Robust Diffusion Models via Divergence-Induced Weighted Denoising

Lei Li[1], Yuexiao Dong[2]
[1]LunarAI LLC, Newark, DE
[2]Fox School of Business, Temple University

Contaminated training data

- Diffusion models are trained with the MSE denoising loss $\sum_t \mathbb{E}\|\epsilon - \epsilon_\theta(x_t, t)\|^2$. Every sample gets the same gradient weight, so a fraction ε of corrupted images pulls the model toward the corruption.
- **Question.** Is there a principled family of robust denoising losses, and how do we choose within it?
- **Answer.** Every f-divergence between the true and the model reverse kernels collapses to a scalar function of the ordinary denoising residual, and its derivative is a bounded, redescending influence weight.

The per-step likelihood ratio is lognormal

Lemma 2.1. For the DDPM posterior q_t and a model kernel $p_{\theta,t}$, both Gaussian with matched covariance Σ_t, the per-step mismatch is

$$m_t = \tfrac{1}{2}(\tilde{\mu}_t - \mu_{\theta,t})^\top \Sigma_t^{-1}(\tilde{\mu}_t - \mu_{\theta,t}) = w_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2$$

and the likelihood ratio $r_t = q_t/p_{\theta,t}$ satisfies $\log r_t \sim N(-m_t, 2m_t)$ given (x_0, x_t). Every conditional f-divergence is therefore a function of the residual alone:

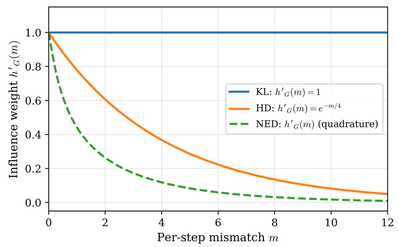
$$D_G(q_t \,\|\, p_{\theta,t}) = h_G(m_t) := \mathbb{E}[G(U_m - 1)], \qquad U_m = e^{-m + \sqrt{2m}Z}, \;\; Z \sim N(0,1)$$

No density-ratio estimate, no discriminator, no second network.

Divergence-induced weighted denoising

Objective $\mathcal{L}_G(\theta) = \sum_t \mathbb{E}[h_G(m_t)]$, with gradient $\sum_t \mathbb{E}[h'_G(m_t)\nabla_\theta m_t]$: the standard DDPM gradient, reweighted per sample by the **influence weight** $h'_G(m_t)$.

Divergence	$G(\delta)$	$h_G(m)$	$h'_G(m)$
KL	$(1+\delta)\log(1+\delta) - \delta$	m	1
HD	$2(\sqrt{(1+\delta)} - 1)^2$	$4(1 - e^{-m/4})$	$e^{-m/4}$
NED	$e^{-\delta} - 1 + \delta$	$\mathbb{E}[\exp(1 - U_m)] - 1$	$\mathbb{E}[\exp(1 - U_m)\, U_m\, (1 - Z/\sqrt{(2m)})]$



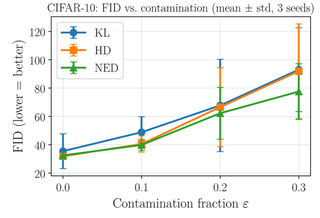
- **One line changes:** replace m_t by $h_G(m_t)$. HD is closed form; NED uses a 64-node Gauss–Legendre quadrature, under 1% overhead. w_t is fixed by the noise schedule, not tuned.

CIFAR-10 under contamination

35M-parameter DDPM, 100K steps, FID at 5K samples. A fraction ε of the training images is replaced by Gaussian noise (noise replacement) or by CIFAR-100 images (structured OOD).

	noise replacement, 3 seeds			structured OOD, 3 seeds			baselines, seed 42			
ε	KL	HD	NED	KL	HD	NED	Huber	Clip	HD	NED
0.00	35.4±12.3	**31.8±1.7**	32.4±2.4	35.4±12.3	**31.8±1.7**	32.4±2.4	**27.8**	**27.7**	32.4	31.6
0.10	48.8±11.0	40.5±5.7	**39.8±3.8**	43.5±10.1	**35.5±10.5**	36.2±6.7	53.6	54.0	43.3	**42.3**
0.20	67.7±32.6	66.5±27.9	**62.2±18.3**	43.9±8.5	**33.7±5.2**	39.1±8.4	87.2	67.5	**45.2**	49.8
0.30	93.0±29.6	91.9±33.4	**77.5±19.7**	49.5±10.0	41.0±4.1	**37.6±3.2**	94.0	90.5	56.2	**55.1**

FID, lower is better. Bold: best per row and group. The $\varepsilon = 0$ runs are shared by the two 3-seed groups.



