DIFFUSION-INFORMED SINGLE-PASS CLASSIFIER

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifiers built from diffusion models (i.e. diffusion classifiers) are reported to be robust and well aligned with human perception. However, a single prediction requires about 10^5 forward passes, which is far too costly for deployment. We ask where this alignment (shape bias, out-of-distribution accuracy, and error consistency with human mistakes) comes from, and whether an ordinary single-pass discriminative classifier can have it too. We carefully decompose the human alignment gap between diffusion classifiers and discriminative classifiers that is reported in the literature into the axes along which their designs differ: training data, resolution, aggregation of evidence across noise levels and training objective (generative vs discriminative). We isolate "aggregation" as an important contributor (more than training data or the generative objective alone) to this alignment. We measure how much each noise level contributes to a diffusion classifier's prediction through reductions in KL divergence to its final prediction. The normalized average contributions define a noise-sampling distribution, w_{gen}, and their cumulative sum, ρ, provides a label-smoothing schedule. We use these two curves to set the noise levels and label confidence when training an ordinary single-pass classifier (e.g., a ResNet-50), transferring aggregation with no change to architecture or inference cost; and, unlike naive noise augmentation, at no cost in accuracy. Our recipe, DISC (Diffusion-informed single-pass classifier), matches the diffusion classifier's out-of-distribution accuracy, exceeds its error consistency, narrows its shape-bias gap, and improves robustness to corruptions and distribution shifts without losing clean accuracy. The w_{gen} curve also serves as an adaptive guidance schedule, improving diffusion sampling quality.

1 INTRODUCTION

A classifier trained with cross-entropy does not see images the way people do (Geirhos et al., 2021). It is rewarded for any feature that separates the training classes, including features no person would use (Geirhos et al., 2020a). It breaks under common corruptions (Hendrycks & Dietterich, 2019), it is fooled by small perturbations (Szegedy et al., 2014; Madry et al., 2018), and its mistakes do not look like human mistakes (Geirhos et al., 2020b).

Diffusion classifiers (Li et al., 2023; Clark & Jaini, 2023) are reportedly much more human aligned. Jaini et al. (2024) report a shape bias close to the human level and the highest error consistency with humans. However, a single prediction costs a forward pass for every class at every noise level, on the order of 10^5 passes per image on ImageNet-1k, which is far too costly for deployment. This raises two questions. Where does the alignment come from, and can an ordinary classifier have it?

Where the alignment comes from. It is tempting to credit the generative objective, but published comparisons confound (1) training data, (2) input resolution, (3) aggregation of evidence across noise levels, and (4) training objective, all at once, so no single axis can be credited on its own for the reported gaps in human alignment. Disentangling them is one of our contributions: we construct a set of models that probe these factors through increasingly controlled comparisons and report the effect of each axis separately in Table 1. We add the rows needed to isolate each axis (R6–R15). Reading one axis at a time (Table 1b), we see a simple picture. At matched data and resolution, switching from the discriminative to the diffusion family raises shape bias in every case but does not consistently raise out-of-distribution (OOD) accuracy or error consistency; those two are instead mostly explained by training data scale, which leaves shape bias untouched. In the comparisons so far, axes (3) aggregation and (4) training objective are still entangled. The Joint Energy Models (JEMs) of Ortega et al. (2026) isolate the objective axis alone (data and resolution fixed, aggregation

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.

Table 1: **The decomposition.** **(a)** Shape bias (SB), out-of-distribution accuracy (OOD) and error consistency (EC) of every model on the model-vs-human benchmark (Geirhos et al., 2021), in percent. (i) Published comparisons, where data, resolution, aggregation and objective are confounded. (ii) Rows we add so that each changes exactly one axis. (iii) Joint energy models (JEMs) of Ortega et al. (2026), read from their figures. (iv) DISC. R7 and R16 are means over three training seeds. **(b)** Each row changes one axis with everything else held fixed, read from (a). **(c)** How the cue-conflict trials are answered (shape, texture, or neither) as test-time Gaussian blur of strength σ increases. The dashed line is the shape fraction of the reference diffusion classifier; DISC reaches EDM's under mild blur, the vanilla ResNet-50 never does.

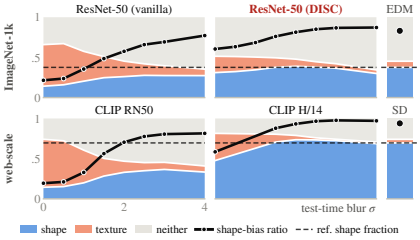
(a) Measured cells

	model	data	loss	input	clean	SB	OOD	EC
(i) Prior art, reproduced: axes (1)–(4) confounded								
R1	ResNet-50	IN-1k	disc.	224	76.1	21.5	55.9	20.8
R2	ResNet-50 + naive diff. noise	IN-1k	disc.	224	67.2	51.2	56.1	24.7
R3	CLIP ViT-B/32	web	disc.	224	—	57.3	70.8	28.1
R4	Stable Diffusion all 250 levels	web	gen.	latent	—	93.9	67.5	25.4
R5	Imagen	web	gen.	64	—	99.0	70.6	30.9
(ii) Settings that enable one-axis-at-a-time controlled experiments								
R6	CLIP ResNet-50	web	disc.	224	—	19.5	62.6	23.7
R7	ResNet-50	IN-1k	disc.	64	73.5	41.6	50.1	18.0
R8	EDM all 250 levels	IN-1k	gen.	64	—	83.1	56.2	15.6
R9	EDM, best single level	IN-1k	gen.	64	—	67.9	34.7	10.0
R10	DiT-XL/2 250 levels	IN-1k	gen.	latent	—	85.9	60.3	20.7
R11	Stable Diffusion, best single level	web	gen.	latent	—	80.6	34.1	9.6
(iii) Ortega et al. (2026): axes (1),(2) fixed, (3) removed, (4) isolated								
R12	JEM, α=0 (WRN-22-8)	IN-1k	disc.	latent	—	51.0	43.8	10.1
R13	JEM, α=0.5	IN-1k	hybrid	latent	—	63.8	41.8	17.5
R14	JEM, α=0.8	IN-1k	hybrid	latent	—	73.0	29.1	10.7
R15	JEM, α=1	IN-1k	gen.	latent	—	63.2	8.4	0.9
(iv) DISC (ours)								
R16	**ResNet-50 (DISC)**	IN-1k	disc.	64	**73.6**	**60.5**	**56.7**	**23.0**

