Robust Diffusion Models via Divergence-Induced Weighted Denoising

Lei Li*
LunarAI LLC
Newark, DE 19702

Yuexiao Dong
Fox School of Business
Temple University

Abstract

We show that replacing the standard MSE denoising loss in diffusion models with a nonlinear transformation induced by an f-divergence yields a simple robust training surrogate that empirically improves performance under data contamination, with small additional computational overhead. The theoretical foundation rests on a local divergence construction: under the Gaussian reverse-kernel structure of DDPM, each per-step likelihood ratio follows a lognormal distribution parameterized by a scalar mismatch, so the conditional f-divergence at each step reduces to a one-dimensional function of the denoising error. Summing these local divergences yields a training objective that unifies diffusion training as divergence-induced weighted denoising, where the derivative of the induced divergence acts as a residual-space influence weight that controls the contribution of each sample. Bounded-influence divergences (Hellinger, negative exponential) suppress large-error samples, with Hellinger yielding an explicit exponential weight, connecting the framework to robust M-estimation. We classify the induced objectives by the decay of their influence weights, give a comparison principle for choosing among them under contamination, and relate the local objective to the path divergence through a one-sided finite-horizon bound. Empirically, on CIFAR-10 under 30% contamination, NED reduces FID from 93.0 (KL) to 77.5, while also outperforming standard robust losses such as Huber and clipped MSE.

1 Introduction and Related Work

Diffusion models have become a dominant paradigm for generative modeling, achieving state-of-the-art performance across a wide range of tasks including image, audio, and multimodal generation. Despite their empirical success, the standard training objective, a weighted mean squared error on the denoising residual, treats every training sample equally regardless of how corrupted or atypical it may be. When training data contains corrupted or low-quality examples, this equal weighting degrades generative quality.

In this paper we show that a simple, principled modification to diffusion training can provide robustness to such contamination. The key idea is to replace the standard per-step MSE loss $m_t = w_t\|\epsilon - \epsilon_\theta\|^2$ with a nonlinear transformation $h_G(m_t)$, where the function h_G is induced by an f-divergence. In practice, this amounts to reweighting each sample's gradient by an influence weight $h'_G(m_t)$ that downweights large residuals. For Hellinger divergence, the reweighting is simply $e^{-m_t/4}$; for the negative exponential divergence (NED), it further suppresses high-mismatch samples. The modification adds small computational overhead and requires changing only one line of code. On CIFAR-10 with 30% data contamination, this single change reduces FID from 93.0 (standard KL/MSE) to 77.5 (NED), while also yielding lower cross-seed variance.

*Corresponding author: lei.li@lunarai.llc

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

These practical gains rest on a statistical framework that we develop in this paper. Under the Gaussian reverse-kernel structure commonly used in diffusion models, we show that the per-step likelihood ratio follows a lognormal distribution parameterized by a scalar mismatch m_t, so that every conditional f-divergence at a step is a one-dimensional function of the denoising error. This gives a local conditional training objective of the form $\sum_t \mathbb{E}[h_G(m_t)]$, where h_G is a one-dimensional function determined entirely by the choice of f-divergence G. The standard MSE objective corresponds to KL divergence ($h_{\mathrm{KL}}(m) = m$), recovering the DDPM mean-matching term (up to endpoint contributions in the full ELBO), while alternative divergences induce different transformations of the same residual. The derivative $h'_G(m)$ acts as an influence weight in the sense of robust M-estimation, governing how strongly different mismatch levels contribute to optimization. This connects diffusion training to classical robustness theory and provides a mechanism for the contamination robustness of bounded-influence divergences. A path-space formulation, in which the marginal divergence is lifted to the joint path space and bounded via data processing, motivates the local construction, and a one-sided finite-horizon bound relates the local objective to the path and marginal divergences without asymptotic assumptions.

Our work is related to several lines of research. Diffusion models were introduced in Ho et al. [2020] and further generalized through score-based formulations Song et al. [2021c]; subsequent developments include improved architectures [Karras et al., 2022, Nichol and Dhariwal, 2021], accelerated sampling [Song et al., 2021a, Lu et al., 2022], and one-step generation [Salimans and Ho, 2022, Song et al., 2023]. Song et al. [2021b] connected score matching to a weighted combination of KL divergences but did not frame diffusion training as general divergence minimization at the path level. In the f-divergence literature, Nowozin et al. [2016] introduced critic-based minimization via the Fenchel dual; related density-ratio approaches include Nguyen et al. [2010], Sugiyama et al. [2012]. Li and Vidyashankar [2025] developed a divergence-minimization framework for latent-structure models with monotone decrease and bounded influence functions. Our work specializes this to diffusion, where the Gaussian structure yields an explicit lognormal likelihood ratio without a learned critic. In flow matching, Lipman et al. [2023] learns a velocity field via quadratic loss; our framework suggests extending such objectives to $h_G(\|v_\theta - v^*\|^2)$ (empirical validation left to future work).

Prior work by Xu et al. [2025] generalizes score distillation to f-divergences (f-distill): a one-step student is trained to reduce the f-divergence between the teacher and student marginal distributions, with a gradient weight $f''(r)\, r^2$ of the teacher/student density ratio r times the score difference, evaluated with two auxiliary models, a discriminator for r and a student score network. Our use of f-divergences addresses a different problem with different machinery: robust training from contaminated data rather than distillation, at the level of conditional divergences between Gaussian kernels at each step, where the analogous weight reduces to the explicit expectation $\mathbb{E}[U^2 G''(U-1)]$ with U the lognormal conditional likelihood ratio of Lemma 2.1 (Theorem 2.1(c)); no ratio estimate and no second model is needed. The two works are complementary instances of one principle, curvature-weighted likelihood ratios, at the distribution level there and at the sample level here. The training divergence is also a design axis in discrete diffusion: score-entropy discrete diffusion [Lou et al., 2024], ratio matching under alternative divergence bounds [Haxholli et al., 2025], generalized ELBO families for masked diffusion [Shi et al., 2024, von Rütte et al., 2025], distillation through an integral form of KL [Zheng et al., 2026], and the divergence view of denoising Markov models [Benton et al., 2024]. Those works choose the divergence for ratio or score learnability or for tighter likelihood bounds; neither they nor f-distill address contaminated training data, and whether analogues of our results exist for discrete state spaces, where the lognormal collapse is unavailable, is open. Closest in motivation is learning diffusion models from corrupted data: Daras et al. [2024] and Bai et al. [2024] address measurement corruption, where every observation is corrupted by a known operator that training inverts or integrates out through consistency constraints or expectation maximization, and Chen et al. [2025] distills a model pretrained on noisy data into a one-step generator. Our setting is the complementary one from robust statistics: an unknown fraction ε of the samples is corrupted arbitrarily, with no operator knowledge and no indicator of which samples are affected. Corrupted samples then carry no information about the clean distribution, and oracle knowledge of the corruption amounts to knowing which samples to drop; Section 3 and Appendix J report that oracle, and a practical filter, as reference baselines.

Our main contributions are as follows. First, we establish a lognormal characterization of the per-step likelihood ratio under Gaussian reverse kernels (Lemma 2.1), which, to our knowledge, has

not been explicitly stated in the diffusion literature; it makes every conditional f-divergence at a step an explicit scalar function of the denoising residual and yields the local objective (1). Second, we classify the induced objectives (Theorem 2.1): local expansion, global monotonicity, an exact influence-weight identity, exact decay rates on the α-family with Hellinger as the critical member, a universal limit on the decay rate that NED attains, and the exclusion of Huber and clipped MSE. Third, we give a comparison principle for choosing among induced objectives under contamination (Proposition 2.1) with a label-free selection rule. Fourth, we relate the local objective to the path and marginal divergences: the path-space bound (Theorem 2.2) motivates the construction, a one-sided finite-horizon bound (Proposition 2.2) controls the path and marginal divergences by the local terms without smallness or independence assumptions, and an asymptotic product-to-sum reduction (Theorem 2.3, Corollary 2.1) gives the small-mismatch interpretation, reported next to a measurement of its assumptions. Finally, we validate the framework on CIFAR-10 (three seeds), CIFAR-100, two contamination regimes, data-cleaning baselines and class-conditional MNIST: HD and NED outperform KL and standard robust losses under contamination.

Section 2 develops the framework: the lognormal characterization and the local objective (2.1), the induced design space (2.2), divergence selection under contamination (2.3), and the relation to the path-space divergence (2.4). Section 3 presents experiments and Section 4 discusses limitations; proofs, audits, and additional experiments are in the appendix.

2 A Divergence Framework for Diffusion Models

2.1 Background and the exact local collapse

Our framework builds on the standard diffusion model setup, which we summarize here to establish notation. A diffusion model defines a forward process that progressively adds Gaussian noise to data $x_0 \sim g$ over T steps,

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t;\, \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t I),$$

with a noise schedule $\{\beta_t\}_{t=1}^T$. For large T, the terminal distribution $q(x_T)$ is approximately $\mathcal{N}(0, I)$. Generation proceeds by learning a reverse process $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1};\, \mu_{\theta,t}(x_t),\, \Sigma_t)$, where Σ_t is typically fixed to $\beta_t I$ or $\tilde{\beta}_t I$ [Ho et al., 2020]. Under the standard ϵ-parameterization [Ho et al., 2020, Song et al., 2021c], the network $\epsilon_\theta(x_t, t)$ predicts the noise component, and the training objective reduces to a weighted mean squared error $\sum_t w_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2$.

Our framework is built on the general class of f-divergences. Given two distributions P and Q with $P \ll Q$, the f-divergence generated by a convex function $G : [-1, \infty) \to \mathbb{R}$ with $G(0) = 0$ is defined as $D_G(P\|Q) = \mathbb{E}_Q[G(dP/dQ - 1)]$. When $G(\delta) = (1+\delta)\log(1+\delta) - \delta$ this recovers the KL divergence; $G(\delta) = 2(\sqrt{1+\delta} - 1)^2$ gives the squared Hellinger distance; and $G(\delta) = e^{-\delta} - 1 + \delta$ defines the negative exponential divergence (NED) [Basu et al., 1997]. The normalization $G'(0) = 0$, $G''(0) = 1$ is adopted throughout (hence $G \geq 0$, since a convex function lies above its tangent at 0, which is the zero line), as is classical for disparity generators [Basu et al., 2011, p. 29].

A key structural observation that underlies our framework is the following characterization of the per-step likelihood ratio.

Lemma 2.1 (Lognormal likelihood ratio)**.** *Consider the DDPM variational posterior $q_t(x_{t-1} \mid x_t, x_0) = \mathcal{N}(\tilde{\mu}_t, \Sigma_t)$ and a model reverse kernel $p_{\theta,t}(x_{t-1} \mid x_t) = \mathcal{N}(\mu_{\theta,t}, \Sigma_t)$ with matched covariance. Define the per-step mismatch $m_t := \frac{1}{2}(\tilde{\mu}_t - \mu_{\theta,t})^\top \Sigma_t^{-1}(\tilde{\mu}_t - \mu_{\theta,t})$. Then, conditional on (x_0, x_t), the per-step likelihood ratio $r_t := q_t/p_{\theta,t}$ evaluated under the local model kernel $p_{\theta,t}(\cdot \mid x_t)$ satisfies*

$$\log r_t \sim N(-m_t,\, 2m_t).$$

In particular, r_t is lognormal with $\mathbb{E}[r_t \mid x_0, x_t] = 1$ and $\mathrm{Var}(r_t \mid x_0, x_t) = e^{2m_t} - 1$.

The proof is a direct calculation (Appendix A). Although the lognormal form of the Gaussian likelihood ratio is classical, this characterization parameterized by m_t has not, to our knowledge, been stated in the diffusion literature. Existing work computes per-step KL divergences without isolating the distributional form of r_t [Ho et al., 2020, Song et al., 2021b]. The lemma is unaffected

by additional conditioning: for a class-conditional reverse kernel $p_{\theta,t}(x_{t-1} \mid x_t, c)$ with matched covariance the same calculation gives $\log r_t \sim N(-m_t, 2m_t)$ with $m_t = m_t(x_0, x_t, c)$, since the proof uses only Gaussianity at fixed conditioning.

We construct the training objective from the local conditional f-divergences at each step,

$$\mathcal{L}_G(\theta) := \sum_{t=1}^{T} \mathbb{E}_{q(x_0,x_t)}[D_G(q(\cdot \mid x_t, x_0) \parallel p_{\theta,t}(\cdot \mid x_t))] .$$

By Lemma 2.1, the conditional divergence at each step is a function of the scalar mismatch $m_t = m_t(x_0, x_t)$: $D_G(q(\cdot \mid x_t, x_0) \,\|\, p_{\theta,t}(\cdot \mid x_t)) = h_G(m_t)$, where $h_G(m) := \mathbb{E}[G(U_m - 1)]$ with $\log U_m \sim N(-m, 2m)$. Hence

$$\mathcal{L}_G(\theta) = \sum_{t=1}^{T} \mathbb{E}_{q(x_0,x_t)}[h_G(m_t(x_0, x_t))] . \tag{1}$$

This objective is mathematically well-defined for any f-divergence G and does not require the path likelihood-ratio factorization. For KL divergence, $h_{\mathrm{KL}}(m) = m$ and the objective recovers the standard DDPM denoising loss: under the variational weights w_t, $\mathcal{L}_{\mathrm{KL}}$ is the θ-dependent part of the negative DDPM ELBO up to θ-independent terms, while the widely used L_{simple} reweights timesteps and loses this exact identity. For general f-divergences, each term of (1) is an exact conditional f-divergence, and the sum is a tractable training surrogate motivated by the path-space upper bound of Section 2.4, where its relation to the path-space divergence is made precise; treating the local criterion as primary when the ideal target is intractable follows generalized variational inference [Li and Turner, 2016, Knoblauch et al., 2022].

