestep re-weighting helps; we complement that observation by decomposing class-conditional likelihood evidence across noise scales and studying how individual contributions change sign and transfer under shift. Other work exploits diffusion score structure for test-time adaptation under corruption (Li et al., 2024), and mines noise-robust diffusion features for domain-generalized detection (He et al., 2025); Fathi et al. (2026) evaluate their in-domain decision biases. What this literature lacks, and what we add, is (i) the per-bucket localization of OOD collapse in low-noise evidence, including its below-chance sign reversal, (ii) the connection to the exact likelihood via the ITD identity (Kong et al., 2023), and (iii) band selection with an empirically calibrated upshift whose direction is motivated by the diagnosis and whose transfer is evaluated separately.

7 Conclusion

For the learned diffusion classifiers and shifts we study, ITD-derived weighting can be a liability. Relative to uniform-t scoring, it emphasizes low-noise errors whose class evidence degrades or reverses under rotation and noise corruptions. Controlled comparisons on the same denoising errors link this weighting to lower OOD accuracy. Restricting the score to a selected band repairs much of the collapse without retraining. An offset developed on Rotated MNIST transfers to CIFAR-10-C, while fog, contrast, and CIFAR-10.1 demonstrate its limits.

Limitations. The identity describes optimal-denoiser likelihood comparisons; our experiments use learned denoisers, finite noise ranges, and numerical quadrature. They establish a score-weighting effect, not that better likelihood estimation or better generative modeling necessarily reduces robustness. The diagnosis motivates the upshift direction, while its magnitude is selected using Rotated MNIST OOD results. CIFAR band calibration uses clean test labels. Our evidence is limited to the tested image datasets and models; ERM remains stronger on Rotated MNIST, and AugMix leads on mean CIFAR-10-C accuracy. A fixed band can hurt when low-noise evidence survives or higher-noise evidence is damaged.

ETHICS STATEMENT

This work studies the robustness of image classifiers on standard public benchmarks (Rotated MNIST, CIFAR-10-C, CIFAR-10.1) and involves no human subjects, personal data, or foreseeable direct misuse. Understanding when generative classifiers fail under distribution shift is relevant to the safe deployment of such models.

REPRODUCIBILITY STATEMENT

All experiments use public datasets; Rotated MNIST follows the DomainBed evaluation protocol, and CIFAR-10-C and CIFAR-10.1 use their standard evaluation splits. Model architectures, noise discretization, band-selection procedures, and evaluation budgets are specified in Section 5 and Appendix A. Code, trained checkpoints, and the per-noise-level loss tensors from which every table and figure is computed will be released upon publication; every table in the paper is generated from those tensors by the released analysis scripts, and every number in the text is checked against them.

AI USE STATEMENT

We used an LLM-based coding assistant throughout this project: to write and run training, evaluation, and analysis code; to orchestrate and monitor experiment queues; to generate figures and tables from the saved result files; to draft and edit the manuscript; to search the literature; and to perform adversarial self-review of the draft (simulated reviewer critiques and an LLM-assisted consistency check of every reported number against the raw result files). The research question, the experimental design and its gating decisions, the interpretation of the results, and the claims made in this paper are the authors' own. All AI-assisted code was reviewed by the authors; every table is generated by released scripts from the raw per-noise-level loss tensors, and every number in the text was checked against those files. The authors take full responsibility for the content of this work.

REFERENCES

Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Your diffusion model is secretly a certifiably robust classifier. *CoRR*, abs/2402.02316, 2024a. doi: 10.48550/ARXIV.2402.02316. URL https://doi.org/10.48550/arXiv.2402.02316.

Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6643–6665. PMLR / OpenReview.net, 2024b. URL https://proceedings.mlr.press/v235/chen24k.html.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/b87bdcf963cad3d0b265fcb78ae7d11e-Abstract-Conference.html.