(b) One axis at a time

axis changed (rows of (a))	ΔSB	ΔOOD	ΔEC
(1) data, R1 → R6	−2.0	**+6.7**	**+2.9**
(2) resolution, R1 → R7	**+20.1**	−5.8	−2.8
(3+4) diffusion family, R7 → R8	**+41.5**	**+6.1**	−2.4
(3+4) diffusion family, R3 → R4	**+36.6**	−3.3	−2.7
(4) objective, R12 → R13	**+12.8**	−2.0	**+7.4**
(3) aggregation, R9 → R8	**+15.2**	**+21.5**	**+5.6**
(3) aggregation, R11 → R4	**+13.3**	**+33.4**	**+15.8**
DISC, R7 → R16	**+18.9**	**+6.6**	**+5.0**

(c) Cue-conflict decisions



absent by construction, Table 1a(iii)): generative mixing helps error consistency and shape bias, but far less than diffusion classifiers achieve, pointing to aggregation as the larger remaining factor. We tested the effect of aggregation directly by keeping the diffusion model fixed and scoring it at a single noise level instead of aggregating over 250 levels (R9→R8, R11→R4). To choose which single noise level to use, all noise levels are evaluated and the one which has best OOD accuracy is selected to give most optimistic picture; nevertheless aggregation seems to be the key factor (more than training data or the generative objective alone) behind this alignment. As we show in Section 3.2, transferring this aggregation is enough to improve all three metrics while holding clean accuracy fixed, unlike naive diffusion noise (Jaini et al., 2024; Gavrikov, 2024) (R2), which costs ≈ 9 points of accuracy.

A separate concern comes from concurrent work. Preferring one cue under conflict does not mean the other cue is unread (Burgert et al., 2025), and the shape-bias ratio ignores trials answered with neither cue, so it can rise by abstention alone (Wolff et al., 2026). Table 1c therefore shows the raw decision fractions under the test-time blur sweep of Wolff et al. (2026). The vanilla ResNet-50 and CLIP ResNet-50 never reach the shape fraction of their reference diffusion classifier (dashed line). Our recipe (DISC) reaches EDM's, and at $\sigma = 2.5$ its ratio passes EDM's.

How we achieve this. We first inspected how a diffusion classifier combines scores across noise levels. Starting from the noisiest level, we add scores toward the clean end and measure each step's reduction in KL divergence to the final prediction. Normalizing these contributions per image and averaging them over the dataset gives w_{gen} (Figure 1a). Its cumulative sum, ρ, records the normalized KL reduction accumulated from the noisiest level down to the current level (Figure 1b). We read both curves from a pretrained diffusion model once per dataset.

We use the two curves to couple input noise and target smoothing. Each training noise level is sampled from w_{gen}, emphasizing levels with larger measured contributions. At that level, ρ interpolates between the one-hot ground-truth label and the uniform vector (Figure 1b–d). Targets are softer where the cumulative attribution is small and sharper where it is large. This transfers the measured noise-level profile into ordinary cross-entropy training, without changing the classifier architecture or inference cost. Once the curves have been estimated, no diffusion model is needed during classifier training or inference. We call the resulting model DISC (Diffusion-Informed Single-pass Classifier), and the recipe DISC training.

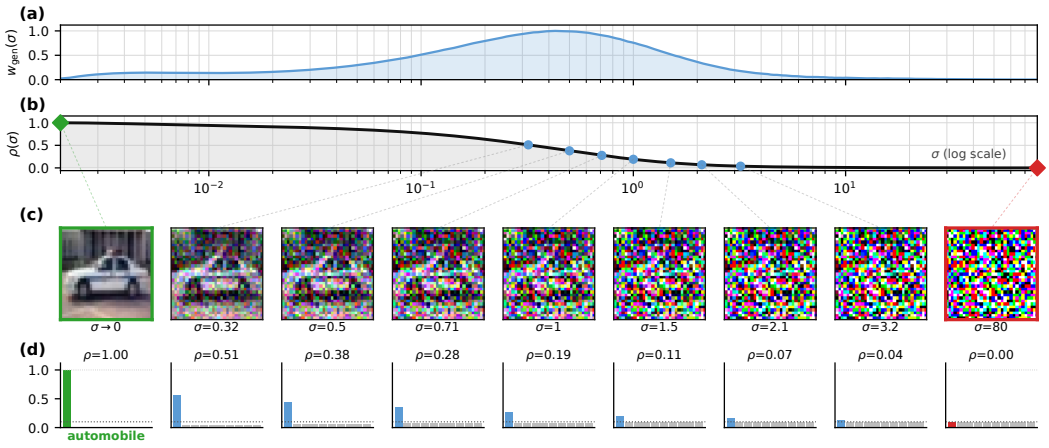


Figure 1: **(a)** $w_{\text{gen}}(\sigma)$: the diffusion classifier's normalized attribution of its decision to each noise level, averaged over the dataset. DISC training draws each noise level with this probability. **(b)** The curve $\rho(\sigma)$ is the cumulative normalized attribution from the noisiest level (right) down to σ. It is anchored at 1 at the lowest noise level and 0 at the highest, and controls the label interpolation in (d). **(c)** Example image noised at the selected σ values. **(d)** The smoothed label $\boldsymbol{q}_\sigma = \rho(\sigma)\,\text{onehot}(c) + (1-\rho(\sigma))\,\boldsymbol{u}$ for each selected noise level. Both curves are measured on ImageNet-64 (Section 3.1). The example image is from CIFAR-10, so that all $K = 10$ bars of the target are visible.

The curves also improve image generation, in two ways. First, DISC can be used unchanged as the guidance classifier of ADM ImageNet-64 sampler (Dhariwal & Nichol, 2021), where it lowers FID from 2.07 to 1.76. Second, the w_{gen} curve itself serves as a guidance schedule: rescaling the classifier-free guidance scale by w_{gen} improves DiT-XL/2 (Peebles & Xie, 2023) from its published FID of 2.27 to 2.11 (Section 4.3).

Contributions. **(1)** A noise-level attribution measured once from a diffusion classifier: its normalized average contributions, w_{gen}, and their cumulative sum, ρ, jointly determine noise sampling and label smoothing. **(2)** DISC, a discriminative classifier trained with a recipe driven by the curve, at no extra inference cost. It keeps clean accuracy, closes the gap in OOD accuracy and error consistency to a diffusion classifier trained on the same dataset (ImageNet-1k), narrows the gap in shape bias, and improves robustness to corruptions and distribution shifts. **(3)** The w_{gen} curve, used as a guidance schedule, improves over constant guidance for the ADM-64 and DiT-XL/2 samplers under their authors' protocols, and DISC is a better guide than the noise-conditional classifier built specifically for guidance. **(4)** A controlled decomposition of human alignment on ImageNet-1k, with the first same-data diffusion-classifier rows on the model-vs-human benchmark.