2.2 The induced design space: divergence-induced weighted denoising

By the local conditional construction (1), the training objective is $\mathcal{L}_G(\theta) = \sum_{t=1}^{T} \mathbb{E}_{q(x_0,x_t)}[h_G(m_t)]$, where $m_t = m_t(x_0, x_t)$ is the one-step Gaussian mismatch. Under the standard ϵ-parameterization, this mismatch takes the form

$$m_t = w_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2, \qquad w_t = \tfrac{\beta_t^2}{2\alpha_t(1-\bar{\alpha}_t)\sigma_t^2},$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$, and σ_t^2 is the reverse-kernel variance. Note that w_t depends only on the noise schedule, not on the divergence, and is not a tuned quantity: m_t is the exact Mahalanobis half-distance between the matched-covariance posterior and model kernels, and x_0- or v-prediction, or another noise schedule, prescribe the corresponding w_t without introducing a separate robustness threshold. In stochastic training, the objective is estimated by sampling x_0, t, ϵ and evaluating $h_G(m_t)$.

To understand the optimization behavior, differentiate $\mathcal{L}_G(\theta)$ with respect to θ:

$$\nabla_\theta \mathcal{L}_G(\theta) = \sum_{t=1}^{T} h'_G(m_t)\, w_t\, \nabla_\theta \|\epsilon - \epsilon_\theta(x_t, t)\|^2.$$

Thus, the effective per-sample weight is $h'_G(m_t)\, w_t$. The quantity $h'_G(m)$ therefore plays the role of an adaptive residual-space influence weight, analogous to the ψ-function in robust M-estimation, and diffusion training under a general divergence can be interpreted as a divergence-induced weighted denoising problem: the residual $\|\epsilon - \epsilon_\theta(x_t, t)\|^2$ is unchanged, but the divergence determines how strongly different mismatch levels are emphasized.

Table 1 summarizes the three cases. The resulting training algorithm (Figure 1) differs from standard DDPM training by a single modification: the MSE loss m_t is replaced by $h_G(m_t)$, or equivalently, each sample's gradient is reweighted by $h'_G(m_t)$. When h_G and h'_G admit closed forms (KL, HD), they are evaluated directly. When no closed form exists (e.g., NED), $h'_G(m)$ is approximated via Gauss–Legendre quadrature over the lognormal variable U_m from Lemma 2.1; our ablations (Appendix J) show that 16 quadrature points suffice.

The design space is the set of scalar maps $h_G(m) = \mathbb{E}[G(U_m - 1)]$, $\log U_m \sim N(-m, 2m)$, indexed by admissible generators G. The following theorem classifies it: (a) and (b) describe the shape of h_G, and (c) to (f) the influence weight h'_G, which governs robustness.

Table 1: Shifted generators, induced local divergences, and influence weights for the three classical divergences; U_m is the lognormal likelihood ratio of Lemma 2.1 and $Z \sim N(0,1)$.

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

Divergence-Induced Diffusion Training

Input: dataset $\mathcal{D}$, divergence generator G, noise schedule $\{\beta_t\}$, learning rate η
Output: trained denoiser ϵ_θ

repeat
1. Sample $x_0 \sim \mathcal{D}$, $t \sim \mathrm{Uniform}\{1,\dots,T\}$, $\epsilon \sim \mathcal{N}(0,I)$
2. Compute $x_t = \sqrt{\bar\alpha_t}\, x_0 + \sqrt{1-\bar\alpha_t}\,\epsilon$
3. Compute per-step mismatch $m_t = w_t\|\epsilon - \epsilon_\theta(x_t,t)\|^2$
4. Evaluate influence weight:
if h'_G has closed form (e.g., KL: $h'_G=1$; HD: $h'_G = e^{-m/4}$): evaluate directly
else (e.g., NED): approximate $h'_G(m_t)$ via quadrature over U_m (Lemma 2.1)
5. Update $\theta \leftarrow \theta - \eta\, h'_G(m_t)\, w_t\, \nabla_\theta\|\epsilon - \epsilon_\theta(x_t,t)\|^2$
until convergence

Figure 1: Divergence-induced diffusion training. The only modification to standard DDPM is step 4: each sample's gradient is reweighted by the influence weight $h'_G(m_t)$.

Theorem 2.1 (Classification of induced objectives)**.** *Let $G:[-1,\infty)\to\mathbb{R}$ be convex with $G(0) = G'(0) = 0$ and $G''(0) = 1$, and let $h_G(m) = \mathbb{E}[G(U_m - 1)]$ with $U_m = \exp(-m + \sqrt{2m}\,Z)$, $Z \sim N(0,1)$, assumed finite for all $m \geq 0$.*

(a) (Local expansion.) *If $G \in C^3$ near 0 and $\mathbb{E}[|G(U_m-1)|\,\mathbf{1}\{|U_m-1|>\eta\}] = O(m^{3/2})$ for some fixed $\eta > 0$, then $h_G(m) = m + O(m^{3/2})$ as $m \to 0$; in particular $h_G(0) = 0$ and $h'_G(0) = 1$. For KL, HD, and NED the sharper rate $h_G(m) = m + O(m^2)$ holds.*

(b) (Global monotonicity.) *h_G is nonnegative and strictly increasing on $[0,\infty)$ with $h_G(0) = 0$.*

(c) (Influence identity.) *If $G \in C^2(-1,\infty)$ has polynomial growth in the sense of Appendix C (as KL, HD, and NED do), then $h'_G(m) = \mathbb{E}[U_m^2\, G''(U_m-1)]$ for every $m > 0$ (the lognormally smoothed curvature of G).*

(d) (α-family.) *For $G''_\alpha(\delta) = (1+\delta)^{\alpha-2}$ under the stated normalization, $h'_\alpha(m) = e^{-\alpha(1-\alpha)m}$: the weight decays to zero if and only if $0 < \alpha < 1$, and the decay coefficient $\alpha(1-\alpha) \leq 1/4$ is maximized uniquely at $\alpha = 1/2$, the Hellinger divergence, giving $e^{-m/4}$; KL is $\alpha = 1$.*

(e) (Universal speed limit.) *Under the hypotheses of (c), there is $c_G > 0$ with $h'_G(m) \geq c_G\, m^{-1/2} e^{-m/4}$ for all large m, so no induced weight decays at a faster asymptotic order; NED attains this order, $h'_{\mathrm{NED}}(m) \sim \frac{e}{4\sqrt{m}}\, e^{-m/4}$ as $m \to \infty$.*

(f) (Exclusion of Huber and clipped MSE.) *Under the growth condition of (c), h_G is infinitely differentiable on $(0,\infty)$. Hence neither the radial Huber loss, whose second derivative jumps at its threshold, nor clipped MSE, which is constant beyond its threshold while h_G is strictly increasing by (b), is induced by any admissible G.*

The proofs are in Appendix C. Part (a) shows that all smooth divergences agree with KL to first order near m=0; the integrability condition holds for KL, HD, and NED because $G(U_m-1)$ grows at most as $U_m \log U_m$ and all polynomial moments of the lognormal U_m are finite (though $\mathbb{E}[e^{cU_m}]=\infty$ for $c>0$). With (b), divergences differ only in their higher-order behavior, which governs robustness, consistent with general divergence-minimization frameworks for latent-structure models [Li and

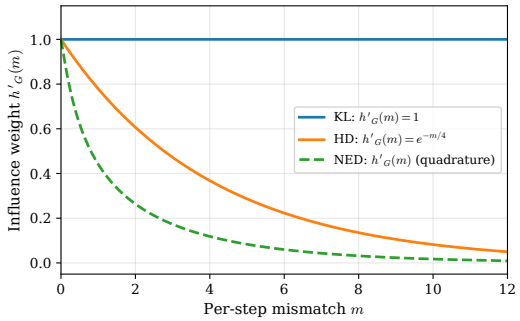

Figure 2: Influence weights $h'_G(m)$ for KL, HD, and NED. KL assigns constant weight to all samples regardless of mismatch; HD and NED both decay exponentially at rate $e^{-m/4}$ (Theorem 2.1(e)), with NED decaying more aggressively due to the polynomial correction.

Vidyashankar, 2025]; since $h'_G(0)=1$, all induced objectives share the same first-order optimization geometry near convergence. Parts (c) to (e) give the classified spectrum: no suppression (KL), the fastest exponential decay within the α-family (HD), and the fastest admissible asymptotic order (NED). HD coincides with the Welsch loss in m up to scale, so redescending weights also arise outside the f-divergence family; what the theorem adds is the derivation and the family-level characterization, and part (f) places the standard Huber and clipping baselines outside the induced class.

The theorem also fixes a trade-off between saturation and gradient retention. Boundedness of h_G and tail decay of h'_G play complementary roles: $h_G(\infty)$ caps an extreme observation's total objective contribution, while $h'_G(m)$ sets its retained gradient signal at mismatch m; they are coupled through $h_G(\infty) = \int_0^\infty h'_G(s)\,ds$, so faster decay suppresses contaminated observations more strongly but may also discard signal from moderately difficult clean samples. Within the α-family the trade is explicit: with $\kappa = \alpha(1-\alpha)$, $h_\alpha(m) = (1-e^{-\kappa m})/\kappa$ and $h_\alpha(\infty) = 1/\kappa$, so Hellinger saturates at 4, while NED saturates at $e-1$ with the additional $m^{-1/2}$ tail suppression of part (e). Appendix D.1 turns measured mismatch quantiles into an admissible range of κ.

2.3 Divergence selection under contamination

The induced objective $\mathcal{L}_G(\theta) = \sum_t \mathbb{E}[h_G(m_t)]$ makes clear that different divergences differ only through their influence weight $h'_G(m)$. For KL divergence, $h'_{\mathrm{KL}}(m) = 1$, so all samples contribute equally regardless of mismatch magnitude: outliers and contaminated samples receive the same weight as clean data, making KL training sensitive to data corruption. For HD, $h'_{\mathrm{HD}}(m) = e^{-m/4}$ decays exponentially with the mismatch m, which provides robustness by strongly downweighting large-error samples, but such rapid decay may also suppress informative samples in early training when mismatches are large. For NED the induced weight has no closed form (Table 1); it decays to zero as $m \to \infty$ at the rate of Theorem 2.1(e): the same exponential rate $e^{-m/4}$ as HD, with an additional $m^{-1/2}$ polynomial correction that makes NED slightly more aggressive. Figure 2 plots the three weights.

Whether the stronger suppression of NED is a gain or a loss depends on where the contaminated samples sit in mismatch, and the following comparison principle makes this exact. Let F_{cl} and F_{ct} denote the mismatch laws of clean and contaminated samples, and for a positive influence profile w define the retention $R(w) = \mathbb{E}_{F_{\mathrm{cl}}}[w(m)]$ (average weight kept on clean samples) and the leakage $L(w) = \mathbb{E}_{F_{\mathrm{ct}}}[w(m)]$ (average weight leaked to contaminated samples).

Proposition 2.1 (Leakage and retention)**.** *Let w_1, w_2 be positive influence profiles with w_2/w_1 nonincreasing in m, so that w_2 is the more aggressive in the tail (numerically verified for $w_2 = h'_{\mathrm{NED}}$, $w_1 = h'_{\mathrm{HD}}$, whose ratio decreases in m and behaves asymptotically as $e/(4\sqrt{m})$). Suppose the likelihood ratio $r = dF_{\mathrm{ct}}/dF_{\mathrm{cl}}$ is nondecreasing in m (contamination separated upward). Then*

$$\frac{L(w_2)}{R(w_2)} \le \frac{L(w_1)}{R(w_1)},$$

that is, the more aggressive profile achieves the better leakage-to-retention ratio.

The proof (Appendix D) is a two-line association inequality under a tilted measure. L/R measures the relative average weight on contaminated versus clean samples. When the two mismatch laws substantially overlap, the likelihood ratio premise does not hold and the proposition is silent; Section 3 reports that regime empirically.

The premise of Proposition 2.1 is measurable before training, and the resulting rule is to inspect the per-sample mismatch distribution, without labels: (1) train a short plain-MSE warmup model (10K steps suffice; on our audit model the 10K-step checkpoint reproduces the leakage-to-retention ratios of the trained model at every timestep slice to within 0.01); (2) log the per-sample m_t at a few noise levels in the first third of the schedule, where the influence weights act (Appendix I); (3) read the pooled histogram. A separated upper tail at large m (tail location q_{ct} with $e^{-q_{\mathrm{ct}}/4} \ll 1$) points to NED; under overlap or a moderate tail either robust member can be used, HD being the closed-form default; all mass at small m means no induced weight can separate the two populations. Our experiments instantiate the regimes. Under noise-replacement contamination the contaminated samples sit far out in mismatch (medians 35.2 against 20.2 on the audit model, an HD weight ratio of $e^{-3.75} \approx 0.02$ at the medians) and NED is best under heavy contamination (Section 3); the per-step leakage-to-retention ratios on the trained model, 0.744 for HD and 0.692 for NED, are ordered as the proposition predicts, at every timestep slice (Appendix I). Under structured out-of-distribution contamination the two laws overlap and no theoretical ordering applies; empirically both robust members improve over KL, and three seeds do not separate them (Section 3). When clean and contaminated samples both remain at small absolute mismatch, no member suppresses contamination materially (class-conditional MNIST, Appendix K); the calibration rule of Appendix D.1 diagnoses this case from the quantiles before training.