Saba Fathi, Fardin Ayar, Maryam Abdolali, Ehsan Javanmardi, Manabu Tsukada, and Mahdi Javanmardi. How do diffusion classifiers decide? A bias-centric evaluation. *CoRR*, abs/2607.03831, 2026. doi: 10.48550/ARXIV.2607.03831. URL https://doi.org/10.48550/arXiv.2607.03831.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Boyong He, Yuxiang Ji, Qianwen Ye, Zhuoyue Tan, and Liaoni Wu. Generalized diffusion detector: Mining robust features from diffusion models for domain-generalized detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1912.02781.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=rmg0qMKYRQ.

Yujin Jeong, Arnas Uselis, Seong Joon Oh, and Anna Rohrbach. Diffusion classifiers understand compositionality, but conditions apply. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/b40004978c235a2dbecbe6d985f78790-Abstract-Datasets_and_Benchmarks_Track.html.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 21696–21707, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b578f2a52a0229873fefc2a4b06377fa-Abstract.html.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=UvmDCdSPDOW.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2206–2217. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00210.

Mingjia Li, Shuang Li, Tongrui Su, Longhui Yuan, Jian Liang, and Wei Li. Exploring structured semantic priors underlying diffusion score for test-time adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL http://proceedings.mlr.press/v139/nichol21a.html.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827. PMLR, 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to cifar-10? *CoRR*, abs/1806.00451, 2018. URL http://arxiv.org/abs/1806.00451.

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J. Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f9fd2624beefbc7808e4e405d73f57ab-Abstract.html.

Yanghao Wang and Long Chen. Noise matters: Optimizing matching noise for diffusion classifiers. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/bef2af7a1199ec7a134b15ac00bd5377-Abstract-Conference.html.

Longxuan Yu, Xing Shi, Xianghao Kong, Tong Jia, and Greg Ver Steeg. MMG: Mutual information estimation via the MMSE gap in diffusion. In *Forty-Second Annual Conference on Uncertainty in Artificial Intelligence*, 2026. URL https://openreview.net/forum?id=qMHdwhu4kb.

Roland S. Zimmermann, Lukas Schott, Yang Song, Benjamin A. Dunn, and David A. Klindt. Score-based generative classifiers. *CoRR*, abs/2110.00473, 2021. URL https://arxiv.org/abs/2110.00473.

A Experimental details

Noise discretization and per-bucket losses. All diffusion models use $T{=}1000$ discrete steps of a cosine variance-preserving schedule (Ho et al., 2020; Nichol & Dhariwal, 2021). For an image $\boldsymbol{x}$ and class y we draw M stratified timesteps (midpoints of M equal-width intervals of $[0, T)$, one Gaussian noise sample each, fixed by a data seed) and compute the conditional denoising error at each; the M errors are grouped into $B{=}16$ equal-width buckets in t (bucket 0 = lowest noise, highest SNR) and averaged within each bucket to give $L_b(\boldsymbol{x}, y)$. The exact score uses weights $w_b \propto \Delta\alpha_b = \alpha(t_b) - \alpha(t_{b+1})$, the log-SNR width of the bucket under the schedule (α ranges from 10.1 at $t{=}0$ to -18.4 at $t{=}T{-}1$; widths 5.69, 1.30, 0.80, 0.61, 0.49, 0.44, 0.40, 0.39, 0.39, 0.41, 0.44, 0.51, 0.62, 0.85, 1.41, 13.75); the uniform-t score uses $w_b = 1/B$. For each fixed model and evaluation budget, the three scores and band selection rules are computed from the same saved tensor $L \in \mathbb{R}^{N\times C\times B}$, without rerunning the model for each score. Appendix D separately evaluates per-timestep errors at the stated budgets before regrouping them into buckets. $M{=}128$ on Rotated MNIST and $M{=}64$ on CIFAR-10-C. Scoring one image costs $C \times M$ conditional U-Net evaluations (640 on CIFAR-10-C, 1,280 on Rotated MNIST) against a single ResNet-18 forward pass — the usual cost of diffusion classifiers, and not the axis compared here; a band uses $|\mathcal{B}|/B$ of that budget (560 for buckets 2–15).