2 Related Work

Diffusion classifiers, alignment, and robustness. A class-conditional diffusion model can be turned into a classifier by scoring each class by how well the model denoises under it (Li et al., 2023; Clark & Jaini, 2023). Such classifiers are reported to be robust (Chen et al., 2024b;a; Li et al., 2023), to avoid shortcut solutions (Li et al., 2025), and to reach near-human shape bias and the highest error consistency with humans on the model-vs-human benchmark (Jaini et al., 2024). They are, however, far too costly for deployment, which raises the question of where these properties come from and whether a single-pass model can have them. Wolff et al. (2026) attribute much of the human alignment to the fact that diffusion classifiers, by design, operate on down-scaled inputs, which acts as a low-pass filter and trades clean accuracy for alignment; this is the "resolution" control in Table 1. Ortega et al. (2026) test the "training objective" control in a setting that removes aggregation over noise levels by construction. Beyond diagnosing the human alignment and robustness, no prior work attempts to make them usable in a single pass. Jaini et al. (2024) briefly report, in a single experiment, that training a discriminative classifier with diffusion noise augmentation transfers part of the alignment, but offer no further analysis, and it costs clean accuracy (Table 1, R2). This work aims to make these gains practical at no clean accuracy or inference cost.

Noise augmentation, label smoothing, and the benchmarks. Gaussian noise training improves corruption robustness (Rusak et al., 2020) and raises shape bias (Hermann et al., 2020), but it costs clean accuracy, which is usually mitigated by mixing clean images into the batch (Rusak et al., 2020). Beyond noise, the augmentation literature offers various techniques: mixing the inputs and labels of image pairs (Zhang et al., 2018; Yun et al., 2019), enforcing consistency across augmented views (Hendrycks et al., 2020), and mixing with fractals (Hendrycks et al., 2022); IPMix (Huang et al., 2023) is a recent tuned combination of these, which does not include noise augmentation. It has also been proposed to tie label smoothing (Szegedy et al., 2016; Müller et al., 2019) to the strength of the augmentation (Liu et al., 2023; Qin et al., 2023). Closest to our method, Heinonen et al. (2025) couple Gaussian noise and blur with label smoothing, softening the label as the input degrades. We derive both schedules from a diffusion classifier's measured noise-level attribution: the normalized contributions determine noise sampling, and their cumulative sum determines label smoothing. Their schedules, in contrast, are handcrafted functions of the signal-to-noise ratio, which only depends on input dimension and is agnostic to data distribution. Their setting is a perfectly isolated testbed for our curves. We show that dropping our measured ρ into their setting improves their results (Table 3). Throughout the paper, we evaluate and compare against baselines on the model-vs-human benchmark (Geirhos et al., 2021), ImageNet-C (Hendrycks & Dietterich, 2019) and the distribution shifts of ImageNet-testbed (Taori et al., 2020).

3 METHOD

DISC uses two curves measured from a diffusion classifier: the normalized average attribution to noise levels, w_{gen}, and its cumulative sum, ρ. Section 3.1 defines both through reductions in KL divergence to the full-score prediction. In training (Section 3.2), noise levels are sampled from w_{gen} and the target interpolation coefficient is ρ; in sampling (Section 3.3), the guidance scale is rescaled by w_{gen}.

3.1 READING THE CURVES OFF A DIFFUSION CLASSIFIER

Diffusion classifier formulation. A class-conditional diffusion model is trained to predict the noise in a noised image. Formulated as: $\boldsymbol{x}_t = \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, for the image noised to level t on a grid $\sigma_0 < \cdots < \sigma_T$, with $t = 0$ the cleanest level and $t = T$ the noisiest, and $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \sigma_t, c)$ for the model's noise prediction under class c. The diffusion classifier (Li et al., 2023) turns the model into a classifier by scoring each class at each noise level by how well the model denoises under that class, where the logit of class c at the noise level t is defined as the negative mean square error between the added noise and the predicted noise conditioned with t and c :

$$\ell_t(c) \;:=\; -\,\mathrm{MSE}\left[\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \sigma_t, c)\right]. \tag{1}$$

It then forms the final logit as the sum of the per-level logits and predicts the class with the largest sum,

$$p_{\text{final}}(c \mid \boldsymbol{x}_0) \;:=\; \mathrm{softmax}_c\Big(\sum_{t=0}^{T} \ell_t(c)\Big), \qquad \hat{c} = \arg\max_c\, p_{\text{final}}(c \mid \boldsymbol{x}_0). \tag{2}$$

No single noise level classifies well on its own (R9 against R8 in Table 1).

Attributing the decision to noise levels. The final logit is a sum, so we can watch the decision form as levels are added. We add levels from the noisiest end toward the clean image and track the resulting prediction. The running prediction once levels $t, \ldots, T$ have joined is

$$q_t(c \mid \boldsymbol{x}) \;:=\; \mathrm{softmax}_c\Big(\sum_{s=t}^{T} \ell_s(c)\Big). \tag{3}$$

At $t = T$ only the noisiest level has joined and q_T is close to uniform; at $t = 0$ every level has joined and $q_0 = p_{\text{final}}$. The distribution q_t is the prediction obtained from the accumulated scores at levels $t, \ldots, T$. The contribution of level t is how far it moves the running prediction toward the final one, measured as the drop in KL divergence to q_0 when level t joins,

$$w_{\boldsymbol{x}}(t) \;:=\; \mathrm{KL}\big(q_0 \,\|\, q_{t+1}\big) - \mathrm{KL}\big(q_0 \,\|\, q_t\big). \tag{4}$$

It is positive when a level moves the prediction toward the decision and negative when it moves it away, and it tracks the decision the classifier actually makes, whether or not that decision is correct.

Summing the contributions of levels t and above telescopes,

$$W_{\boldsymbol{x}}(t) \;:=\; \sum_{s=t}^{T-1} w_{\boldsymbol{x}}(s) \;=\; \mathrm{KL}\big(q_0 \,\|\, q_T\big) - \mathrm{KL}\big(q_0 \,\|\, q_t\big), \tag{5}$$

so $W_{\boldsymbol{x}}(0) = \mathrm{KL}(q_0 \,\|\, q_T)$ is the total KL reduction from the highest-noise reference to the full-score prediction. The ratio $W_{\boldsymbol{x}}(t)/W_{\boldsymbol{x}}(0)$ is the normalized cumulative reduction after levels t and above have been included. Averaging over a class-balanced dataset gives the curves,

$$w_{\mathrm{gen}}(t) \;:=\; \mathbb{E}_{(\boldsymbol{x},c)}\left[\frac{w_{\boldsymbol{x}}(t)}{W_{\boldsymbol{x}}(0)}\right], \qquad \rho(t) \;:=\; \mathbb{E}_{(\boldsymbol{x},c)}\left[\frac{W_{\boldsymbol{x}}(t)}{W_{\boldsymbol{x}}(0)}\right] \;=\; \sum_{s=t}^{T-1} w_{\mathrm{gen}}(s), \tag{6}$$

so ρ is the cumulative sum of w_{gen} from the noisiest level down, with $\rho(0) = 1$ and $\rho(T) = 0$. These curves describe the selected diffusion model under the scoring convention used to compute q_t.