2.4 From path-space motivation to finite-horizon control

The natural goal of diffusion training is to minimize the divergence $D_G(g\|f_\theta)$ between the data distribution g and the model marginal f_θ, which is generally intractable because f_θ is defined only implicitly through the reverse chain. A natural approach is to lift the problem to the joint path space $x_{0:T}$, which is the motivation for the local objective (1).

Theorem 2.2 (Lifted path-space divergence is a tangent upper bound)**.** *Let D_G be a general divergence, g, f_θ the marginal data/model distributions on x_0, and Q, P_θ the corresponding lifted path laws on $x_{0:T}$, with $Q \ll P_\theta$ (otherwise the divergence may be infinite). Then $D_G(g\|f_\theta) \le D_G(Q\|P_\theta)$ for every θ, with equality when $Q(x_{1:T} \mid x_0) = P_\theta(x_{1:T} \mid x_0)$ for g-a.e. x_0. In particular, if θ_0 denotes an exact lifted model satisfying $P_{\theta_0} = Q$, then $D_G(g\|f_{\theta_0}) = D_G(Q\|P_{\theta_0})$.*

The proof, with the explicit reduction of the path likelihood ratio under the stated sufficient condition, is in Appendix B. For KL divergence, the lifted divergence yields the standard ELBO, where the path likelihood ratio factorizes exactly into per-step terms plus endpoint contributions. For general f-divergences, the path likelihood ratio contains an additional endpoint factor $g(x_0)q(x_T \mid x_0)/p(x_T)$ that does not separate cleanly from the per-step product under a nonlinear G, and a nonlinear G does not turn a product of dependent path ratios into a sum. Rather than attempting to bound this factor, we use the path-space perspective as motivation and construct the training objective from the local conditional divergences at each step. The division of labor is as follows: Theorem 2.2 motivates the local construction through its tightness condition; the quantitative link to the path divergence is given at finite T by Proposition 2.2 on the actual path law, and in the asymptotic small-mismatch regime by Theorem 2.3 and Corollary 2.1 inside an auxiliary local construction.

Proposition 2.2 (Finite-horizon one-sided control)**.** *Let Q and P_θ be the path laws of Theorem 2.2, q_T, p_T their laws of x_T, and H the Hellinger distance, $H^2(P,R) = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dR})^2$. Write $\phi_G(u) = G(u-1)$ and suppose $c_-\,\phi_{\mathrm{HD}} \le \phi_G \le c_+\,\phi_{\mathrm{HD}}$ on $(0,\infty)$ for constants $0 < c_- \le c_+ < \infty$. Let $\ell_{G,t} = \mathbb{E}_{q(x_0,x_t)}[h_G(m_t)]$ for $t \ge 2$ be the per-step terms of* (1)*, and let $\Delta_{G,1} = \mathbb{E}_{q_1}\big[D_G\big(q(x_0 \mid x_1)\,\|\,p_{\theta,1}(x_0 \mid x_1)\big)\big]$ be the decoder discrepancy. Then*

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

The comparability condition holds for HD with $c_- = c_+ = 1$ and for NED with $c_- = (e-2)/2 \approx 0.359$ and $c_+ \approx 1.073$; KL is outside the class but has the exact ELBO identity of Section 2.1.

The proof (Appendix E) has four steps: Jensen's inequality shows that each optimized, x_0-conditioned term $\ell_{G,t}$ upper bounds the discrepancy $\Delta_{G,t}$ between the model kernel and the true reverse kernel; a hybrid chain of path laws and the triangle inequality for the Hellinger metric bound $H(Q, P_\theta)$ by per-step Hellinger discrepancies plus the terminal term; and the two comparability inequalities convert Hellinger to D_G and back. The three terms have different jobs: $H(q_T, p_T)$ measures how far the noised data are from the prior (at most 0.13 on our schedule); $\Delta_{G,1}$ is the decoder error at $t = 1$, which cannot be estimated, as it involves the unknown $q(x_0 \mid x_1)$, but is bounded in advance for HD and NED ($D_{\mathrm{HD}} \le 4$, $D_{\mathrm{NED}} \le e - 1$), and stopping the chain at $t = 1$ removes it (Appendix E); and $\sum_{t\ge 2} \sqrt{\ell_{G,t}}$ adds up the per-step losses. Hence, at fixed T and without smallness or independence assumptions, vanishing local, terminal and decoder terms imply vanishing path and marginal divergences. At trained losses the bound is loose: on the audited checkpoints $\sum_t \sqrt{\ell_{\mathrm{HD},t}}$ is 698, 529 and 527, giving right-hand sides of order 10^5 against the ceilings $D_{\mathrm{HD}} \le 4$ and $D_{\mathrm{NED}} \le e - 1$ (Appendix E); it becomes informative once $\sum_t \sqrt{\ell_{G,t}}$ is of order one. Part of every $\ell_{G,t}$ is a floor that does not measure model error, because the loss compares the model kernel to the x_0-conditioned posterior rather than to the true reverse kernel: for KL this is exact, $\ell_{\mathrm{KL},t} = \Delta_{\mathrm{KL},t} + I(x_0; x_{t-1} \mid x_t)$, so a large loss need not mean a poor fit, and the same proof gives the bound with the tighter $\Delta_{G,t}$ in place of $\ell_{G,t}$.

Beyond the finite-horizon bound there is an asymptotic small-mismatch regime in which the sum of per-step divergences approximates the divergence of a product likelihood ratio; the following results, whose proofs are in the appendix, make this precise under regularity conditions.

Theorem 2.3 (Product-to-sum reduction)**.** *Let $\delta_t \ge -1$ for all t, $R_T := \prod_{t=1}^T (1 + \delta_t)$, $S_T := \sum_{t=1}^T \delta_t$, $B_T := R_T - 1 - S_T$. Assume $a_T > 0$ is deterministic with $a_T \to 0$, and $G \in C^3$ near 0 with $G(0) = G'(0) = 0$, $G''(0) = 1$. Under the moment conditions (A1)–(A6) on S_T, B_T and $\{\delta_t\}$ and the tail integrability conditions stated in Appendix F,*

$$\mathbb{E}\Big[G\Big(\textstyle\prod_{t=1}^T (1 + \delta_t) - 1\Big)\Big] = \sum_{t=1}^T \mathbb{E}[G(\delta_t)] + o(a_T).$$

Corollary 2.1 (Diffusion Gaussian specialization)**.** *Under the setting of Lemma 2.1, let $\mathbb{E}_{\mathrm{loc}}$ denote expectation under the auxiliary joint construction in which, conditional on (x_0, x_t), each likelihood-ratio increment r_t is evaluated under $p_{\theta,t}(\cdot \mid x_t)$ as in Lemma 2.1. Define $\delta_t := r_t - 1$ and $a_T := \sum_{t=1}^T \mathbb{E}[m_t]$. Under conditions (D1)–(D4) of Appendix G, which require in particular $a_T \to 0$, $\max_t m_t \to 0$ almost surely, and pairwise uncorrelated increments, and under the tail integrability conditions of Theorem 2.3, $\mathbb{E}_{\mathrm{loc}}[G(R_T - 1)] = \sum_{t=1}^T \mathbb{E}_{\mathrm{loc}}[h_G(m_t)] + o(a_T)$.*

For KL the reduction is the exact chain rule of the ELBO; for general f-divergences no chain rule holds, and the first-order reduction provides asymptotic justification for the local construction (1) inside the auxiliary construction. Condition (D1) is the natural local regime in which the per-step mismatches are uniformly small and spread across many steps; condition (D2), pairwise uncorrelatedness of the increments, is an approximation whose quality depends on the model architecture and training stage, and is needed only for the asymptotic justification, not for training.

Whether trained models are near this regime is an empirical question: the conditions above are asymptotic and cannot be verified at fixed T, but the quantities they constrain can be measured on trained models (Appendix H). At the 100K-step checkpoints, $a_T = \sum_t \mathbb{E}[m_t]$ is 2.0×10^4 (KL, clean), 2.0×10^4 (KL, ε=0.30) and 2.5×10^4 (NED, clean), $\max_t m_t$ is about 1.6×10^3, and the per-image mean mismatch has clean median/99th percentile $19.9/30.1$, $20.2/30.7$ and $25.3/35.8$, with the contaminated-sample median at 35.2 against 20.2 on the contaminated run. Re-mapping the trained ε=0.30 model to coarser grids gives $a_T = 2.58/2.20/2.01 \times 10^4$ at $T = 250/500/1000$, so a fourfold change in T moves a_T by less than a factor 1.3. The trained models are therefore far from the $a_T \to 0$ regime; Theorem 2.3 and Corollary 2.1 describe the surrogate in the asymptotic regime, and Proposition 2.2 is the statement that applies to the trained models.

Remark 2.1 (Local objective vs. path-space bound)**.** *Since $h_G(m) = m + o(m)$ (Theorem 2.1(a)), all divergences collapse to nearly the same value when m_t is small (at m=0.01 the three divergences agree to four digits); the choice of divergence matters primarily in the large-mismatch regime where robustness is needed. The local construction is used for* objective design*, not as a convergence guarantee; Section 3 validates effectiveness at finite mismatch.*

Table 2: CIFAR-10 FID (lower is better). Left and middle: mean ± std over 3 seeds under noise replacement and under structured OOD contamination (CIFAR-100 images); the ε=0 runs are shared by both groups. Right: robust-loss baselines under noise replacement, seed 42. Bold: best per row and group.

	noise replacement, 3 seeds			structured OOD (CIFAR-100), 3 seeds			robust 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**

3 Experiments

We evaluate whether the theoretical robustness properties derived in Section 2 translate into measurable differences in generative quality under data contamination. The experiments are controlled mechanism tests rather than a scale benchmark: all methods share identical architecture, hyperparameters, and training budget; only the divergence-induced loss function differs. We use a 35M-parameter DDPM U-Net (base channels 128, channel multipliers $(1, 2, 2, 2)$, self-attention at 16×16) with a linear β_t schedule (β_1=10^{-4}, β_T=0.02), T=1000 forward steps, Adam optimizer at learning rate 2×10^{-4}, batch size 128, and bfloat16 mixed-precision. All runs last 100K gradient steps. FID is computed with 5K generated samples against the clean test split via a 50-step DDIM sampler; the absolute values reflect this deliberately small model and short training budget, and the relative ordering across divergences is the quantity of interest. The HD runs use a 5,000-step plain-MSE warmup before switching to h_{HD}, a precaution against early gradient suppression when all mismatches are large; KL, NED, and the robust baselines use no warmup. A single-seed ablation over warmup lengths 0 to 10,000 on clean data varies FID by less than 3 points, and a cosine-schedule replicate at ε=0.30 reproduces the KL-to-NED ordering (Appendix J). To simulate contamination, a fraction ε of the training images is replaced by Gaussian noise images (each pixel drawn from $\mathcal{N}(0, 1)$ on the $[-1, 1]$ intensity scale and clipped to $[-1, 1]$); we call this the noise-replacement regime. Since the model is unconditional, label information is not used. Each setting is run with three seeds (42, 123, 456); we report mean ± standard deviation.

Table 2 (left) reports CIFAR-10 [Krizhevsky, 2009] FID scores for the three divergence-induced objectives under noise replacement. On clean data, all three divergences achieve comparable FID, consistent with the first-order equivalence $h_G(m) = m+O(m^2)$ (Theorem 2.1(a)). Notably, HD and NED exhibit substantially lower cross-seed variance than KL (1.7 and 2.4 vs. 12.3), suggesting that bounded-influence objectives stabilize training. Under contamination (ε=0.30), NED achieves the best mean FID (77.5), improving over KL by 15.5 points. In this regime the advantage of bounded-influence divergences grows with the contamination level, consistent with the theoretical prediction that the divergence choice matters primarily in the large-mismatch regime; per-seed variability is large for every objective at ε=0.30 (Appendix J), and NED's advantage is clearest in the mean and at the worst seed (92.4 against 124.8 for KL).

We also compare the divergence-induced objectives with two standard robust losses: Huber loss (δ=1.0) and clipped MSE (threshold τ=10). Table 2 (right) reports the results using seed 42 for a controlled single-seed comparison. At moderate to high contamination ($\varepsilon \geq 0.10$), both HD and NED substantially outperform Huber and clipped MSE on this seed: at ε=0.30, HD and NED achieve FID scores of 56.2 and 55.1, compared to 94.0 and 90.5 for the standard baselines. On clean data the four objectives are comparable: the 3–4 FID differences at ε=0 are small relative to run-to-run variation, and additional clean-data seeds for Huber (27.8/33.4/44.5) and clipped MSE (27.7/27.1/47.3, seeds 42/123/456) span ranges that exceed the gaps. The robustness conclusion therefore rests on within-seed degradation as ε increases, which controls for each run's clean baseline: on seed 42, Huber and clipped MSE degrade from clean to ε=0.30 by 66.2 and 62.8 FID against 23.8 and 23.5 for HD and NED, and across matched seeds NED degrades by 45.1 against 57.6 for KL. A second seed at ε=0.30 gives Huber 92.4 and clipped MSE 72.5, so the three-seed KL-to-NED comparison of Table 2 is the primary evidence. The thresholds δ=1.0 and τ=10 are the default values; tuning them is left for future work.

The results so far concern noise replacement, a far-mismatch regime in which contaminated images are perfectly separable from clean ones by loss ranking. Table 2 (middle) reports a second regime in which the contaminated fraction of CIFAR-10 is replaced by CIFAR-100 images (structured, out-of-distribution contamination) that are not trivially separable from clean data. Here both robust objectives improve over KL in the three-seed mean and at the worst seed at every level, by 5 to 12 FID in the means, while three seeds do not separate HD from NED, consistent with the overlap regime, in which Proposition 2.1 expresses no preference: their means differ by at most 5.4 FID and the lower of the two changes with the seed (per-seed values in Appendix J). The two regimes give the practical guidance: NED under heavy far-mismatch contamination (Table 2, left); either robust member near the data distribution, HD being the closed-form default; and the label-free rule of Section 2.3 to tell the regimes apart.