Two regimes, one rule

- **Noise replacement (far mismatch).** Contaminated samples sit at mismatch 35 against 20 for clean ones (HD weight ratio $e^{-3.75} \approx 0.02$). NED is best under heavy contamination (77.5 against 93.0 for KL at $\varepsilon = 0.30$), and HD and NED beat Huber and clipped MSE.
- **Structured OOD (overlap).** Both robust members beat KL in the mean and at the worst seed at every level, by 5 to 12 FID; HD and NED are not separated, as Proposition 2.1 predicts.
- **Cosine schedule.** The ordering holds: KL 86.2, NED 66.1.

Cleaning composes with the robust objectives

At $\varepsilon = 0.30$ (seed 42), a loss-rank filter followed by NED (32.7) matches oracle removal followed by KL (33.7); filter then KL gives 47.9. Under structured OOD the filter runs at chance (precision and recall 0.31) and filter then KL is worse than KL alone (77.9 against 38.9).

Boundary case

Class-conditional MNIST, $\varepsilon = 0.30$. Conditional accuracy falls from about 97% to about 68% for KL, HD and NED alike. Contaminated mismatch stays small (median 2.3 against 0.17), so no induced weight can separate the populations, as the label-free rule predicts before training.

Classification of the induced objectives

Theorem 2.1. For convex G with $G(0) = G'(0) = 0$ and $G''(0) = 1$:

- **(a) Local expansion.** $h_G(m) = m + O(m^{3/2})$: all divergences agree with KL to first order; the choice matters only at large mismatch.
- **(d) α-family.** $h'_\alpha(m) = e^{-\alpha(1-\alpha)m}$: redescending iff $0 < \alpha < 1$, with the fastest decay uniquely at $\alpha = 1/2$ (Hellinger), $e^{-m/4}$.
- **(e) Universal speed limit.** $h'_G(m) \geq c_G\, m^{-1/2}\, e^{-m/4}$ for large m: no induced weight decays faster, and NED attains it, $h'_{\mathrm{NED}}(m) \sim (e/4)\, m^{-1/2}\, e^{-m/4}$.
- **(f) Exclusion.** h_G is C^∞ and strictly increasing, so Huber (curvature jump) and clipped MSE (flat tail) are induced by no admissible G.

Which member? Leakage against retention

Proposition 2.1. Retention $R(w) = \mathbb{E}_{\mathrm{cl}}[w(m)]$, leakage $L(w) = \mathbb{E}_{\mathrm{ct}}[w(m)]$. If the contaminated mismatch law dominates the clean one in likelihood ratio and w_2/w_1 is nonincreasing (NED against HD), then $L(w_2)/R(w_2) \leq L(w_1)/R(w_1)$.

CIFAR-10, $\varepsilon = 0.30$: 0.692 (NED) ≤ 0.744 (HD); under overlap it is silent.

Label-free rule, before training. Train a 10K-step MSE warmup model and log per-sample m_t at a few early noise levels: a separated far tail points to NED, overlap allows either robust member (HD by default), and all mass at small m means no induced weight can separate the populations.

From path space to finite horizon

- **Motivation (Theorem 2.2).** $D_G(g\|f_\theta) \leq D_G(Q\|P_\theta)$, with equality when the conditional path laws agree.
- **Control (Proposition 2.2).** HD and NED satisfy $c_-\, \varphi_{\mathrm{HD}} \leq \varphi_G \leq c_+\, \varphi_{\mathrm{HD}}$, so

$$D_G(g \,\|\, f_\theta) \leq D_G(Q \,\|\, P_\theta) \leq 4c_+\Big[H(q_T, p_T) + \frac{1}{2\sqrt{c_-}}\Big(\sqrt{\Delta_{G,1}} + \sum_{t=2}^{T} \sqrt{\ell_{G,t}}\Big)\Big]^2$$

Vanishing local, terminal and decoder terms force vanishing path and marginal divergences. Trained models have $a_T \approx 2 \times 10^4$, far from the regime of Theorem 2.3; Proposition 2.2 applies, but loosely (order 10^5).

Selected References

Ho, Jain, Abbeel (2020). "Denoising diffusion probabilistic models." In: NeurIPS.
Huber (1964). "Robust estimation of a location parameter." In: The Annals of Mathematical Statistics 35(1).
Basu, Shioya, Park (2011). "Statistical inference: the minimum distance approach." Chapman and Hall/CRC.
Xu, Nie, Vahdat (2025). "One-step diffusion models with f-divergence distribution matching." arXiv:2502.15681.

Code and result files: supplementary material.