Using w_{gen} as a sampling distribution requires non-negative entries; together with their unit sum, this also places the cumulative label coefficient ρ in $[0, 1]$. Non-negativity is not guaranteed by the per-image construction: a level can move the running prediction away from the full-score prediction, giving a negative contribution. On 59% of the estimation images, the cumulative quantity $W_{\boldsymbol{x}}$ is negative at some level. Nevertheless, every entry of the measured dataset-average w_{gen} is non-negative without clipping. The resulting ρ stays in $[0, 1]$ and monotonically decreases with σ (Figure 4, Appendix A.1).

We estimate both curves on a class-balanced subset of the ImageNet-1k training split with $12{,}000$ images. The diffusion model is queried once, for this estimate, and never again: the training recipe has no diffusion model in the loop, and all it takes as input is the table of w_{gen} and ρ. The diffusion models we read the curves from (Karras et al., 2022; Peebles & Xie, 2023) are trained on the same data as our classifier (ImageNet-1k), so no gain can be credited to extra data.

3.2 Discriminative training with the curves

We train a standard classifier f_θ with ordinary cross-entropy and change two things: how noisy the input is, and how confident the target is. At each step we draw $(\boldsymbol{x}_0, c)$ from the training set and a noise level $\sigma \sim w_{\mathrm{gen}}$, and then estimate model parameters θ by minimizing the objective:

$$\begin{aligned} &\min_\theta \; \mathbb{E}_{(\boldsymbol{x}_0,c)\sim\mathcal{D},\, \sigma\sim w_{\mathrm{gen}},\, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\Big[\mathrm{CE}\big(\boldsymbol{q}_\sigma,\, \mathrm{softmax} f_\theta(\boldsymbol{x}_\sigma)\big)\Big], \quad \text{where} \\ &\boldsymbol{x}_\sigma = \frac{(\boldsymbol{x}_0 - \mu_{\mathrm{data}}) + \sigma\boldsymbol{\epsilon}}{\sqrt{\sigma^2 + \sigma_{\mathrm{data}}^2}}, \qquad \boldsymbol{q}_\sigma = \rho(\sigma)\,\mathrm{onehot}(c) + \big(1 - \rho(\sigma)\big)\boldsymbol{u}. \end{aligned} \tag{7}$$

As shown in Figure 1d, the label map couples target sharpness to cumulative attribution: where the accumulated scores have made little progress toward the full-score prediction, targets are softer; where the cumulative attribution is larger, targets are sharper. This gives targets close to one-hot at low σ and close to uniform at high σ. The sampler and label map are evaluated separately and jointly in Section 4.1. Figure 2a illustrates the effect on toy data. A vanilla CE classifier is fully confident everywhere, also far from the data (Hein et al., 2019); its input gradient looks like noise (Szegedy et al., 2014; Tsipras et al., 2019), and its decision boundaries are brittle (Goodfellow et al., 2015; Madry et al., 2018). With DISC training, confidence is high near the data and falls to uniform away from it. Moreover, the confidence $p(c \mid x)$ seems to be monotonically increasing towards the data manifold, which suggests that the gradient in input space in the direction that maximizes confidence, i.e., $\nabla_{\boldsymbol{x}_t} \log p(c \mid \boldsymbol{x}_t)$, is perceptually aligned and can guide the diffusion sampling process (Kaur et al., 2019; Srinivas et al., 2023; Kawar et al., 2023; Ma et al., 2024). We test these properties on ImageNet in Sections 4.1 and 4.3.

Figure 2b ablates the w_{gen} sampler and the ρ label schedule against vanilla diffusion noise augmentation, which uses a uniform sampler and hard labels (Jaini et al., 2024). Noise augmentation aims to cover the density around the training samples instead of the samples alone. The w_{gen} sampler covers this density better, and with ρ-coupled labels it requests less confidence near decision boundaries, which avoids a sharp clash of labels there. On ImageNet, only the w_{gen} + ρ combination preserves clean accuracy while gaining human alignment and adversarial robustness (Section 4.1).

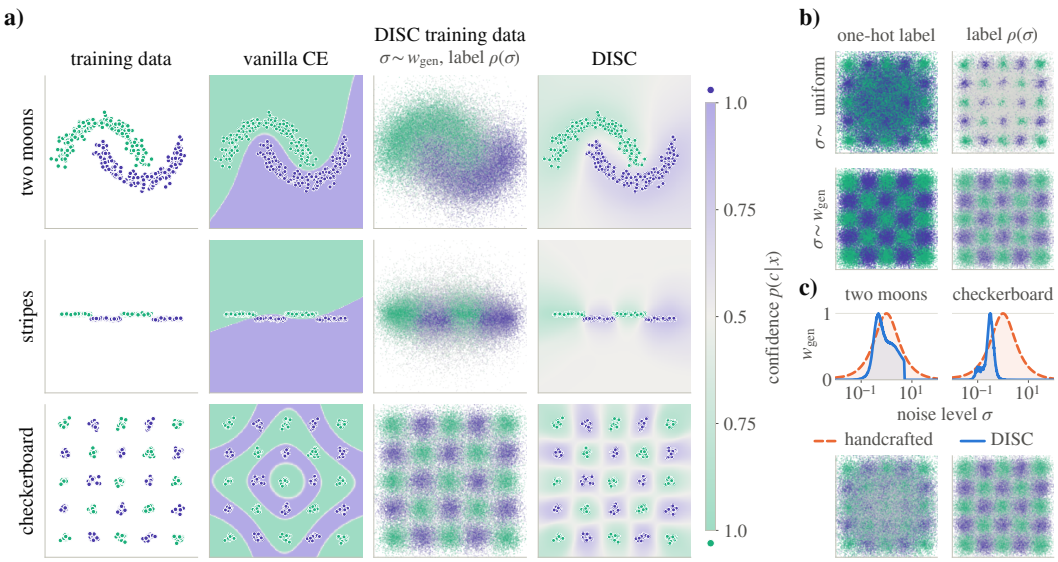


Figure 2: **Illustration of DISC training on toy examples. (a)** Three toy settings where we trained a diffusion model and obtained w_{gen} and ρ from it. Columns: the training data; the confidence map of a classifier trained with standard cross-entropy (vanilla CE), which sees the same points every epoch; the DISC training data, i.e. the inputs noised with $\sigma \sim w_{\text{gen}}$ over 150 epochs, colored by their label confidence $\rho(\sigma)$; and the confidence map of DISC trained on them, in one clean forward pass. Color is the model's confidence $p(c \mid x)$; dots are the training points. **(b)** The checkerboard training data under each noise-level sampler and label choice. The measured pair (w_{gen}, ρ) concentrates high-confidence targets around the class clusters while softening targets in the intervening regions. **(c)** The noise-level sampler w_{gen} from the measured curve (DISC) and from the handcrafted schedule of Heinonen et al. (2025). They nearly agree on two moons and differ significantly on the checkerboard, whose cell size the handcrafted schedule cannot know. Below: the checkerboard training data each one induces.