Data cleaning is a complementary remedy (Appendix J): at ε=0.30, oracle removal of the contaminated samples gives FID 33.7 and a practical loss-rank filter followed by NED matches it (32.7), while under structured contamination the filter runs at chance and filtering then KL is worse than KL alone (77.9 against 38.9); HD and NED have no such failure mode.

Additional results, including CIFAR-100 (where NED reduces FID from 98.0 to 75.3 at ε=0.30), a quadrature resolution ablation for NED, the per-seed values behind Table 2, and a class-conditional MNIST study with a conditional-fidelity metric, are reported in Appendices J and K.

The overhead is small: HD adds one `exp` per sample, and NED's 64-node Gauss–Legendre quadrature, vectorized over the batch, adds under 1% to the per-step wall-clock time of the U-Net forward–backward pass, so all three divergences train in about 3h per 100K-step run on a single GPU.

4 Discussion and Limitations

This work provides a divergence-based foundation for diffusion training: the lognormal characterization of the per-step likelihood ratio yields a unified view of diffusion training as divergence-induced weighted denoising, and the experiments support the conclusion that bounded-influence divergences outperform both KL and standard robust losses under contamination. The connection to robust statistics runs deeper than an analogy. In classical M-estimation, robustness is governed by the residual score $\psi(r) = \rho'(r)$; boundedness of ψ is the standard criterion [Huber, 1964, Hampel et al., 1986]. In our setting, with residual $r = \epsilon - \epsilon_\theta$ and mismatch $m = w\|r\|^2$, the residual-space score is $\psi_G(r) = 2w\, h'_G(w\|r\|^2)\, r$. For KL, $h'_{\mathrm{KL}}(m) = 1$, so $\psi_{\mathrm{KL}}(r) = 2wr$ is unbounded in $\|r\|$. For HD and NED, $h'_G(m)$ decays exponentially, making the residual-space score bounded and redescending, thereby suppressing high-mismatch samples. A full estimator-level influence-function analysis, quantifying the breakdown point of the induced training procedure, remains an open direction. The relation to the end target parallels standard practice: KL training optimizes the ELBO, a one-sided bound whose gap to the marginal likelihood is not quantified in practice, usually through L_{simple}, which also drops the exact identity; HD and NED keep exact conditional-divergence semantics at every step, satisfy the one-sided bound of Proposition 2.2, and come with measured diagnostics (Appendices H and I).

From a practitioner's perspective, the recipe is simple: replace the MSE loss with $h_G(w_t\|\epsilon-\epsilon_\theta\|^2)$ for a bounded-influence G. HD is the closed-form default; NED suppresses more strongly at the cost of a quadrature evaluation and is preferred under heavy far-mismatch contamination, the regime the warmup audit of Section 2.3 detects before training; near the data distribution both improve over KL. Higher-resolution validation is the first item of future work; integrating the induced objectives into ambient-type training [Daras et al., 2024, Bai et al., 2024] and distillation from a robustly trained teacher [Chen et al., 2025] are natural next steps. Extension to flow matching and further limitations are discussed in Appendix L.

Acknowledgements and Disclosures

This work received no external funding. The authors declare no competing interests.

References

Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. An expectation-maximization algorithm for training clean diffusion models from corrupted observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ayanendranath Basu, Sahadeb Sarkar, and A. N. Vidyashankar. Minimum negative exponential disparity estimation in parametric models. *Journal of Statistical Planning and Inference*, 58(2): 349–370, 1997.

Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical Inference: The Minimum Distance Approach*. Chapman and Hall/CRC, 2011.

Joe Benton, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. From denoising diffusions to denoising Markov models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):286–301, 2024.

Tianyu Chen, Yasi Zhang, Zhendong Wang, Ying Nian Wu, Oscar Leong, and Mingyuan Zhou. Denoising score distillation: From noisy diffusion pretraining to one-step high-quality generation. *arXiv preprint arXiv:2503.07578*, 2025.

Giannis Daras, Alex Dimakis, and Constantinos Daskalakis. Consistent diffusion meets Tweedie: Training exact ambient diffusion models with noisy data. In *International Conference on Machine Learning (ICML)*, 2024.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

Etrit Haxholli, Yeti Z. Gurbuz, Oğul Can, and Eli Waxman. Efficient perplexity bound and ratio matching in discrete diffusion language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peter J. Huber. Robust estimation of a location parameter. In *The Annals of Mathematical Statistics*, volume 35, pages 73–101, 1964.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lei Li and Anand N. Vidyashankar. Divergence-minimization for latent-structure models: Monotone operators, contraction guarantees, and robust inference. *arXiv preprint arXiv:2511.17974*, 2025.

Yingzhen Li and Richard E. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning (ICML)*, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2021.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021c.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Dimitri von Rütte, Janis Fluri, Yuhui Ding, Antonio Orvieto, Bernhard Schölkopf, and Thomas Hofmann. Generalized interpolating discrete diffusion. In *International Conference on Machine Learning (ICML)*, 2025.

Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

Haoyang Zheng, Xinyang Liu, Cindy Xiangrui Kong, Nan Jiang, Zheyuan Hu, Weijian Luo, Wei Deng, and Guang Lin. Ultra-fast language generation via discrete diffusion divergence instruct. In *International Conference on Learning Representations (ICLR)*, 2026.

Appendix

A Proof of Lemma 2.1

Conditional on (x_0, x_t), the means $\tilde{\mu}_t = \tilde{\mu}_t(x_0, x_t)$ and $\mu_{\theta,t} = \mu_{\theta,t}(x_t)$ are fixed. Under matched covariance Σ_t, the log likelihood ratio between $q_t = \mathcal{N}(\tilde{\mu}_t, \Sigma_t)$ and $p_{\theta,t} = \mathcal{N}(\mu_{\theta,t}, \Sigma_t)$ is

$$\log r_t = \log \frac{q_t(x_{t-1})}{p_{\theta,t}(x_{t-1})} = -\tfrac{1}{2}\|\Sigma_t^{-1/2}(x_{t-1} - \tilde{\mu}_t)\|^2 + \tfrac{1}{2}\|\Sigma_t^{-1/2}(x_{t-1} - \mu_{\theta,t})\|^2.$$

Write $\Delta := \Sigma_t^{-1/2}(\tilde{\mu}_t - \mu_{\theta,t})$ so that $m_t = \frac{1}{2}\|\Delta\|^2$. Under the local model kernel $p_{\theta,t}(\cdot \mid x_t)$, set $z := \Sigma_t^{-1/2}(x_{t-1} - \mu_{\theta,t}) \sim N(0, I)$. Expanding and simplifying gives $\log r_t = \Delta^\top z - \frac{1}{2}\|\Delta\|^2$. Since $\Delta^\top z \sim N(0, \|\Delta\|^2) = N(0, 2m_t)$, we obtain $\log r_t \sim N(-m_t, 2m_t)$ conditional on (x_0, x_t).

The moments follow from standard lognormal calculations: $\mathbb{E}[r_t \mid x_0, x_t] = e^{-m_t + m_t} = 1$ and $\mathrm{Var}(r_t \mid x_0, x_t) = e^{2m_t} - 1$.

The calculation uses only that both kernels are Gaussian with the same covariance at fixed conditioning. If the model kernel carries additional conditioning, for instance a class label c in $p_{\theta,t}(x_{t-1} \mid x_t, c) = \mathcal{N}(\mu_{\theta,t}(x_t, c), \Sigma_t)$, the same steps apply conditionally on (x_0, x_t, c) and give $\log r_t \sim N(-m_t, 2m_t)$ with $m_t = \frac{1}{2}(\tilde{\mu}_t - \mu_{\theta,t}(x_t, c))^\top \Sigma_t^{-1}(\tilde{\mu}_t - \mu_{\theta,t}(x_t, c))$; the objective (1) becomes $\sum_t \mathbb{E}[h_G(m_t(x_0, x_t, c))]$, and the influence weight acts on the denoising residual exactly as in the unconditional case. □

B Proof of Theorem 2.2

The map $x_{0:T} \mapsto x_0$ is a measurable projection from path space to observation space. The marginals of Q and P_θ under this projection are exactly g and f_θ, respectively. Since f-divergences are monotone under measurable maps, we obtain

$$D_G(g\|f_\theta) \le D_G(Q\|P_\theta).$$

For equality, note that equality in the data-processing inequality holds when the likelihood ratio dQ/dP_θ depends only on the projected variable x_0. A sufficient condition is that the conditional latent laws agree given x_0:

$$Q(x_{1:T} \mid x_0) = P_\theta(x_{1:T} \mid x_0) \quad \text{for } g\text{-a.e. } x_0.$$

We now carry out the reduction explicitly, writing everything with densities (the same steps hold at the measure level with regular conditional probabilities). Let $q(x_{0:T})$ and $p_\theta(x_{0:T})$ denote the densities of Q and P_θ. Define the x_0 marginals $g(x_0) = \int q(x_{0:T})\, dx_{1:T}$ and $f_\theta(x_0) = \int p_\theta(x_{0:T})\, dx_{1:T}$, and the conditionals $q(x_{1:T} \mid x_0) = q(x_{0:T})/g(x_0)$ wherever $g(x_0) > 0$ and $p_\theta(x_{1:T} \mid x_0) = p_\theta(x_{0:T})/f_\theta(x_0)$ wherever $f_\theta(x_0) > 0$. With these definitions,

$$q(x_{0:T}) = g(x_0)\, q(x_{1:T} \mid x_0), \qquad p_\theta(x_{0:T}) = f_\theta(x_0)\, p_\theta(x_{1:T} \mid x_0),$$

are identities: each is the chain rule of densities, holding wherever the marginal is positive and almost everywhere overall since densities are defined up to null sets. The integral over $x_{1:T}$ is present inside the definitions of g and f_θ, and each conditional integrates to one in $x_{1:T}$. Note that the model factorization runs opposite to the generative direction $p(x_T)\prod_t p_\theta(x_{t-1} \mid x_t)$: the conditional given x_0 is defined by the ratio above (a disintegration), is generally intractable, and is never evaluated in the argument; it appears only inside the sufficient condition. Wherever the densities are positive,

$$\frac{dQ}{dP_\theta}(x_{0:T}) = \frac{g(x_0)}{f_\theta(x_0)} \cdot \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{1:T} \mid x_0)},$$

and under the displayed sufficient condition the second factor equals 1 for g-almost every x_0; where $g(x_0) = 0$ the numerator $q(x_{0:T})$ vanishes, so those points contribute $G(-1)$ to either side alike.

The path ratio is then the function $g(x_0)/f_\theta(x_0)$ of x_0 alone, and the last step carries out the $x_{1:T}$ integral explicitly: writing $h(x_0) = G\big(g(x_0)/f_\theta(x_0) - 1\big)$,

$$D_G(Q\|P_\theta) = \int h(x_0)\, p_\theta(x_{0:T})\, dx_{0:T} = \int \Big(\int h(x_0)\, p_\theta(x_0, x_{1:T})\, dx_{1:T} \Big) dx_0$$
$$= \int h(x_0)\, f_\theta(x_0)\, dx_0 = D_G(g\|f_\theta).$$

The first equality is the definition of D_G; the second rewrites the multiple integral over the $T+1$ coordinates in iterated form, integrating over the block $x_{1:T}$ first and over x_0 second, which Tonelli's theorem justifies because the integrand is nonnegative ($G \geq 0$ under the normalization $G(0) = G'(0) = 0$, since a convex function lies above its tangent at 0); the third holds because $h(x_0)$ does not depend on $x_{1:T}$ and the inner integral is the definition of f_θ; the last is the definition of $D_G(g\|f_\theta)$. Hence, under the displayed sufficient condition (not unconditionally), $D_G(g\|f_\theta) = D_G(Q\|P_\theta)$.

The final claim follows: if θ_0 gives an exact lifted model ($P_{\theta_0} = Q$), then in particular the conditional path laws agree, so the bound is tight (and both sides vanish). Conversely, conditional path equality alone does not imply $P_\theta = Q$, since it says nothing about the x_0 marginals. □

C Proof of Theorem 2.1

Throughout, $\phi(u) := G(u-1)$ for $u > 0$, $U_m = \exp(-m + \sqrt{2m}\, Z)$ with $Z \sim N(0,1)$, and $h_G(m) = \mathbb{E}[\phi(U_m)]$. We say that G has *polynomial growth* if

$$|G(u-1)| + u\,|G'(u-1)| + u^2\,|G''(u-1)| \;\leq\; C\,(u^p + u^{-p}) \qquad \text{for all } u > 0, \tag{2}$$

for some constants C, p. KL, HD, and NED have polynomial growth: for KL, $\phi(u) = u\log u - u + 1$, $u\phi'(u) = u\log u$, $u^2\phi''(u) = u$; for HD, $\phi(u) = 2(\sqrt{u}-1)^2$, $u\phi'(u) = 2(u - \sqrt{u})$, $u^2\phi''(u) = \sqrt{u}$; for NED, $\phi(u) = e^{1-u} + u - 2$, $u\phi'(u) = u(1 - e^{1-u})$, $u^2\phi''(u) = u^2 e^{1-u}$. Since all positive and negative moments of a lognormal variable are finite, every expectation below is finite under (2). Parts (c), (e), and (f) are stated under (2).