Rotated MNIST. The 70k MNIST images are split into six disjoint chunks, one per angle in $\{0°, 15°, \ldots, 75°\}$, following DomainBed. Per-domain experts and the joint model are class-conditional U-Nets trained for 30 epochs (AdamW, learning rate $2{\times}10^{-4}$, batch 256, cosine decay); the ERM reference is a CNN trained for 10 epochs (AdamW, 10^{-3}) on the five training domains, with the final checkpoint used as is (no test-domain model selection). Every classifier, ERM included, is scored on the same 2,000 test images from the held-out domain (single training seed); LODO and in-domain band selection use 400 validation images per training domain.

CIFAR-10-C. The class-conditional DDPM is a 13.3M-parameter U-Net (base width 128, self-attention at 16×16 and 8×8, class and time embeddings of width 512) trained for 400 epochs on the 50k clean training images (AdamW, $2{\times}10^{-4}$, batch 128, EMA weights used at evaluation).

The ERM reference is a ResNet-18 trained for 100 epochs (SGD, learning rate 0.1, momentum 0.9, cosine decay, batch 128, standard crop/flip augmentation); clean test accuracy 95.1%. The AugMix reference uses the identical recipe with torchvision's AugMix transform applied on top of crop/flip (single seed; clean accuracy 95.4%). For seed 0 every corruption/severity setting is evaluated on its full 10,000-image CIFAR-10-C split; seeds 1 and 2 retrain both models from scratch and evaluate a fixed 2,048-image subset with the noise seed tied to the training seed. Bands are re-selected for each seed. Selection rules: *clean* maximizes accuracy over the 136 contiguous bands using the labeled clean test split as calibration data; *clean*+2 shifts that selected band's endpoints upward by two buckets, clipping at 15, with the offset fixed from Rotated MNIST. The clean split therefore does not provide independent test accuracy for selected bands. *Hold-out* searches the same bands using the four hold-out corruptions at all severities; *oracle* searches them using the 15 standard test corruptions and is reported only as a reference.

B THE EXACT-LIKELIHOOD SCORE: TRUNCATION SENSITIVITY AND ABLATIONS

Table 5: The exact score under alternative log-SNR truncations of the cosine schedule (the $\bar{\alpha}$ clamp that sets the lowest reachable SNR) and with individual buckets removed from the quadrature (remaining weights renormalized). Weight shares move with the truncation, but severe OOD degradation persists, and removing bucket 0 is the deletion that repairs it on both benchmarks (removing buckets 1–2 helps on Rotated MNIST only). MNIST: 6-fold mean OOD accuracy of the joint classifier; CIFAR-10-C: seed 0, N=10,000.

	weight share (%)			MNIST	CIFAR-10-C		
Score	b=0	b=1–2	b=15	OOD mean	noise	mean (15)	clean
Exact, clamp 10^{-8} (paper)	19.9	7.4	48.2	27.2	9.8	55.9	91.9
Exact, clamp 10^{-5}	26.3	9.7	31.7	25.8	9.7	55.3	91.6
Exact, clamp 10^{-3}	33.5	12.3	13.2	24.9	9.6	54.9	91.5
Exact without bucket 15	38.5	14.2	0.0	24.5	9.6	54.7	91.5
Exact without bucket 0	0.0	9.2	60.2	50.1	40.0	74.1	93.8
Exact without buckets 0 and 15	0.0	23.1	0.0	46.6	38.9	73.2	93.7
Exact without buckets 1–2	21.5	0.0	52.1	46.4	10.7	54.2	90.2
Uniform-t	6.3	12.5	6.3	57.9	36.3	74.2	94.0
Uniform-t without bucket 0	0.0	13.3	6.7	60.6	57.6	78.3	93.4

Offset sweep on Rotated MNIST. Shifting the in-domain-selected band by $k = 0, \ldots, 5$ buckets gives mean OOD accuracy 90.7, 92.3, 92.7, 92.5, 91.3, and 88.4% (75° fold: 75.0, 81.4, 84.4, 86.3, 86.9, 85.6%); k=2 maximizes the mean and matches LODO (92.7%), which is how the offset was fixed before it was applied to CIFAR-10-C. A cross-fold sensitivity check selects k to maximize the mean over the other five folds and applies it to the remaining fold (k^*=3 for 0°, 1 for 75°, 2 otherwise), giving 91.9%. The global choice is 0.8 points higher, and all cross-fold selections shift toward higher noise. This is an offset sensitivity analysis, not a claim of fully nested training-domain isolation. A fixed band 10–14 with no selection at all reaches 92.2%. The reverse transfer fails: the LODO bands selected on Rotated MNIST (8–12, 9–12, 9–13, 10–12), applied unchanged to CIFAR-10-C, give 65.1–72.0% over the 15 corruptions, below the matched ERM (75.0%); it is the offset rule, not a fixed band, that transfers.