Figure 2c shows how our measured curve adapts to the dataset, unlike the handcrafted signal-to-noise schedule of Heinonen et al. (2025). Their schedule depends only on the input dimension and the noise level, so it induces the same $w_{\text{gen}}(\sigma)$ for the two moons and for the checkerboard, and it fails to cover the data distribution of the checkerboard.

3.3 Diffusion guidance with the $w_{\text{gen}}(\sigma)$ curve

Diffusion guidance steers a sampler toward class c by adding $s \cdot \nabla_{\boldsymbol{x}_t} \log p(c \mid \boldsymbol{x}_t)$ to the model's score at every noise level, with a constant *guidance scale* s. Classifier guidance (Dhariwal & Nichol, 2021) obtains this gradient from a separate noise-conditional classifier trained on noised images. Classifier-free guidance (Ho & Salimans, 2021) obtains it as the weighted difference between a conditional $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, c)$ and an unconditional $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \varnothing)$ denoiser. We contribute here in two ways:

1) Drop-in replacement for ADM classifier guidance. In the setting of ADM (Dhariwal & Nichol, 2021), in place of their noise-level conditioned classifier that is tuned specifically for guidance, we plug in DISC, a single-forward-pass ResNet-50 with no noise-level condition, and use its gradients $\nabla_{\boldsymbol{x}_t} \log p(c \mid \boldsymbol{x}_t)$ for classifier-guidance. DISC takes the sampler state $\boldsymbol{x}_t$ directly, with no time input. This tests at ImageNet scale what the toy experiments of Section 3.2 suggest: the gradient of DISC visually points towards the data manifold.

2) Rescaling the guidance scale by $w_{\text{gen}}(\sigma)$. A constant scale applies the same guidance at every noise level. Our measured w_{gen} is concentrated at intermediate noise levels, identifying where score accumulation makes the largest average normalized progress toward the full-score prediction. This profile is consistent with the empirical finding of Kynkäänniemi et al. (2024) that guidance is harmful at high noise levels and largely unnecessary at low ones. We therefore use w_{gen} to concentrate guidance at levels with larger measured contributions, where $w_{\text{gen}}(\sigma)$ is large: the constant scale s

Table 2: **The sampler $\times$ label** 2×2 at 64px, flagship recipe, mean$\pm$std over three seeds; the w_{gen}+ρ-soft cell is DISC (R16) and the no-noise control is R7 of Table 1. mCA: mean corruption accuracy; AA: AutoAttack at $\varepsilon = 1/255$; clean top-1 is the training-script validation number.

sampler	label	clean ↑	mCA ↑	AA ↑	shape bias	OOD acc.	error cons.
w_{gen}	ρ-soft	$\mathbf{73.64}_{\pm.06}$	$\mathbf{50.89}_{\pm.14}$	$18.2_{\pm1.4}$	$60.5_{\pm0.1}$	$56.7_{\pm0.3}$	$23.0_{\pm0.1}$
w_{gen}	hard	$70.98_{\pm.13}$	$49.51_{\pm.10}$	$25.3_{\pm0.7}$	$65.3_{\pm0.6}$	$\mathbf{57.3}_{\pm0.3}$	$\mathbf{23.6}_{\pm0.3}$
uniform	hard	$62.40_{\pm.07}$	$43.45_{\pm.17}$	$\mathbf{27.7}_{\pm0.8}$	$\mathbf{67.1}_{\pm2.7}$	$54.0_{\pm0.3}$	$23.3_{\pm0.7}$
uniform	ρ-soft	$63.98_{\pm.03}$	$43.14_{\pm.21}$	$23.3_{\pm1.0}$	$59.8_{\pm2.8}$	$52.6_{\pm0.5}$	$22.5_{\pm1.1}$
no noise (control)		$73.49_{\pm.08}$	$40.37_{\pm.20}$	$0.0_{\pm0.0}$	$41.6_{\pm3.2}$	$50.1_{\pm0.6}$	$18.0_{\pm0.4}$

is replaced by $s \cdot w_{\text{gen}}(\sigma_t)$, with the density $w_{\text{gen}}(\sigma)$ of Equation 6 normalized to unit mean over the steps the sampler visits, so the average guidance strength along the trajectory is unchanged and only its distribution over noise levels changes. In classifier-free guidance the guidance weight $w - 1$ is replaced by $(w - 1)\, w_{\text{gen}}(\sigma_t)$ in the same way.

4 Experiments

Setting. We keep the training data controlled: every model we train uses ImageNet-1k only, and so do the diffusion models we read the curves from, so no gain can come from extra data. We obtain w_{gen} and ρ from two diffusion models trained on ImageNet-1k, one in pixel space (EDM (Karras et al., 2022)) and one in latent space (DiT-XL/2 (Peebles & Xie, 2023)). The DiT-XL/2 curve is used only for guidance on DiT-XL/2 itself; every classifier is trained with the EDM curves. All discriminative recipes are ResNet-50s (He et al., 2016) trained on ImageNet-1k. Our 64×64 models replace the ResNet-50 stem with a single 3×3 stride-1 convolution and no max-pooling, so the residual stages see feature maps of about the size they see at 224 pixels and all later layers are unchanged (a common choice for small inputs; e.g., He et al., 2016). We use the GeLU activation function and train for 300 epochs with a batch size of 2048 (SGD, momentum 0.9, learning rate 0.64, cosine schedule, 5 warm-up epochs, weight decay 5×10^{-5}). The noised input is scaled by $1/\sqrt{\sigma^2 + \sigma_{\text{data}}^2}$ with $\sigma_{\text{data}} = 0.5$ as in Equation 7. The matched control is the identical recipe with no noise.