C.1 Part (a): local expansion

Let $\delta = U_m - 1$. Since $G \in C^3$ near 0 with $G(0) = G'(0) = 0$, $G''(0) = 1$, Taylor expansion gives

$$G(\delta) = \frac{1}{2}\delta^2 + O(|\delta|^3).$$

The stated tail condition $\mathbb{E}[|G(U_m - 1)|\, \mathbf{1}\{|U_m - 1| > \eta\}] = O(m^{3/2})$ directly controls the contribution outside the Taylor neighborhood. Note that U_m is lognormal, so all its polynomial moments are finite; however, its moment generating function diverges ($\mathbb{E}[e^{cU_m}] = \infty$ for $c > 0$), so generators with exponential or faster growth require case-by-case verification of the tail condition. Taking expectation on the Taylor region,

$$h_G(m) = \frac{1}{2}\mathbb{E}[\delta^2] + O\big(\mathbb{E}|\delta|^3\big)\,.$$

By Lemma 2.1, $\log U_m \sim N(-m, 2m)$, so standard lognormal moment calculations give

$$\mathbb{E}[\delta^2] = e^{2m} - 1 = 2m + O(m^2), \qquad \mathbb{E}|\delta|^3 = O(m^{3/2}).$$

Substituting yields

$$h_G(m) = m + O(m^{3/2}).$$

We now verify the sharper $O(m^2)$ rate for KL, HD, and NED. Each generator is C^∞ near 0, so we can write $G(\delta) = \frac{1}{2}\delta^2 + \frac{1}{6}G^{(3)}(0)\delta^3 + \frac{1}{24}G^{(4)}(0)\delta^4 + R_5(\delta)$ where $|R_5(\delta)| \leq C|\delta|^5$ for $|\delta| \leq \eta$. Using the exact lognormal moments $\mathbb{E}[\delta^2] = 2m + 2m^2 + O(m^3)$, $\mathbb{E}[\delta^3] = 12m^2 + O(m^3)$, $\mathbb{E}[\delta^4] = 12m^2 + O(m^3)$, and $\mathbb{E}|\delta|^5 = O(m^{5/2})$, we obtain $h_G(m) = m + O(m^2)$ provided the tail condition $\mathbb{E}[|G(U_m - 1)|\, \mathbf{1}\{|U_m - 1| > \eta\}] = O(m^2)$ holds. We verify this via a lognormal Chernoff bound: since $U_m \to 1$ sharply as $m \to 0$, for fixed $\eta > 0$, $P(|U_m - 1| > \eta)$ decays like $\exp(-c_\eta/m)$. For the three generators: KL has $|G(U_m - 1)| \leq C\, U_m \log U_m$; HD has $|G(U_m - 1)| \leq C\, U_m$; NED has $|G(U_m - 1)| \leq C\, U_m$. In each case, the growth is at most linear-times-logarithmic in U_m, and the super-exponential decay of the tail probability dominates, so the $O(m^2)$ tail condition is satisfied. □

C.2 Part (b): global monotonicity

Since G is convex on $[-1,\infty)$, ϕ is convex on $(0,\infty)$. Also,

$$\phi(1) = G(0) = 0, \qquad \phi'(1) = G'(0) = 0.$$

First, by convexity of G,

$$G(\delta) \geq G(0) + G'(0)\delta = 0 \qquad \text{for all } \delta \geq -1.$$

Hence $h_G(m) = \mathbb{E}[G(U_m - 1)] \geq 0$. When $m = 0$, we have $U_0 = 1$ almost surely, so $h_G(0) = G(0) = 0$. Now let $(B_t)_{t\geq 0}$ be a standard Brownian motion and define

$$M_t := \exp(\sqrt{2}\, B_t - t), \qquad t \geq 0.$$

Then M_t is a positive martingale, and for each fixed $m \geq 0$, $M_m \overset{d}{=} \exp(\sqrt{2m}\, Z - m) = U_m$. Therefore, $h_G(m) = \mathbb{E}[\phi(M_m)]$. Let $0 \leq s \leq t$. Since M_t is a martingale and ϕ is convex, conditional Jensen's inequality gives

$$\mathbb{E}[\phi(M_t) \mid \mathcal{F}_s] \geq \phi\big(\mathbb{E}[M_t \mid \mathcal{F}_s]\big) = \phi(M_s).$$

Thus $(\phi(M_t))_{t\geq 0}$ is a submartingale. Taking expectations yields $\mathbb{E}[\phi(M_t)] \geq \mathbb{E}[\phi(M_s)]$, that is, $h_G(t) \geq h_G(s)$, so h_G is nondecreasing on $[0,\infty)$.

The inequality is in fact strict for $s < t$. Conditional on $\mathcal{F}_s$, $M_t = M_s \exp(\sqrt{2}(B_t - B_s) - (t-s))$ is a nondegenerate lognormal variable with full support $(0,\infty)$ and conditional mean M_s. Equality in Jensen's inequality for a convex ϕ and a nondegenerate X requires ϕ to coincide almost surely with an affine function on the support of X; here the support is all of $(0,\infty)$, and ϕ is not affine there because $\phi''(1) = G''(0) = 1 > 0$. Hence $\mathbb{E}[\phi(M_t) \mid \mathcal{F}_s] > \phi(M_s)$ almost surely, and $h_G(t) > h_G(s)$: h_G is strictly increasing on $[0,\infty)$, and in particular $h_G(m) > h_G(0) = 0$ for every $m > 0$. (When G is strictly convex the last claim also follows directly from strict Jensen, $h_G(m) = \mathbb{E}[\phi(U_m)] > \phi(\mathbb{E}[U_m]) = \phi(1) = 0$.) □

C.3 Part (c): the influence identity

Let $p_m(x) = (4\pi m)^{-1/2} \exp\big(-(x+m)^2/(4m)\big)$ be the density of $\log U_m \sim N(-m, 2m)$, so that $h_G(m) = \int_{\mathbb{R}} \psi(x)\, p_m(x)\, dx$ with $\psi(x) := \phi(e^x)$. A direct computation gives

$$\partial_x p_m = -\tfrac{x+m}{2m}\, p_m, \quad \partial_x^2 p_m = \Big(\tfrac{(x+m)^2}{4m^2} - \tfrac{1}{2m}\Big) p_m, \quad \partial_m p_m = \Big(\tfrac{(x+m)^2}{4m^2} - \tfrac{x+m}{2m} - \tfrac{1}{2m}\Big) p_m,$$

hence $\partial_m p_m = \partial_x^2 p_m + \partial_x p_m$ for $m > 0$ (the Fokker–Planck equation of $dX = -dm + \sqrt{2}\, dB$, of which $\log M_m$ in part (b) is a solution). Under (2), $|\psi(x)| \leq C(e^{px} + e^{-px})$ and likewise for $\psi'(x) = e^x \phi'(e^x)$ and $\psi''(x) = e^{2x}\phi''(e^x) + e^x \phi'(e^x)$, while $\partial_m p_m$ and $\partial_x^k p_m$ are Gaussian densities times polynomials in x with coefficients bounded on compact m-intervals of $(0,\infty)$; dominated convergence therefore permits differentiation under the integral sign, and the boundary terms in the integrations by parts below vanish. For $m > 0$,

$$\begin{aligned} h_G'(m) = \int \psi\, \partial_m p_m\, dx &= \int \psi\, (\partial_x^2 p_m + \partial_x p_m)\, dx = \int (\psi'' - \psi')\, p_m\, dx \\ &= \int e^{2x} \phi''(e^x)\, p_m(x)\, dx = \mathbb{E}\big[U_m^2\, G''(U_m - 1)\big], \end{aligned}$$

where the third equality integrates by parts twice for the first term and once for the second, and the fourth uses $\psi'' - \psi' = e^{2x}\phi''(e^x)$. □

C.4 Part (d): the α-family

For $G_\alpha''(\delta) = (1+\delta)^{\alpha-2}$ (with $G_\alpha(0) = G_\alpha'(0) = 0$, so that $G_\alpha''(0) = 1$; $\alpha = 1$ is KL, $\alpha = 1/2$ is HD), part (c) gives

$$h_\alpha'(m) = \mathbb{E}\big[U_m^2\, U_m^{\alpha-2}\big] = \mathbb{E}\big[U_m^\alpha\big] = \exp\big(-\alpha m + \alpha^2 m\big) = e^{-\alpha(1-\alpha)m},$$

using the lognormal moment formula $\mathbb{E}[U_m^a] = e^{-am + a^2 m}$ for $\log U_m \sim N(-m, 2m)$ (the growth condition (2) holds for every real α). The exponent $\kappa = \alpha(1-\alpha)$ is positive exactly when $0 < \alpha < 1$, is nonpositive for $\alpha \leq 0$ and $\alpha \geq 1$ (constant weight 1 at $\alpha \in \{0,1\}$, and a weight that grows with m otherwise), and is maximized uniquely at $\alpha = 1/2$ with $\kappa = 1/4$, giving $h_{\mathrm{HD}}'(m) = e^{-m/4}$. Integrating, $h_\alpha(m) = (1 - e^{-\kappa m})/\kappa$ and $h_\alpha(\infty) = 1/\kappa$ for $0 < \alpha < 1$. □

C.5 Part (e): the universal speed limit and the NED rate

Lower bound. By part (c), $h'_G(m) = \mathbb{E}[U_m^2 G''(U_m - 1)]$, and the integrand is nonnegative because G is convex. Since G'' is continuous at 0 with $G''(0) = 1$, there is $\eta \in (0, 1]$ such that $G''(e^x - 1) \geq 1/2$ for $|x| \leq \eta$. Restricting the expectation to $\{|\log U_m| \leq \eta\}$,

$$h'_G(m) \;\geq\; \tfrac{1}{2}\, e^{-2\eta}\; P\big(|\log U_m| \leq \eta\big) = \tfrac{1}{2}\, e^{-2\eta}\; P\Big(Z \in \Big[\tfrac{m-\eta}{\sqrt{2m}}, \tfrac{m+\eta}{\sqrt{2m}}\Big]\Big).$$

For $m \geq \eta$ the interval lies in $[0, \infty)$, where the standard normal density is decreasing, so the probability is at least the interval length $2\eta/\sqrt{2m}$ times the density at the right endpoint:

$$P\Big(Z \in \Big[\tfrac{m-\eta}{\sqrt{2m}}, \tfrac{m+\eta}{\sqrt{2m}}\Big]\Big) \;\geq\; \frac{2\eta}{\sqrt{2m}} \cdot \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(m+\eta)^2}{4m}\Big) = \frac{\eta}{\sqrt{\pi m}}\, e^{-\eta/2 - \eta^2/(4m)}\, e^{-m/4}.$$

Hence for $m \geq 1$, $h'_G(m) \geq c_G\, m^{-1/2} e^{-m/4}$ with $c_G = \frac{\eta}{2\sqrt{\pi}}\, e^{-5\eta/2 - \eta^2/4} > 0$. No induced influence weight can therefore decay faster than the order $m^{-1/2} e^{-m/4}$.

NED attains the order. Since $\mathbb{E}[U_m] = 1$, we have $h_{\mathrm{NED}}(m) = e\,\mathbb{E}[e^{-U_m}] - 1$. Differentiating under the integral sign via $\partial U_m/\partial m = U_m(-1 + Z/\sqrt{2m})$ gives

$$h'_{\mathrm{NED}}(m) = e\,\mathbb{E}\Big[e^{-U_m}\, U_m\Big(1 - \tfrac{Z}{\sqrt{2m}}\Big)\Big]$$

(equivalently, by part (c), $h'_{\mathrm{NED}}(m) = \mathbb{E}[U_m^2 e^{1-U_m}]$). **Girsanov shift.** Since $U_m = e^{-m + \sqrt{2m} Z}$ with $Z \sim N(0,1)$, the change of measure $\mathbb{E}[U_m\, f(Z)] = \mathbb{E}[f(Z + \sqrt{2m})]$ transforms the expectation to

$$h'_{\mathrm{NED}}(m) = \frac{e}{\sqrt{2m}}\, \mathbb{E}\Big[|Z|\, e^{-\exp(m - \sqrt{2m}|Z|)}\, \mathbf{1}\{Z < 0\}\Big] + o(m^{-1/2} e^{-m/4}),$$

where the $Z \geq 0$ contribution is super-exponentially small and hence negligible at the $m^{-1/2} e^{-m/4}$ scale. **Laplace analysis.** Substitute $|Z| = \sqrt{m/2} + s$ (the transition point where $\exp(m - \sqrt{2m}|Z|) = 1$):

$$h'_{\mathrm{NED}}(m) \sim \frac{e}{\sqrt{2m}} \cdot \sqrt{\tfrac{m}{2}}\, \varphi\Big(\sqrt{\tfrac{m}{2}}\Big) \int_{-\infty}^{\infty} e^{-e^{-\sqrt{2m}\, s}}\, e^{-s\sqrt{m/2}}\, ds,$$

where φ is the standard normal density. Set $u = e^{-\sqrt{2m}\, s}$, giving $e^{-s\sqrt{m/2}} = u^{1/2}$ and $ds = -du/(u\sqrt{2m})$. The integral becomes $\frac{1}{\sqrt{2m}} \int_0^\infty e^{-u} u^{-1/2}\, du = \frac{\Gamma(1/2)}{\sqrt{2m}} = \frac{\sqrt{\pi}}{\sqrt{2m}}$. Combining all factors and using $\varphi(\sqrt{m/2}) = \frac{1}{\sqrt{2\pi}} e^{-m/4}$:

$$h'_{\mathrm{NED}}(m) \sim \frac{e}{\sqrt{2m}} \cdot \sqrt{\tfrac{m}{2}} \cdot \frac{e^{-m/4}}{\sqrt{2\pi}} \cdot \frac{\sqrt{\pi}}{\sqrt{2m}} = \frac{e}{4\sqrt{m}}\, e^{-m/4}.$$

Thus NED decays at the same exponential rate $e^{-m/4}$ as HD, with the additional $m^{-1/2}$ factor of the universal limit: it attains the fastest admissible asymptotic order. □

C.6 Part (f): smoothness and the exclusion of Huber and clipped MSE

With ψ and p_m as in part (c), every m-derivative of p_m has the form $\partial_m^k p_m(x) = P_k(x, m)\, p_m(x)$ with P_k a polynomial in x whose coefficients are rational functions of m without poles on $(0, \infty)$. Under (2), $|\psi(x)|\, |P_k(x, m)|\, p_m(x)$ is dominated, locally uniformly in $m \in (0, \infty)$, by an integrable function of x, so $h_G(m) = \int \psi\, p_m\, dx$ can be differentiated under the integral sign any number of times: $h_G \in C^\infty(0, \infty)$.