C PER-FOLD AND PER-CORRUPTION RESULTS

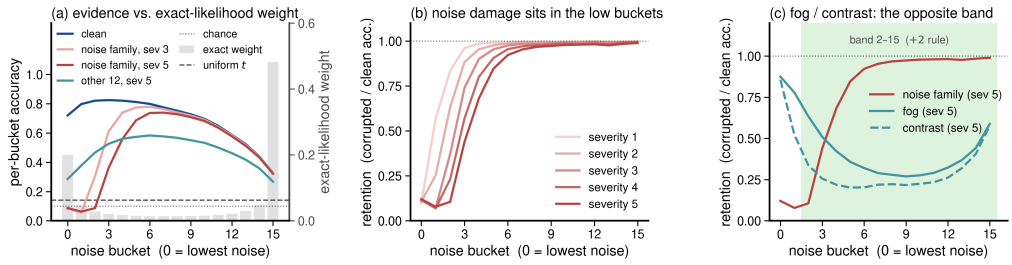


Figure 6: CIFAR-10-C per-bucket analysis. (a) Per-bucket accuracy on clean data and under corruption, with the exact-likelihood quadrature weight per bucket overlaid (right axis; dashed: uniform-t): relative to the uniform-t average, the exact likelihood over-weights the lowest-noise bucket, exactly where noise-family corruptions drive accuracy to chance. (b) Per-bucket retention (corrupted/clean accuracy) for the noise family: damage concentrates in low-noise buckets, deepens and widens with severity, and vanishes by mid-band — the CIFAR analogue of the Rotated MNIST retention structure. (c) The mirror image: fog and contrast (severity 5) damage the mid/high-noise band and leave the lowest-noise bucket nearly intact, so the $+2$ band (shaded) discards their surviving evidence. All panels: N=10,000, seed 0.

Table 6: Rotated MNIST, per-fold OOD accuracy (%) with the LODO-selected band of each fold. Binomial standard error at N=2,000 is at most 1.1 points.

Test angle	ERM	Joint exact	Joint uniform-t	Joint band	Best expert band	Posterior band	LODO band
0°	93.0	29.5	45.6	82.9	87.7	57.2	8–12
15°	98.4	40.8	88.2	97.0	90.8	87.6	9–12
30°	98.8	28.7	66.7	97.0	89.6	92.9	10–12
45°	98.0	32.7	74.6	95.8	88.8	92.0	10–12
60°	98.9	15.7	41.2	96.7	91.1	87.1	9–13
75°	95.6	16.1	31.4	86.8	89.8	62.4	9–12
Mean	97.1	27.2	57.9	92.7	89.6	79.9	

Table 7: Rotated MNIST severing experiment, per fold: band and OOD accuracy (%) under each selection rule. In-domain selection picks wide bands that reach into the fragile low buckets; every rule that succeeds on the 75° fold excludes buckets below 7.

	In-domain		In-domain +2		LODO		Oracle	
Test angle	band	acc	band	acc	band	acc	band	acc
0°	8–14	84.1	10–15	84.1	8–12	82.9	9–14	84.9
15°	5–14	96.8	7–15	97.5	9–12	97.0	7–14	97.5
30°	8–13	97.6	10–15	96.9	10–12	97.0	8–14	97.7
45°	6–15	95.8	8–15	96.5	10–12	95.8	8–15	96.5
60°	6–14	95.1	8–15	97.2	9–13	96.7	8–15	97.2
75°	5–15	75.0	7–15	84.4	9–12	86.8	9–15	86.9
Mean		90.7		92.7		92.7		93.4