The human-alignment gap. The main result is already in Section 1: after setting the right target by adding diffusion classifiers at controlled data and resolution, DISC retains clean accuracy, matches the diffusion classifier's OOD accuracy, exceeds its error consistency, and narrows its shape-bias gap (Table 1). The rest of this section presents ablations and controlled experiments.

4.1 The 2×2 ablation and a drop-in test inside prior art

The 2×2 ablation. Vanilla noise augmentation (uniform noise-level sampler, hard labels), which is the recipe of Jaini et al. (2024) at 64 pixels, already achieves the human alignment gains they report. However, this costs 11.1 points of clean accuracy against the control. Introducing w_{gen} noise-level sampler alone recovers +8.6 points of clean accuracy and +3.3 points of out-of-distribution accuracy while preserving the error consistency. The combination of w_{gen}+ρ is the only setting that preserves clean accuracy compared to the baseline with no augmentation at all and achieves maximum mCA (mean corruption accuracy on ImageNet-C) while roughly preserving other metrics. This is consistent with the toy examples in Figure 2: w_{gen} emphasizes levels with larger measured contributions, while ρ softens targets as the cumulative attribution decreases. Table 2 tests the practical effect of this coupling.

Our schedule alone, inside prior work. Both ingredients (w_{gen} and ρ) above are ours. However, the mechanism of coupling input noise with label smoothing is already published by Heinonen et al. (2025). They use the SNR heuristic $\gamma(\sigma) = 1/(1 + \text{SNR}(\sigma))^n$, which only depends on the input dimension and a free parameter n that is tuned, which means it is agnostic to the data distribution where it will be applied; induced problems are already discussed in Figure 2c, and now in Table 3, we drop our ρ in place of their heuristic and keep every other hyperparameter unchanged. This improves the reported mean on every metric they used in their setting. Their batches are one third clean, one third noised, and one third blurred. Since we only modify the noised third, increasing noise share to 60% further amplifies the improvements, including clean accuracy.

Table 3: **Isolating the label schedule inside prior art.** PreResNet-50 on TinyImageNet in the pipeline of Heinonen et al. (2025) (their sampler, blur branch, FCR+TrivAug augmentation, 300 epochs; their Table 5). In our rows, their SNR label map is replaced by our measured $1-\rho(\sigma)$, and nothing else is changed. Their batches are one third clean, one third noised, and one third blurred (`xnb`). Since we only modify the noised third, we add a row where 60% of the batch is noised (`xnnnb`). Cells are mean$_{(\pm\text{std})}$ over three seeds; error in %; lower is better. Bold is the best value in each column; green marks whether there is an improvement over the Heinonen baseline.

	Error ↓		NLL ↓		ECE ↓	
Label schedule	clean	corr	clean	corr	clean	corr
No mollification (Heinonen et al., 2025)	$33.8_{\pm.13}$	$69.9_{\pm.35}$	$1.57_{\pm.01}$	$4.64_{\pm.12}$	$0.13_{\pm.00}$	$0.30_{\pm.00}$
Heinonen, as reported	$32.4_{\pm.21}$	$59.2_{\pm.62}$	$1.43_{\pm.01}$	$3.04_{\pm.05}$	$0.11_{\pm.00}$	$0.17_{\pm.00}$
Heinonen repro. (γ_{SNR} labels)	$32.43_{\pm.29}$	$59.18_{\pm.10}$	$1.440_{\pm.015}$	$3.050_{\pm.010}$	$0.112_{\pm.001}$	$0.171_{\pm.000}$
+ **matched ρ labels (ours)**	$31.51_{\pm.08}$	$59.11_{\pm.61}$	$1.402_{\pm.008}$	$3.018_{\pm.047}$	$0.110_{\pm.001}$	$0.168_{\pm.004}$
+ ρ labels, 60% noised (`xnnnb`)	$\mathbf{30.66}_{\pm.31}$	$\mathbf{58.38}_{\pm.68}$	$\mathbf{1.374}_{\pm.006}$	$\mathbf{2.935}_{\pm.047}$	$\mathbf{0.108}_{\pm.002}$	$\mathbf{0.159}_{\pm.003}$

Table 4: ResNet-50 with input resolution 224×224, evaluated on clean ImageNet, ImageNet-C (Hendrycks & Dietterich, 2019), and the natural distribution shifts of ImageNet-testbed (Taori et al., 2020). Upper block: the authors' released checkpoints as distributed with ImageNet-testbed. mCE uses ImageNet-C predictions with AlexNet normalization. Bold is best, underline second.

	Classification	Corruptions	Natural distribution shifts					
	Clean Acc.(↑)	ImageNet-C mCE(↓)	IN-V2 Acc.(↑)	IN-A Acc.(↑)	IN-R Acc.(↑)	Sketch Acc.(↑)	ObjectNet Acc.(↑)	Stylized Acc.(↑)
From the released predictions of Taori et al. (2020)								
Vanilla (torchvision V1)	76.1	76.7	63.2	0.0	36.2	24.1	35.8	20.5
MixUp (Zhang et al., 2018)	77.5	69.2	65.0	6.5	39.6	26.9	37.1	25.2
CutMix (Yun et al., 2019)	**78.6**	73.2	65.8	**7.3**	34.8	23.7	38.1	18.9
Gaussian-noise aug. (Taori et al., 2020)	73.0	75.5	60.6	2.5	37.4	23.3	30.4	21.3
AugMix (Hendrycks et al., 2020)	77.5	65.9	65.1	3.8	41.0	28.5	36.4	26.5
Recent baselines and augmentation methods, using their released checkpoints								
Diffusion noise (Jaini et al., 2024; Gavrikov, 2024)	67.2	73.1	54.7	2.4	36.6	20.0	26.0	23.0
IPMix (Huang et al., 2023)	77.7	62.7	65.7	4.1	42.6	31.1	38.3	28.5
our setting								
Matched control (our hyperparameters, no noise)	76.1	78.2	63.6	3.5	34.7	22.4	34.5	17.0
DISC	76.6	69.3	64.7	4.0	40.5	26.2	34.6	24.2
DISC + IPMix	**78.6**	**60.0**	**66.9**	5.4	**43.3**	**31.4**	**38.8**	**30.1**

4.2 Robustness at 224×224 resolution

So far, all experiments were run at 64×64 pixels, where our target diffusion classifier operates. In this section we apply DISC training to ResNet-50 at its native resolution (224×224) and compare it with augmentation methods on the ImageNet-testbed suite (Taori et al., 2020) in Table 4. The curves were measured at 64 pixels. We therefore stretch them by $\sqrt{224/64}$ (i.e. $\rho^{224}(\sigma) := \rho^{64}\big(\sigma/\sqrt{224/64}\big)$ and $w_{\text{gen}}^{224}(\sigma) := w_{\text{gen}}^{64}\big(\sigma/\sqrt{224/64}\big)$). We keep the architecture and augmentation of the torchvision V1 baseline, and only update its hyperparameters: 300 epochs, batch 2048, cosine schedule, weight EMA, learning rate 1.28, and weight decay 2.5×10^{-5}.