The radial Huber loss with threshold $\tau > 0$, written as a function of the mismatch m with $\rho_H(m) = m$ for $m \leq \tau$ and $\rho_H(m) = 2\sqrt{\tau m} - \tau$ for $m > \tau$ (the Huber loss $r^2/2$ for $|r| \leq \delta$, $\delta|r| - \delta^2/2$ otherwise, evaluated at $|r|^2 \propto m$ and normalized to slope one at the origin), is C^1 but its second derivative jumps from 0 to $-1/(2\tau)$ at $m = \tau$; it cannot equal a C^∞ function. Clipped MSE, $\rho_C(m) = \min(m, \tau)$, is constant on $[\tau, \infty)$, whereas h_G is strictly increasing by part (b). Hence neither is h_G for any admissible G. We note that the baselines implemented in Section 3 are not functions of m_t at all: the Huber baseline is applied elementwise to the residual coordinates, and the clipped baseline clips the unweighted per-sample mean squared residual; the argument above shows that even their radial analogues lie outside the induced class. □

D Proof of Proposition 2.1 and the calibration rule

Proof of Proposition 2.1. Since w_1 is positive and $R(w_1) = \mathbb{E}_{F_{\mathrm{cl}}}[w_1] \in (0, \infty)$, the tilted probability $d\nu = w_1\, dF_{\mathrm{cl}}/R(w_1)$ is well defined. Write $a = w_2/w_1$ and let $r = dF_{\mathrm{ct}}/dF_{\mathrm{cl}}$. Then

$$R(w_2) = R(w_1)\,\mathbb{E}_\nu[a], \qquad L(w_1) = R(w_1)\,\mathbb{E}_\nu[r], \qquad L(w_2) = R(w_1)\,\mathbb{E}_\nu[ra],$$

using $L(w_i) = \int w_i\, r\, dF_{\mathrm{cl}}$. Hence $L(w_2)/R(w_2) \le L(w_1)/R(w_1)$ is equivalent to $\mathbb{E}_\nu[ra] \le \mathbb{E}_\nu[r]\,\mathbb{E}_\nu[a]$, that is, to $\mathrm{Cov}_\nu(r, a) \le 0$. Both r and a are functions of the same scalar m, r nondecreasing by hypothesis and a nonincreasing; Chebyshev's association inequality (for a nondecreasing and a nonincreasing function of one variable, the covariance under any probability measure is nonpositive) gives the claim. □

Two remarks delimit the scope. First, L/R compares the average influence weight kept on clean samples with the average weight leaked to contaminated samples; it is a proxy for relative gradient influence, not a generalization guarantee. Second, the monotone likelihood ratio premise requires the contaminated mismatch law to dominate the clean one upward; when the two laws substantially overlap the premise fails and the proposition expresses no preference between profiles.

D.1 A quantitative calibration rule

Fix a required clean retention $\rho \in (0, 1)$ and a contamination suppression target $\delta \in (0, 1)$, and read two quantiles from the per-step mismatch histogram at a reference noise level: a high clean quantile q_{cl} and a low contaminated quantile q_{ct}. The training loop computes m_t at no extra cost, and the reference level can be taken in the first third of the schedule, where the influence weights act (Appendix I). Requiring $h'_G(q_{\mathrm{cl}}) \ge \rho$ and $h'_G(q_{\mathrm{ct}}) \le \delta$ for a member of the α-family, whose weight is $e^{-\kappa m}$ with $\kappa = \alpha(1 - \alpha)$ (Theorem 2.1(d)), gives the window

$$\frac{\ln(1/\delta)}{q_{\mathrm{ct}}} \;\le\; \kappa \;\le\; \frac{\ln(1/\rho)}{q_{\mathrm{cl}}}.$$

Two readings follow. The window is nonempty exactly when $q_{\mathrm{ct}}/q_{\mathrm{cl}} \ge \ln(1/\delta)/\ln(1/\rho)$: which targets are achievable is set by the relative separation of the two mismatch laws. And $\kappa \le 1/4$ on the family (attained at Hellinger), while by Theorem 2.1(e) no induced weight decays faster than the order $m^{-1/2}e^{-m/4}$, so a target with $\ln(1/\delta)/q_{\mathrm{ct}}$ much above $1/4$ is out of reach for the entire class: the absolute scale of the contaminated mismatch, $q_{\mathrm{ct}} \ge 4\ln(1/\delta)$, is a second constraint separate from overlap.

Worked examples at $t = 99$ with $\rho = 0.5$ and $\delta = 0.05$, so that the targets need $q_{\mathrm{ct}}/q_{\mathrm{cl}} \ge \ln 20/\ln 2 \approx 4.3$ and $q_{\mathrm{ct}} \ge 4\ln 20 \approx 12$, using the per-step medians of Appendix I as the quantiles. CIFAR-10 at $\varepsilon = 0.30$ (noise replacement): the medians are 2.4 (clean) and 22.5 (contaminated); separation 9.4 and scale 22.5 both pass, so the targets are achievable and aggressive weighting is justified; the measured per-step leakage-to-retention ratios at this slice are 0.007 (HD) and 0.002 (NED). Class-conditional MNIST at $\varepsilon = 0.30$ (Appendix K): the medians are 0.6 and 8.1; the separation 13.5 passes but the scale fails ($8.1 < 12$), so 5% suppression is unreachable for the whole family at this level, which is the boundary case diagnosed there by arithmetic before training. Hellinger ($\kappa = 1/4$) saturates at $h_{\mathrm{HD}}(\infty) = 4$; NED saturates at $h_{\mathrm{NED}}(\infty) = e - 1$ and adds the $m^{-1/2}$ tail suppression, so it is the more aggressive choice, attractive when the contaminated mismatches are clearly separated and more likely to suppress useful moderate-mismatch observations otherwise.

E Proof of Proposition 2.2 and its measured magnitudes

Setting. Both path laws admit reverse-time factorizations, since the time reversal of a Markov chain is Markov:

$$Q(dx_{0:T}) = q_T(dx_T) \prod_{t=T}^{1} K_t(dx_{t-1} \mid x_t), \qquad P_\theta(dx_{0:T}) = p_T(dx_T) \prod_{t=T}^{1} P_{\theta,t}(dx_{t-1} \mid x_t),$$

where $K_t(\cdot \mid x_t) = q(x_{t-1} \in \cdot \mid x_t)$ is the true reverse kernel of the forward process. Marginalizing x_0 gives the mixture representation $K_t(\cdot \mid x_t) = \int q(\cdot \mid x_t, x_0)\, q(dx_0 \mid x_t)$ for $t \ge 2$, with

$q(\cdot \mid x_t, x_0)$ the Gaussian DDPM posterior of Lemma 2.1; at $t = 1$ the posterior $q(x_0 \mid x_1, x_0)$ is a point mass, so Lemma 2.1 does not apply there and the true kernel is $K_1(\cdot \mid x_1) = q(x_0 \in \cdot \mid x_1)$. The optimized per-step term is $\ell_{G,t} = \mathbb{E}_{q(x_0,x_t)}[D_G(q(\cdot \mid x_t, x_0) \,\|\, P_{\theta,t}(\cdot \mid x_t))] = \mathbb{E}[h_G(m_t)]$, and the decoder discrepancy is $\Delta_{G,1} = \mathbb{E}_{q_1}[D_G(K_1 \,\|\, P_{\theta,1})]$. We use the normalization identity $D_{\mathrm{HD}}(P\|R) = \mathbb{E}_R[2(\sqrt{dP/dR} - 1)^2] = 2\int(\sqrt{dP} - \sqrt{dR})^2 = 4H^2(P, R)$.

Step 1 (convexity). D_G is jointly convex in its two arguments, in particular convex in the first. Jensen's inequality applied to the mixture representation of K_t gives, for every x_t and $t \geq 2$, $D_G(K_t(\cdot \mid x_t) \,\|\, P_{\theta,t}(\cdot \mid x_t)) \leq \int D_G(q(\cdot \mid x_t, x_0) \,\|\, P_{\theta,t}(\cdot \mid x_t))\, q(dx_0 \mid x_t)$; integrating over $q_t(dx_t)$,

$$\Delta_{G,t} := \mathbb{E}_{q_t}\big[D_G(K_t \,\|\, P_{\theta,t})\big] \;\leq\; \ell_{G,t}, \qquad t \geq 2.$$

The optimized, x_0-conditioned loss therefore upper bounds the discrepancy between the model kernel and the true reverse kernel.

Step 2 (hybrid chain). For $t = 0, \ldots, T$ define the hybrid path laws $R^{(t)} := q_T\, K_T \cdots K_{t+1}\, P_{\theta,t} \cdots P_{\theta,1}$, so that $R^{(0)} = Q$ and $R^{(T)} = q_T\, P_{\theta,T} \cdots P_{\theta,1}$. Since the Hellinger distance is a metric,

$$H(Q, P_\theta) \;\leq\; \sum_{t=1}^{T} H\big(R^{(t-1)}, R^{(t)}\big) + H\big(R^{(T)}, P_\theta\big).$$

The last term contracts to $H(q_T, p_T)$: $R^{(T)}$ and P_θ share all kernels and differ only in the initial law, and H is nonincreasing under a common Markov kernel (data processing). For the swap terms, $R^{(t-1)}$ and $R^{(t)}$ share the law of $x_{t:T}$ (with x_t marginal q_t) and the downstream kernels $P_{\theta,t-1}, \ldots, P_{\theta,1}$, differing only in the kernel from x_t to x_{t-1}, K_t versus $P_{\theta,t}$. By data processing the downstream kernels can only decrease H, and for two laws $\mu \otimes K$ and $\mu \otimes K'$ with a common first marginal μ, $H^2(\mu \otimes K, \mu \otimes K') = \frac{1}{2}\int \mu(dx) \int(\sqrt{K(dy \mid x)} - \sqrt{K'(dy \mid x)})^2 = \mathbb{E}_\mu[H^2(K(\cdot \mid x), K'(\cdot \mid x))]$. Hence

$$H^2\big(R^{(t-1)}, R^{(t)}\big) \;\leq\; \mathbb{E}_{q_t}\big[H^2\big(K_t(\cdot \mid x_t), P_{\theta,t}(\cdot \mid x_t)\big)\big].$$

Step 3 (domination below). From $c_- \phi_{\mathrm{HD}} \leq \phi_G$ pointwise, $c_- D_{\mathrm{HD}} \leq D_G$ for every pair of laws, so $4c_- H^2(K_t, P_{\theta,t}) \leq D_G(K_t \| P_{\theta,t})$. Combined with Steps 1 and 2, $H^2(R^{(t-1)}, R^{(t)}) \leq \ell_{G,t}/(4c_-)$ for $t \geq 2$ and $H^2(R^{(0)}, R^{(1)}) \leq \Delta_{G,1}/(4c_-)$. Therefore

$$H(Q, P_\theta) \;\leq\; H(q_T, p_T) + \frac{1}{2\sqrt{c_-}}\Big(\sqrt{\Delta_{G,1}} + \sum_{t=2}^{T} \sqrt{\ell_{G,t}}\Big).$$

Step 4 (domination above). From $\phi_G \leq c_+ \phi_{\mathrm{HD}}$, $D_G(Q\|P_\theta) \leq c_+ D_{\mathrm{HD}}(Q\|P_\theta) = 4c_+ H^2(Q, P_\theta)$. Substituting Step 3 gives the stated bound on $D_G(Q\|P_\theta)$, and Theorem 2.2 (data processing under the projection $x_{0:T} \mapsto x_0$) transfers the same right-hand side to $D_G(g\|f_\theta)$. □

Constants. For HD, $c_- = c_+ = 1$. For NED, $\phi_{\mathrm{NED}}(u) = e^{1-u} + u - 2$ and $\phi_{\mathrm{HD}}(u) = 2(\sqrt{u} - 1)^2$; the ratio $\phi_{\mathrm{NED}}/\phi_{\mathrm{HD}}$ extends continuously to $(0, \infty)$ with limits $(e-2)/2 \approx 0.3591$ at $u \to 0$, 1 at $u = 1$ (both generators have $\phi''(1) = 1$), and $1/2$ at $u \to \infty$; numerically its infimum is the limit $(e-2)/2$ at $u \to 0$ and its supremum is 1.0731 at $u \approx 2.18$, so $c_- = (e-2)/2$ and $c_+ = 1.074$ are valid. KL is excluded because $\phi_{\mathrm{KL}}(u) = u \log u - u + 1$ grows like $u \log u$, so no finite c_+ exists; KL needs nothing here, having the exact ELBO identity of Section 2.1. If $q_T = p_T$, Cauchy–Schwarz gives the weaker corollary $D_G(Q\|P_\theta) \leq (c_+/c_-)\,(\sum_t \sqrt{\ell_{G,t}})^2 \leq (c_+/c_-)\, T\, \mathcal{L}_G(\theta)$; the per-step form of the proposition is sharper.