The bucket-0 anomaly. Retention on the lowest-noise bucket (0.37–0.46) exceeds that of buckets 1–3 on all six folds. At the highest SNR the denoising error is dominated by pixel-level reconstruction that is nearly class-independent, so both the in-domain and OOD margins are tiny and their ratio is not a reliable measure of transfer; we draw no conclusion from its retention value. The

claims about bucket 0 in Section 3 rest on its OOD accuracy (14% on Rotated MNIST; 8.0% on the CIFAR-10-C noise family averaged over severities, where bucket 0 alone is the worst single bucket: 8.0%, 16.6%, and 35.6% for buckets 0, 1, 2) and on the ablation of Table 5, not on retention.

Table 8: CIFAR-10-C accuracy (%) per corruption, mean over five severities (N=10,000 per setting, seed 0). †: noise family.

Corruption	ERM	Exact lik.	Uniform-t	Band +2	Oracle band
gaussian_noise †	48.6	9.3	32.8	86.7	89.0
shot_noise †	60.9	10.3	45.8	87.4	89.6
impulse_noise †	54.7	9.8	30.3	59.3	78.3
defocus_blur	83.1	78.9	88.0	87.1	86.3
glass_blur	54.7	35.0	69.3	74.5	78.6
motion_blur	77.2	72.2	83.4	82.7	82.0
zoom_blur	78.5	71.7	85.9	85.7	84.9
snow	83.5	69.1	85.3	85.5	85.0
frost	80.1	63.7	86.4	87.6	86.7
fog	88.9	84.6	81.1	72.2	67.2
brightness	93.7	86.7	92.5	91.2	90.2
contrast	79.2	84.5	76.6	64.8	59.6
elastic_transform	85.8	68.9	84.5	84.5	83.8
pixelate	76.2	59.6	85.4	88.2	88.4
jpeg_compression	79.5	33.8	86.6	89.4	88.9
Mean (15)	75.0	55.9	74.2	81.8	82.6

Table 9: CIFAR-10-C accuracy (%) per severity, mean over the 15 standard corruptions (N=10,000 per setting, seed 0).

Severity	ERM	Exact lik.	Uniform-t	Band +2	Oracle band
1	88.1	66.6	87.3	89.9	89.2
2	82.7	61.1	80.8	87.7	87.2
3	76.7	57.6	74.3	84.5	84.3
4	69.5	51.5	68.4	77.3	79.3
5	57.9	42.5	60.4	69.5	72.9

D PER-TIMESTEP QUADRATURE: VALIDATING THE 16-BUCKET EXACT SCORE

The main-text ITD-weighted score is a bucketwise approximation using bucket-mean ϵ-errors weighted by log-SNR width. Its class differences approximate the learned-denoiser counterpart of Eq. 1 over the evaluated noise range. To check that the ordering exact $<$ uniform-t $<$ band is a property of the log-SNR weighting and not of that discretization, we re-evaluate with one bucket per stratified timestep (M buckets, one noise draw each) and form the exact score by an M-point trapezoid rule in α, the uniform-t score as the plain mean over the M timesteps, and the band score as the mean over the timesteps that fall in the band (buckets 2–15 on CIFAR-10-C; the fold's LODO band on Rotated MNIST); the 16-bucket score of the main text is recovered from the same tensor by re-grouping the timesteps. On CIFAR-10-C we run M=64 and M=128 (seed 0, a fixed 2,048-image subset of every setting) and on Rotated MNIST M=128 (joint model, all six folds, held-out test images and in-domain validation images).