As before, DISC training costs no clean accuracy, although we train on noised images only without clean-image mixing; clean accuracy is 76.6 against 76.1 for the vanilla checkpoint. To check that the improvement comes from the recipe and not from the changed hyperparameters, we train a matched control with the same hyperparameters and no noise. It lands at 76.1, the vanilla number, and DISC improves on it in every column. Despite these gains, DISC falls behind IPMix (Huang et al., 2023), a strong recent augmentation method that is a tuned mix of previous augmentation methods (e.g., mixing set from PixMix, image-level mixing from MixUp, patch-level mixing from CutMix, JSD loss from AugMix), yet does not use noise. There are several ways to combine our recipe with theirs, e.g. applying IPMix to noised images or mixing within the batch. We choose the simplest way: the released IPMix script fine-tunes the torchvision V1 checkpoint for 180 epochs, and we feed it our checkpoint instead, changing nothing else. The combination is the best on every column of Table 4 except ImageNet-A.

4.3 DISC-GUIDANCE AND $w_{\text{gen}}(\sigma)$ GUIDANCE-RESCALING

DISC can replace the noise-conditional classifier in classifier guidance without retraining, and $w_{\text{gen}}(\sigma)$ can replace the constant guidance scale in both classifier guidance and classifier-free guidance. We test both on released samplers under their authors' protocols (50k samples for ImageNet-1k). For classifier guidance, we use the ADM ImageNet 64×64 model of Dhariwal & Nichol (2021) in their setting; our pipeline reproduces their unguided FID of 2.07. We sweep the guidance scale s for three classifiers, their noise-conditional classifier, DISC, and the vanilla ResNet-50 trained at 64 pixels without noise, each under a constant scale and under the rescaled scale $s \cdot w_{\text{gen}}(\sigma_t)$.

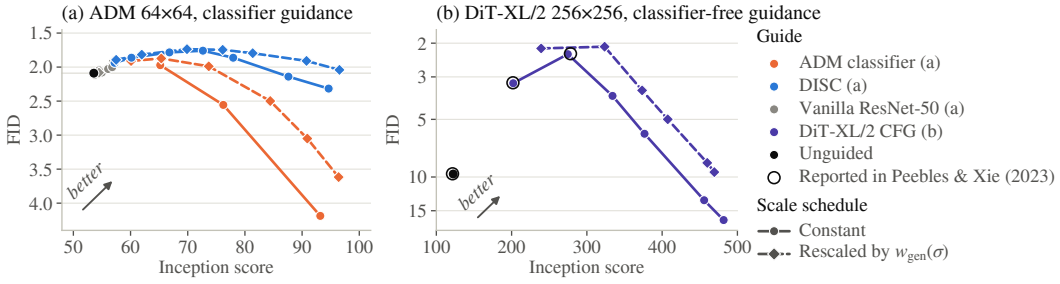


Figure 3: **DISC as the guide, and $w_{\text{gen}}(\sigma)$ as the guidance schedule.** 250 steps and 50k samples. **(a)** Classifier guidance on the released ADM ImageNet 64×64 model, swept over the guidance scale s. **(b)** Classifier-free guidance on the released DiT-XL/2 at 256×256 under the protocol of Peebles & Xie (2023), swept over w. Hollow circles are their reported values. The rescaled series replaces $w-1$ by $(w-1)\,w_{\text{gen}}(\sigma_t)$, with the curve measured on DiT itself in its latent space (Section 3.1).

Classifier guidance. ADM's own classifier under its constant scale improves FID only at the smallest scale and degrades from there. Rescaling its scale by $w_{\text{gen}}(\sigma)$ repairs most of the damage. DISC under a constant scale is better still at every matched Inception score. Both together lower FID from 2.07 to **1.74** at s=1, raise Inception score from 53 to 70, and dominate every other series at every scale we swept.

Classifier-free guidance. We run the DiT-XL/2 (Peebles & Xie, 2023) at 256×256 under its authors' protocol, whose published anchors our constant-scale runs reproduce, and replace the guidance weight $w-1$ by $(w-1)\,w_{\text{gen}}(\sigma_t)$ with the curve measured on DiT itself in its latent space. The rescaled weight wins at every w (Figure 3b). At w=1.5 it reaches FID **2.11** against 2.30 with Inception score 322 against 275. A second seed moves both series by at most 0.04, well below the gap between them.

5 CONCLUSION

We asked where the human alignment of diffusion classifiers comes from, and whether an ordinary classifier can have it. First, we decomposed the alignment gap into the axes along which the two model families compared in prior work differ, and set the right target at matched training data and resolution (EDM against a ResNet-50, both trained on ImageNet at 64×64). We then measured the contribution of each noise level through reductions in KL divergence to the diffusion classifier's full-score prediction, and used the resulting normalized attribution and its cumulative sum to set the noise-sampling and label-smoothing schedules. Training with input noise and label smoothing coupled through these curves keeps clean accuracy and closes the gap to the target diffusion classifier in OOD accuracy and error consistency, and narrows it in shape bias. Unlike noise augmentation over the full diffusion noise range, it costs no clean accuracy. The curves are measured on the dataset and nothing in it is tuned on the task. Finally, the w_{gen} curve works as a guidance schedule that improves over a constant scale on two released samplers, and DISC guides the ADM sampler better than the noise-conditional classifier trained for that purpose.

AI USE STATEMENT

We used LLMs to draft and revise sections of the paper, and to polish the parts that we wrote ourselves. In both cases we edited and verified the output before submission, and all claims and

perspectives in this paper are our own. We also used LLMs as coding assistants to speed up the experiments. The design of the experiments is ours, and we verified and corrected the output of the coding agents. Finally, in addition to the related work we were already aware of, we used LLMs to find relevant literature to cite and position our work against.

Reproducibility statement

We will release the source code, the measured ρ tables for EDM ImageNet-64 and DiT-XL/2, cached diffusion-classifier predictions, and the DISC checkpoints. Every external model we use is public.

References

Tom Burgert, Oliver Stoll, Paolo Rota, and Begüm Demir. Imagenet-trained CNNs are not biased towards texture: Revisiting feature reliance through controlled suppression. *arXiv preprint arXiv:2509.20234*, 2025.

Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Your diffusion model is secretly a certifiably robust classifier. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. URL https://arxiv.org/abs/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 *International Conference on Machine Learning (ICML)*, 2024b. URL https://arxiv.org/abs/2305.15241.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2303.15233.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2105.05233.