The terminal term. On the linear schedule of Section 3, $\bar{\alpha}_T = 4.0 \times 10^{-5}$, so q_T is the mixture over x_0 of $\mathcal{N}(\sqrt{\bar{\alpha}_T}\, x_0, (1 - \bar{\alpha}_T) I)$ and $p_T = \mathcal{N}(0, I)$. By joint convexity of H^2, $H^2(q_T, p_T) \leq \mathbb{E}_{x_0}[H^2(\mathcal{N}(\sqrt{\bar{\alpha}_T} x_0, (1 - \bar{\alpha}_T) I), \mathcal{N}(0, I))]$, and for isotropic Gaussians in dimension d with variances σ_1^2, σ_2^2, $H^2 = 1 - \big(2\sigma_1\sigma_2/(\sigma_1^2 + \sigma_2^2)\big)^{d/2} \exp\big(-\|\mu_1 - \mu_2\|^2/(4(\sigma_1^2 + \sigma_2^2))\big)$. With $d = 3072$ and $\|x_0\|^2 \leq d$ for pixels in $[-1, 1]$, this gives $H^2(q_T, p_T) \leq 0.016$, that is, $H(q_T, p_T) \leq 0.13$.

Two remarks. (i) Steps 2 to 4 use only $\Delta_{G,t}$, the divergence between the model kernel and the true reverse kernel; Step 1 merely certifies $\Delta_{G,t} \le \ell_{G,t}$ with the computable loss. The proposition therefore holds verbatim with $\Delta_{G,t}$ in place of $\ell_{G,t}$, strictly tighter, and $\ell_{G,t}$ carries an irreducible floor even for a perfect model because the x_0-conditioned posterior is not the marginal reverse kernel. For KL this is exact: since $\mathbb{E}_{q(x_0|x_t)}\, q(x_{t-1} \mid x_t, x_0) = q(x_{t-1} \mid x_t)$, the compensation identity gives $\ell_{\mathrm{KL},t} = \Delta_{\mathrm{KL},t} + \mathbb{E}\,\mathrm{KL}\big(q(x_{t-1} \mid x_t, x_0) \,\|\, q(x_{t-1} \mid x_t)\big) = \Delta_{\mathrm{KL},t} + I(x_0; x_{t-1} \mid x_t)$, so part of a large training loss is this conditional mutual information rather than model error. (ii) Restricting the hybrid chain to $t \ge 2$ (the $t = 1$ swap is never performed) bounds the divergence between the laws q_1 and $p_{\theta,1}$ of x_1, the σ_1-smoothed data and model laws ($\beta_1 = 10^{-4}$, so $\sigma_1 \approx 0.01$ on the $[-1, 1]$ pixel scale): $D_G(q_1 \| p_{\theta,1}) \le 4c_+\big[H(q_T, p_T) + (2\sqrt{c_-})^{-1} \sum_{t=2}^{T} \sqrt{\ell_{G,t}}\big]^2$, in which every term is measurable from training or analytic. This is the early-stopping convention of the diffusion convergence literature; the last step is a decoder question, handled by early stopping or decoder smoothness.

Measured magnitudes. Table 3 evaluates the right-hand side on three 100K-step checkpoints of the audit runs of Appendix H (CIFAR-10, noise replacement), computing $\ell_{G,t}$ on 128 training images for every $t = 1, \dots, 1000$ under the model's own training distribution, with the terminal term omitted ($H(q_T, p_T) \le 0.13$) and the $t = 1$ term evaluated as the implemented $h_G(m_1)$ for illustration. Against the ceilings $D_{\mathrm{HD}} \le 4$ and $D_{\mathrm{NED}} \le e - 1$ the right-hand sides are five orders of magnitude too large to be informative, and better constants would not change this: since $(\sum_t \sqrt{\ell_{G,t}})^2 \ge \sum_t \ell_{G,t} = \mathcal{L}_G(\theta)$, the right-hand side always exceeds the training loss, which is itself far above the ceilings. The bound becomes informative once $\sum_t \sqrt{\ell_{G,t}}$ is of order one (for HD, once it is below $2(1 - H(q_T, p_T))$). Its content at the current training scale is the direction of control at finite T, not the size of the gap.

Table 3: Proposition 2.2 evaluated on trained checkpoints (CIFAR-10, noise replacement, 100K steps, 128 training images, $T = 1000$). $\mathcal{L}_G = \sum_t \ell_{G,t}$; RHS $= 4c_+[(2\sqrt{c_-})^{-1} \sum_t \sqrt{\ell_{G,t}}]^2$ with $(c_-, c_+) = (1, 1)$ for HD and $(0.359, 1.074)$ for NED; ceilings $D_{\mathrm{HD}} \le 4$, $D_{\mathrm{NED}} \le e - 1 \approx 1.72$.

Checkpoint	$\mathcal{L}_{\mathrm{HD}}$	$\sum_t \sqrt{\ell_{\mathrm{HD},t}}$	RHS (HD)	$\mathcal{L}_{\mathrm{NED}}$	$\sum_t \sqrt{\ell_{\mathrm{NED},t}}$	RHS (NED)
KL, $\varepsilon = 0.30$	830	698	4.9×10^5	456	546	8.9×10^5
KL, clean	546	529	2.8×10^5	327	443	5.9×10^5
NED, clean	549	527	2.8×10^5	328	441	5.8×10^5

F Proof of Theorem 2.3

Conditions. With $R_T = \prod_{t=1}^{T}(1 + \delta_t)$, $S_T = \sum_{t=1}^{T} \delta_t$ and $B_T = R_T - 1 - S_T$ as in the theorem, the moment conditions are: *(A1)* $\mathbb{E}[S_T^2] = O(a_T)$; *(A2)* $\mathbb{E}[S_T^2] - \sum_t \mathbb{E}[\delta_t^2] = o(a_T)$; *(A3)* $\mathbb{E}|S_T|^3 = o(a_T)$; *(A4)* $\mathbb{E}|S_T B_T| = o(a_T)$ and $\mathbb{E}[B_T^2] = o(a_T)$; *(A5)* $\mathbb{E}|B_T|^3 = o(a_T)$; *(A6)* $\sum_t \mathbb{E}|\delta_t|^3 = o(a_T)$; and the tail integrability conditions $\mathbb{E}[|G(S_T+B_T)|\,\mathbf{1}\{|S_T+B_T| > \eta\}] = o(a_T)$ and $\sum_t \mathbb{E}[|G(\delta_t)|\,\mathbf{1}\{|\delta_t| > \eta\}] = o(a_T)$ for any $\eta > 0$. In the proof, $\Phi_T := G(R_T - 1) = G(S_T + B_T)$ and $\Psi_T := \sum_t G(\delta_t)$.

Proof. Since $G \in C^3$ near 0 with $G(0) = G'(0) = 0$ and $G''(0) = 1$, there exist constants $\eta > 0$, $C > 0$ such that for all $|x| \le \eta$,

$$G(x) = \tfrac{1}{2}x^2 + \mathcal{R}(x), \qquad |\mathcal{R}(x)| \le C|x|^3.$$

Because $a_T \to 0$, assumptions (A1), (A3), (A4), and (A5) imply

$$S_T \to 0 \quad \text{in } L^2, \qquad B_T \to 0 \quad \text{in } L^2,$$

hence

$$S_T + B_T \to 0 \quad \text{in probability.}$$

The tail integrability assumption ensures that the contribution from the event $\{|S_T + B_T| > \eta\}$ is $o(a_T)$. On the complementary event, the Taylor expansion applies. By (A3) and (A5),

$$\mathbb{E}|S_T + B_T|^3 \le C\big(\mathbb{E}|S_T|^3 + \mathbb{E}|B_T|^3\big) = o(a_T).$$

The quadratic term on the tail is also controlled: $\mathbb{E}[(S_T + B_T)^2 \, \mathbf{1}\{|S_T + B_T| > \eta\}] \le \eta^{-1}\mathbb{E}|S_T + B_T|^3 = o(a_T)$. Hence

$$\mathbb{E}[\Phi_T] = \tfrac{1}{2}\mathbb{E}[(S_T + B_T)^2] + o(a_T).$$

Expanding the square gives

$$\mathbb{E}[(S_T + B_T)^2] = \mathbb{E}[S_T^2] + 2\mathbb{E}[S_T B_T] + \mathbb{E}[B_T^2] = \mathbb{E}[S_T^2] + o(a_T)$$

by (A4). Hence

$$\mathbb{E}[\Phi_T] = \tfrac{1}{2}\mathbb{E}[S_T^2] + o(a_T).$$

On the other hand, for each t,

$$G(\delta_t) = \tfrac{1}{2}\delta_t^2 + \mathcal{R}(\delta_t), \qquad |\mathcal{R}(\delta_t)| \le C|\delta_t|^3$$

for $|\delta_t|$ sufficiently small. Similarly, $\sum_t \mathbb{E}[\delta_t^2 \, \mathbf{1}\{|\delta_t| > \eta\}] \le \eta^{-1} \sum_t \mathbb{E}|\delta_t|^3 = o(a_T)$ by (A6), and the per-step tail integrability assumption controls the G contribution on the tails. Summing yields

$$\mathbb{E}[\Psi_T] = \tfrac{1}{2} \sum_{t=1}^{T} \mathbb{E}[\delta_t^2] + o(a_T)$$

by (A6). Finally, assumption (A2) gives

$$\mathbb{E}[S_T^2] = \sum_{t=1}^{T} \mathbb{E}[\delta_t^2] + o(a_T),$$

so

$$\mathbb{E}[\Phi_T] = \mathbb{E}[\Psi_T] + o(a_T).$$

□

Interpretation of conditions (A1)–(A6). Conditions (A1)–(A6) are abstract moment conditions on the increments δ_t and their partial sums. In the context of diffusion models, they admit a natural interpretation as Lindeberg-type regularity conditions ensuring that the product likelihood ratio $R_T = \prod_t (1 + \delta_t)$ is well approximated by its linearization $1 + S_T$.

Conditions (A1) and (A2) together require that $\mathbb{E}[S_T^2]$ is $O(a_T)$ and that the variance of S_T is dominated by the sum of individual variances $\sum_t \mathbb{E}[\delta_t^2]$, up to $o(a_T)$ error. In the diffusion setting, (A2) holds exactly when the increments are pairwise uncorrelated, which is precisely what the orthogonality assumption (D2) provides.

Condition (A3) controls the third moment of the sum S_T, ensuring that S_T does not have heavy tails. Together with (A1), this is a standard Lindeberg-type condition: it guarantees that no single increment dominates the sum, which is supported by the moment conditions in (D1).

Conditions (A4) and (A5) bound the cross-interaction and higher moments of the product remainder $B_T = R_T - 1 - S_T$. Since B_T captures the nonlinear terms in the product $\prod_t (1 + \delta_t)$, these conditions ensure that the product-to-sum approximation $R_T \approx 1 + S_T$ is valid in L^2 and L^3. When the per-step mismatches are uniformly small (D1), the remainder B_T is of order $O(\sum_{s<t} \delta_s \delta_t)$, and its moments are controlled by those of the individual increments.

Finally, (A6) requires $\sum_t \mathbb{E}|\delta_t|^3 = o(a_T)$, which ensures that the per-step Taylor remainders $\mathcal{R}(\delta_t)$ are negligible in aggregate. Under (D1), the lognormal structure of Lemma 2.1 gives $\mathbb{E}[|\delta_t|^3 \mid x_0, x_t] = O(m_t^{3/2})$, so $\sum_t \mathbb{E}|\delta_t|^3 = O(\sum_t \mathbb{E}[m_t^{3/2}]) = o(a_T)$ by the moment condition in (D1).

In summary, (A1)–(A3) and (A6) are Lindeberg-type conditions on the sum S_T, while (A4)–(A5) control the product remainder B_T. In the diffusion specialization, some of these conditions follow from the lognormal moment bounds (specifically (A1), (A2), and (A6) from (D1) and (D2)), while the third-moment condition (A3) and the product-remainder conditions (A4)–(A5) are imposed explicitly via (D3) and (D4) in Corollary 2.1.

G Proof of Corollary 2.1

Conditions. In the setting of the corollary ($\delta_t = r_t - 1$, $a_T = \sum_{t=1}^{T} \mathbb{E}[m_t]$, expectations $\mathbb{E}_{\mathrm{loc}}$ under the auxiliary construction): *(D1)* $a_T \to 0$, $\max_t m_t \le c_T \to 0$ almost surely, $\sum_t \mathbb{E}[m_t^2] = o(a_T)$, and $\sum_t \mathbb{E}[m_t^{3/2}] = o(a_T)$; *(D2)* the increments $\{\delta_t\}$ are pairwise uncorrelated: $\mathbb{E}_{\mathrm{loc}}[\delta_s \delta_t] = 0$ for $s \neq t$; *(D3)* $\mathbb{E}_{\mathrm{loc}}|S_T|^3 = o(a_T)$; *(D4)* with R_T, S_T, B_T as in Theorem 2.3, $\mathbb{E}_{\mathrm{loc}}|S_T B_T| = o(a_T)$, $\mathbb{E}_{\mathrm{loc}}[B_T^2] = o(a_T)$, $\mathbb{E}_{\mathrm{loc}}|B_T|^3 = o(a_T)$; together with the tail integrability conditions of Theorem 2.3. Condition (D1) is the natural local regime in which the per-step mismatches are uniformly small ($\max_t m_t \to 0$) and spread across many steps, the expected setting for a well-trained diffusion model. Condition (D2) is a pairwise uncorrelatedness assumption on the likelihood-ratio increments, used to eliminate cross terms in S_T^2; it is an approximation whose quality depends on the model architecture and training stage, and is most natural when each δ_t is small. Even when (D2) is violated, the local objective (1) remains well-defined and can be used for training; (D2) is only needed for the asymptotic justification. Condition (D3) controls the third moment of S_T, ensuring no single increment dominates, and (D4) controls the higher-order product remainder $B_T = R_T - 1 - S_T$; both are supported by the lognormal moment bounds from (D1) but are imposed explicitly. Appendix H measures the quantities constrained by (D1) on trained models.