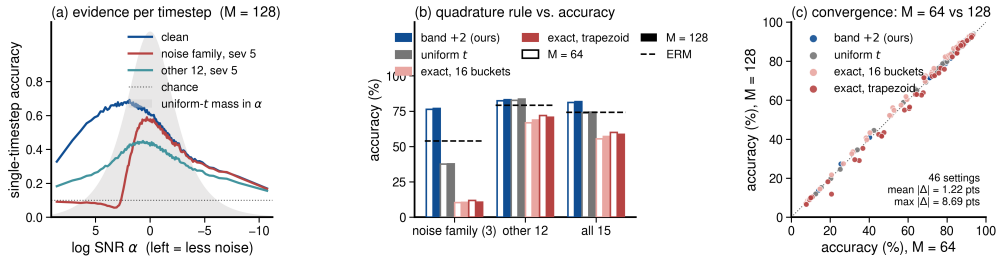


Figure 7: Per-timestep validation on CIFAR-10-C (N=2,048, seed 0). (a) Accuracy of the single-timestep classifier $\arg\min_y \|\epsilon - \hat{\epsilon}_y(\boldsymbol{x}, t_m)\|^2$ at each of M=128 timesteps against log-SNR α, for clean data, the noise family at severity 5, and the other 12 corruptions at severity 5; the gray density is where a uniform-in-t sampler places its mass on the α axis, whereas the exact likelihood is flat in α. On the noise family every timestep with $\alpha \gtrsim 3$ is at or below chance. (b) Accuracy of the four estimators by corruption family for M=64 (hollow) and M=128 (filled); dashed: ERM. (c) Per-setting accuracy at M=128 against M=64 for the same estimators (46 settings).

Table 10: CIFAR-10-C accuracy (%) of the four estimators computed from per-timestep errors, M=64 and M=128 timesteps (N=2,048 per setting, seed 0). "noise", "other 12", "all 15": means over the corresponding standard corruptions at severities 1, 3, and 5.

	M=64				M=128			
Score	noise	other 12	all 15	clean	noise	other 12	all 15	clean
Exact, trapezoid in α, M points	11.9	72.0	60.0	93.3	10.7	70.7	58.7	93.2
Exact, 16-bucket rule (paper)	10.4	66.8	55.5	91.4	10.4	68.8	57.1	92.9
Uniform-t average	37.6	82.7	73.7	93.7	37.7	83.6	74.4	94.0
Band +2	76.4	82.4	81.2	91.7	76.9	83.1	81.8	92.8

Table 11: Rotated MNIST accuracy (%) of the joint classifier from M=128 per-timestep errors, per fold, out of distribution (2,000 held-out test images) and in domain (400 validation images per training domain).

	OOD (held-out rotation)				in-domain (validation)			
Fold	exact, trap.	exact, 16 b.	uniform-t	band (LODO)	exact, trap.	exact, 16 b.	uniform-t	band (LODO)
0°	33.9	29.5	45.6	82.9	96.3	95.7	97.0	97.4
15°	54.2	40.8	88.2	97.0	96.1	95.8	97.2	97.7
30°	37.4	28.7	66.7	97.0	95.3	95.1	97.1	96.8
45°	45.0	32.7	74.6	95.8	96.0	95.5	97.3	96.8
60°	20.5	15.7	41.2	96.7	95.3	94.7	96.6	97.5
75°	19.3	16.1	31.4	86.8	96.1	95.6	97.2	97.2
Mean	35.0	27.2	57.9	92.7	95.8	95.4	97.0	97.2

The ordering survives the discretization. With the 128-point trapezoid rule the exact score reaches 10.7% on the CIFAR-10-C noise family, 58.7% over the 15 corruptions, and 35.0% mean OOD on Rotated MNIST, against 37.7%, 74.4%, and 57.9% for the uniform-t average and 76.9%, 81.8%, and 92.7% for the band (Tables 10 and 11). The full ordering holds on all six Rotated MNIST folds. On CIFAR-10-C at M=128 the exact score is below the uniform-t average on 41 of the 45 corruption settings and below the band on 37; the four settings on which it beats the uniform-t average are fog and contrast at severities 3 and 5, the two corruptions whose damage the main text locates in the mid-band rather than at low noise (Section 5.3). (The band is below the uniform-t average on 20 of the 45 settings, all outside the noise family, as in the main text: the fix

targets the noise family.) The 16-bucket rectangle rule and the trapezoid rule differ by 1.9 points per setting on average (largest difference 9.0) on CIFAR-10-C at $M{=}128$, while on Rotated MNIST the rectangle rule understates the exact score by 7.8 points OOD (27.2% vs. 35.0%), so the finer quadrature improves this score's accuracy without changing its ordering relative to uniform-t and band scoring. This is evidence for a weighting effect, not for decreasing robustness with increasing numerical accuracy. The trapezoid rule involves no truncation choice beyond the schedule itself and places 19.5% ($M{=}64$) and 21.5% ($M{=}128$) of its weight on the range of bucket 0, bracketing the rectangle rule's 19.9%.