Paul Gavrikov. diffusion-noise-cls-pytorch: Unofficial PyTorch port of the diffusion-noise frontend of Jaini et al. https://github.com/paulgavrikov/diffusion-noise-cls-pytorch, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020a. URL https://arxiv.org/abs/2004.07780.

Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. 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*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/9f6992966d4c363ea0162a056cb45fe5-Abstract.html.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems 34*, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6572.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Markus Heinonen, Ba-Hien Tran, Michael Kampffmeyer, and Maurizio Filippone. Robust classification by coupling data mollification with label smoothing. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025. URL https://arxiv.org/abs/2406.01494.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1903.12261.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16762–16771. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01628.

Katherine L Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Zhenglin Huang, Xiaoan Bao, Na Zhang, Qingqi Zhang, Xiao Tu, Biao Wu, and Xi Yang. Ipmix: Label-preserving data augmentation method for training robust 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.

Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.16779.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2206.00364.

Simran Kaur, Jeremy Cohen, and Zachary C. Lipton. Are perceptually-aligned gradients a general property of robust classifiers? In *NeurIPS 2019 Workshop on Science Meets Engineering of Deep Learning*, 2019.

Bahjat Kawar, Roy Ganz, and Michael Elad. Enhancing diffusion-based image synthesis with robust classifier guidance. *Transactions on Machine Learning Research*, 2023.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

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. URL https://arxiv.org/abs/2303.16203.

Alexander C. Li, Ananya Kumar, and Deepak Pathak. Generative classifiers avoid shortcut solutions. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2512.25034.

Yang Liu, Shen Yan, Laura Leal-Taixé, James Hays, and Deva Ramanan. Soft augmentation for image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jiajun Ma, Tianyang Hu, Wenjia Wang, and Jiacheng Sun. Elucidating the design space of classifier-guided diffusion generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1706.06083.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.02629.

Jorge Chang Ortega, Bastien Le Lan, Thomas Serre, and Victor Boutin. Not too generative, not too discriminative: The human alignment sweet spot, 2026. URL https://arxiv.org/abs/2605.23819.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yao Qin, Xuezhi Wang, Balaji Lakshminarayanan, Ed H. Chi, and Alex Beutel. What are effective labels for augmented data? improving calibration and robustness with AutoLabel. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.

Evgenia Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pp. 53–69. Springer, 2020. doi: 10.1007/978-3-030-58580-8_4.

Suraj Srinivas, Sebastian Bordt, and Himabindu Lakkaraju. Which models have perceptually-aligned gradients? an explanation via off-manifold robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6199.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://arxiv.org/abs/1512.00567.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1805.12152.

Max Wolff, Thomas Klein, Evgenia Rusak, Felix A. Wichmann, and Wieland Brendel. Low-pass filtering improves behavioral alignment of vision models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YhgBy6jTR8.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

A EMPIRICAL BEHAVIOR AND ESTIMATION OF THE ATTRIBUTION CURVES

This appendix examines the measured attribution curves used by DISC: the behavior of individual images and their dataset average, and the stability of the estimate across data subsets. The curves depend on the diffusion model and the scoring convention. Figure 5 shows similar average curves from 1,000 and 12,000 training images and from a separate 1,000-image validation subset. We do not test whether different diffusion models in the same representation space yield the same curves; the DiT-XL/2 curve is measured in latent space and is not directly comparable to the EDM pixel-space curve.

A.1 PER-IMAGE CONTRIBUTIONS AND THE DATASET AVERAGE

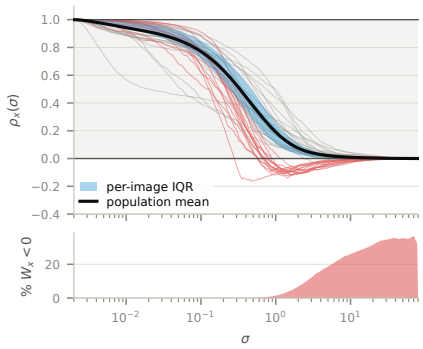


Figure 4: Nothing in the formulation bounds $W_{\boldsymbol{x}}(t)/W_{\boldsymbol{x}}(0)$ to $[0, 1]$ or makes it monotone in σ. Top: per-image curves for 12,000 training images (grey; blue: interquartile band; red: the 15 deepest excursions below zero); 59% of the images go negative at some level. Their average, ρ (black) stays in $[0, 1]$ and decreases at every level. Bottom: percentage of images with $W_{\boldsymbol{x}}(t) < 0$ at each level.

The per-image construction does not guarantee non-negative contributions or a cumulative curve in $[0, 1]$. In particular, $W_{\boldsymbol{x}}(t) < 0$ when $\mathrm{KL}(q_0 \,\|\, q_t) > \mathrm{KL}(q_0 \,\|\, q_T)$: the accumulated prediction is farther from the full-score reference than the highest-noise prediction is. This is a statement about agreement with the reference prediction, regardless of whether that reference predicts the true class. On 59.3% of the 12,000 estimation images, $W_{\boldsymbol{x}}$ is negative at some level, mostly at high σ (Figure 4). After per-image normalization and averaging, every measured entry of w_{gen} is non-negative without clipping. Together with its unit sum, this makes it a sampling distribution and makes ρ a valid label-interpolation coefficient. These are observed properties of the estimated curves, not guarantees for arbitrary diffusion models or datasets.

A.2 ESTIMATION DETAILS

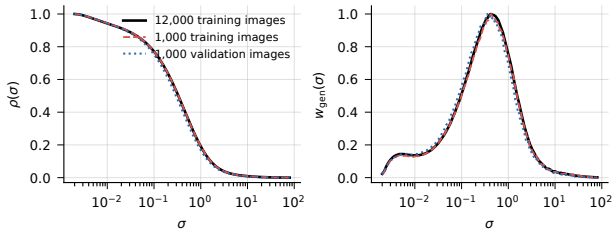


Figure 5: $\rho(\sigma)$ and $w_{\mathrm{gen}}(\sigma)$ read from 12,000 training images (the table used throughout), from 1,000 training images, one per class, and from 1,000 validation images.

We estimate the curves on 12,000 training images, 12 per class, on the 250 step noise level grid in $[0.002, 80]$ (Karras et al., 2022). $\rho = 1$ is anchored at the smallest grid node, $\sigma_{\min} = 0.002$, rather than at $\sigma = 0$, and the pure-noise reference q_T is the prediction at the highest node rather than exactly uniform (and is close to uniform in practice). The estimate saturates well before the full set: 1,000 training images give the curve of 12,000 (Figure 5).