Convergence in M. Doubling the number of timesteps from 64 to 128 moves the per-setting accuracy by 1.7 points on average for the two exact scores and by 0.7–0.8 points for the uniform-t and band scores (largest single change: the trapezoid exact score on shot noise at severity 1, 8.7 points; Figure 7c), and the family means by 2.0 points at most (Table 10), with no change in the ordering of the four estimators within any corruption family. The exact scores converge more slowly because their weight is concentrated on a few timesteps at the two ends of the schedule, where the log-SNR spacing is largest, so a single noise draw at those timesteps carries more of the score.

In domain the exact score is fine. On the Rotated MNIST validation images of the training domains the exact trapezoid score reaches 95.8% (uniform-t 97.0%, band 97.2%; Table 11), and on clean CIFAR-10 93.2% (94.0%, 92.8%). The OOD gap between the exact score and the band (57.7 points on Rotated MNIST, 66.2 on the CIFAR-10-C noise family) is therefore created by the shift, not by the exact weighting being a poor classifier per se.

E A NATURAL DATASET SHIFT: CIFAR-10.1

The two shifts in the main text are synthetic transformations. As a check on a natural shift we score the same three CIFAR-10 models, unchanged, on CIFAR-10.1 (v6; 2,000 new test images collected with the CIFAR-10 protocol; Recht et al., 2018), with the same $M{=}64$ budget and the band 2–15 locked from the main text (no selection of any kind on CIFAR-10.1).

Table 12: CIFAR-10.1 accuracy (%), three model seeds, with clean CIFAR-10 accuracy for reference (seed 0 on the full 10k test split, seeds 1–2 on their 2,048-image subsets). Nothing is selected on CIFAR-10.1.

	seed 0	seed 1	seed 2	mean $\pm$ sd	clean (mean)
ERM ResNet-18	87.6	88.9	87.8	88.1 $\pm$ 0.7	95.3
Diffusion classifier, exact likelihood (ITD quadrature)	84.9	83.7	84.0	84.2 $\pm$ 0.6	91.5
Diffusion classifier, uniform-t average	88.0	86.8	85.5	86.8 $\pm$ 1.3	94.1
Band 2–15 (clean+2, locked)	84.2	83.7	82.3	83.4 $\pm$ 1.0	92.2
ERM ResNet-18 + AugMix (seed 0)	89.6	–	–	–	95.4

A complementary failure profile explains why the band does not help. CIFAR-10.1 is a mild shift that largely preserves low-noise evidence and degrades high-noise evidence somewhat more: per-bucket retention (CIFAR-10.1/clean accuracy, 3-seed mean) is 0.87–0.89 on buckets 0–2 and 0.77–0.80 on buckets 13–15, so the low-noise evidence that noise-family corruptions destroy is here the best-preserved part of the schedule, and the mild damage sits at high noise (the fog/contrast direction of Figure 6c). Accordingly the locked band 2–15 is 3.4 points *below* the uniform-t average (83.4% vs. 86.8%), and the band that would be selected on CIFAR-10.1 itself keeps the low buckets (0–13, 0–15, 0–15 across seeds; test-selected, reported only as a reference). The uniform-t diffusion classifier stays within 1.3 points of the matched ERM (86.8% vs. 88.1%). The exact score is again below the uniform-t average on every seed (84.2% vs. 86.8%) and slightly above the band (83.4%), with a far smaller gap than under the severe CIFAR-10-C noise corruptions. CIFAR-10.1 thus exposes a boundary condition of the mechanism: when low-noise evidence remains intact, removing it costs accuracy, consistent with its measured per-bucket profile.