Proof. All expectations below are under $\mathbb{E}_{\mathrm{loc}}$ (the auxiliary local-product construction defined in Corollary 2.1). By Lemma 2.1, conditional on (x_0, x_t), $\log r_t \sim N(-m_t, 2m_t)$ under the local model kernel $p_{\theta,t}(\cdot \mid x_t)$. Defining $h_G(m) := \mathbb{E}[G(U_m - 1)]$ with $\log U_m \sim N(-m, 2m)$, the tower property gives $\mathbb{E}_{\mathrm{loc}}[G(\delta_t)] = \mathbb{E}[h_G(m_t)]$, and hence

$$\mathbb{E}_{\mathrm{loc}}[\Psi_T] = \sum_{t=1}^{T} \mathbb{E}[h_G(m_t)].$$

Standard lognormal moment calculations give

$$\mathbb{E}_{\mathrm{loc}}[\delta_t^2 \mid x_0, x_t] = e^{2m_t} - 1 = 2m_t + O(m_t^2).$$

By the uniform-smallness condition $\max_t m_t \le c_T \to 0$ in (D1), the local expansion $\mathbb{E}_{\mathrm{loc}}[|\delta_t|^3 \mid x_0, x_t] = O(m_t^{3/2})$ holds uniformly over all steps (the $O(\cdot)$ constant depends only on c_T, which is bounded). Taking expectations gives

$$\sum_{t=1}^{T} \mathbb{E}_{\mathrm{loc}}[\delta_t^2] = 2a_T + O\left(\sum_{t=1}^{T} \mathbb{E}[m_t^2]\right) = 2a_T + o(a_T) = O(a_T),$$

by the moment condition $\sum_t \mathbb{E}[m_t^2] = o(a_T)$ in (D1), and

$$\sum_{t=1}^{T} \mathbb{E}_{\mathrm{loc}}|\delta_t|^3 = O\left(\sum_{t=1}^{T} \mathbb{E}[m_t^{3/2}]\right) = o(a_T),$$

by the moment condition $\sum_t \mathbb{E}[m_t^{3/2}] = o(a_T)$ in (D1).

Next, assumption (D2) directly gives $\mathbb{E}_{\mathrm{loc}}[\delta_s \delta_t] = 0$ for $s \neq t$. Hence

$$\mathbb{E}_{\mathrm{loc}}[S_T^2] = \sum_{t=1}^{T} \mathbb{E}_{\mathrm{loc}}[\delta_t^2].$$

The argument above verifies (A1), (A2), and (A6) of Theorem 2.3 from (D1) and (D2). The third-moment condition (A3), $\mathbb{E}_{\mathrm{loc}}|S_T|^3 = o(a_T)$, is imposed directly as assumption (D3). The tail integrability condition is assumed explicitly in Corollary 2.1; per-step tails are natural for KL, HD, and NED because their generators have at most linear-times-logarithmic growth in U_m (see Appendix C.1), and the super-exponential decay of $P(|U_m - 1| > \eta)$ as $m \to 0$ ensures the tail conditions hold. The product-remainder conditions (A4)–(A5) are imposed as (D4). Applying Theorem 2.3 gives

$$\mathbb{E}_{\mathrm{loc}}[G(R_T - 1)] = \sum_{t=1}^{T} \mathbb{E}_{\mathrm{loc}}[G(\delta_t)] + o(a_T).$$

Since $\mathbb{E}_{\mathrm{loc}}[G(\delta_t)] = \mathbb{E}[h_G(m_t)]$, the conclusion $\mathbb{E}_{\mathrm{loc}}[G(R_T - 1)] = \sum_t \mathbb{E}[h_G(m_t)] + o(a_T)$ follows. □

H Audit of the small-mismatch regime on trained models

The conditions of Theorem 2.3 and Corollary 2.1 are asymptotic and cannot be verified at a fixed T; what can be measured is how far trained models are from the regime they describe. We instrumented three CIFAR-10 runs under the protocol of Section 3 (noise replacement; KL at $\varepsilon = 0.30$, KL on clean data, NED on clean data; 100K steps) and, at the final checkpoint, evaluated the per-step mismatch m_t on the full grid $t = 1, \dots, 1000$ for batches of training images, separately for the clean and the contaminated population. Table 4 reports the quantities constrained by condition (D1): $a_T = \sum_t \mathbb{E}[m_t]$ (for KL this is exactly the training objective, since $h_{\mathrm{KL}}(m) = m$), $\max_t m_t$, the ratios $\sum_t \mathbb{E}[m_t^2]/a_T$ and $\sum_t \mathbb{E}[m_t^{3/2}]/a_T$ that (D1) requires to be $o(1)$, and percentiles of the per-image mean mismatch $T^{-1}\sum_t m_t$. On the contaminated run the contaminated samples have median per-image mean mismatch 35.2 against 20.2 for the clean samples. Re-mapping the same trained models to coarser grids gives $a_T = 2.58/2.20/2.01 \times 10^4$ (KL, $\varepsilon = 0.30$), $2.55/2.17/1.99 \times 10^4$ (KL, clean) and $3.24/2.75/2.51 \times 10^4$ (NED, clean) at $T = 250/500/1000$, so a fourfold change in T moves a_T by less than a factor 1.3; we did not retrain at other T, and a genuine T-trend would require retraining under matched continuous-time schedules, since increasing T alone changes the discretization, the schedule weights and the optimization problem. The trained models are therefore far from $a_T \to 0$, and the asymptotic reduction is a local interpretation of the surrogate rather than a description of the trained models; Proposition 2.2 is the statement that applies at finite T.

Table 4: Regime diagnostics at the 100K-step checkpoints (CIFAR-10, noise replacement; clean population unless stated). Percentiles refer to the per-image mean mismatch $T^{-1}\sum_t m_t$.

	KL, $\varepsilon = 0.30$	KL, clean	NED, clean
$a_T = \sum_t \mathbb{E}[m_t]$ (clean; contaminated)	2.01×10^4; 3.52×10^4	1.99×10^4	2.51×10^4
$\max_t m_t$	1642	1646	2051
$\sum_t \mathbb{E}[m_t^2]/a_T$ (required $o(1)$)	931	920	1176
$\sum_t \mathbb{E}[m_t^{3/2}]/a_T$ (required $o(1)$)	29.6	29.5	33.5
per-image mean mismatch, p50/p90/p95/p99	20.2/25.5/27.0/30.7	19.9/25.4/26.9/30.1	25.3/30.7/32.3/35.8

I Leakage and retention audit

Proposition 2.1 is stated in terms of the retention $R_G = \mathbb{E}_{\mathrm{cl}}[h'_G(m_t)]$ and leakage $L_G = \mathbb{E}_{\mathrm{ct}}[h'_G(m_t)]$, expectations of the influence weight over the clean and contaminated populations with t uniform as in training. Because h'_G is nonlinear, $h'_G(\mathrm{med}(m_t)) \neq \mathbb{E}[h'_G(m_t)]$ in general, so plug-in values at median mismatches (such as the $e^{-3.75} \approx 0.02$ quoted in Section 2.3) illustrate scale but are not the statistic of the proposition. Table 5 reports the statistic itself on the KL run at $\varepsilon = 0.30$ of Appendix H (512 images per population, m_t evaluated on the training grid), both at the 100K-step checkpoint and at the 10K-step checkpoint used as a warmup model by the selection rule, together with the per-slice values $L_G(t)/R_G(t)$ at fixed t and the clean and contaminated medians of m_t at those slices. The predicted ordering $L_{\mathrm{NED}}/R_{\mathrm{NED}} \leq L_{\mathrm{HD}}/R_{\mathrm{HD}}$ holds in aggregate and at every slice, and the 10K-step checkpoint matches the ratios of the trained model at every slice to within 0.01, which is what makes a short warmup model sufficient for the label-free rule.

The aggregate ratios are far above the median plug-in value because the uniform-t average is dominated by high-noise steps, where both populations sit near $m_t = 0$ and every weight is near 1; there the contaminated target is itself close to pure noise, so equal weighting does no harm. Suppression acts where contamination is informative: at $t \leq 199$ contaminated samples keep at most 13% of the clean weight under HD and 5% under NED, falling below 10^{-4} at $t = 49$. This is the basis of the label-free rule of Section 2.3: train a short MSE warmup, log the per-sample m_t at a few noise levels in the first third of the schedule, and read the pooled histogram. The diagnostic needs only the pooled shape: under structured out-of-distribution contamination, where the per-sample loss-rank filter of Appendix J runs at chance, the histogram shows no separated far tail; there the proposition expresses no preference, and either robust member can be used (Table 2, middle).

Table 5: Per-step leakage and retention on the CIFAR-10 KL run at $\varepsilon = 0.30$ (noise replacement). Top: aggregate R_G, L_G and L_G/R_G with t uniform. Bottom: medians of m_t (clean / contaminated) and $L_G(t)/R_G(t)$ at fixed t, for the 100K-step and the 10K-step checkpoints.

	100K steps			10K steps		
G	R_G	L_G	L_G/R_G	R_G	L_G	L_G/R_G
HD	0.891	0.663	0.744	0.885	0.662	0.749
NED	0.776	0.537	0.692	0.766	0.536	0.699

	median m_t		100K steps		10K steps	
t	clean	contam.	$L_{\mathrm{HD}}/R_{\mathrm{HD}}$	$L_{\mathrm{NED}}/R_{\mathrm{NED}}$	$L_{\mathrm{HD}}/R_{\mathrm{HD}}$	$L_{\mathrm{NED}}/R_{\mathrm{NED}}$
49	7.46	45.8	6×10^{-5}	2×10^{-5}	7×10^{-5}	2×10^{-5}
99	2.37	22.5	0.007	0.002	0.007	0.003
149	1.20	13.7	0.04	0.01	0.05	0.02
199	0.75	8.92	0.13	0.05	0.13	0.05
299	0.37	3.92	0.41	0.17	0.41	0.18
399	0.22	1.73	0.68	0.37	0.69	0.38
499	0.13	0.72	0.86	0.60	0.86	0.61
599	0.07	0.26	0.95	0.82	0.95	0.82
799	0.01	0.02	1.00	0.99	1.00	0.99
999	0.00	0.00	1.00	1.00	1.00	1.00

J Extended experimental results

All experiments use the same 35M-parameter DDPM U-Net architecture described in Section 3, trained for 100K steps on a single NVIDIA RTX 5090 GPU ($\sim$3h per run). FID is computed with 5K generated samples against the clean test split via a 50-step DDIM sampler. Training uses Adam at $\mathrm{lr} = 2{\times}10^{-4}$, batch size 128, bfloat16 mixed-precision, and gradient clipping at norm 1.0.

Contamination protocols. In the noise-replacement regime a fraction ε of the training images, chosen uniformly at random once per run, is replaced by Gaussian noise images: each pixel is drawn independently from $\mathcal{N}(0,1)$ on the $[-1,1]$ intensity scale and clipped to $[-1,1]$. In the structured out-of-distribution regime the same fraction is replaced by images drawn uniformly from the CIFAR-100 training set. All contaminated images keep their position in the training set; no label information is used by the unconditional models.

Warmup. The HD runs use a 5,000-step plain-MSE warmup before switching to h_{HD}, as stated in the released scripts, a precaution against early gradient suppression when all mismatches are large; KL, NED, Huber and clipped MSE use no warmup. Table 6 reports a separate single-seed sweep over the warmup length on clean CIFAR-10.

Table 6: HD warmup ablation on clean CIFAR-10 (single seed, 100K steps).

MSE warmup steps	0	2,500	5,000	10,000
FID	26.3	25.3	27.0	27.8

Cosine schedule replicate. To test sensitivity to the noise schedule, we trained KL and NED at $\varepsilon = 0.30$ (noise replacement, seed 42, 100K steps, no warmup, all other settings unchanged) under the cosine schedule of Nichol and Dhariwal [2021], with w_t prescribed by that schedule as in Section 2.2 and no other change. Training required no retuning and the losses converged normally. Sampling required one standard adjustment: under the cosine schedule the terminal $\bar{\alpha}_T \approx 2.4{\times}10^{-9}$ (against 4.0×10^{-5} for the linear schedule), so the uniform 50-step DDIM grid used for the linear runs starts in a degenerate region where the first reduced step amplifies the ϵ-prediction error by a factor of about 660, and without clipping the predicted x_0 the sampler returns saturated noise (FID above 390 for both objectives) despite converged training losses. We therefore evaluate with a schedule-safe version of the same sampler that clips the predicted x_0 to $[-1,1]$, as in improved DDPM [Nichol and Dhariwal, 2021], and starts the 50-step grid at the largest t with $\bar{\alpha}_t \geq 4 \times 10^{-5}$,

the linear terminal value ($t = 994$ under the cosine schedule; the grid is unchanged for linear runs). Applied to a linear-schedule KL run at $\varepsilon = 0.30$, the schedule-safe sampler reproduces that run's original-sampler FID within sampling noise (88.1 against 85.6 at 5,000 samples), so the two samplers are comparable. Table 7 reports the result: under the cosine schedule KL scores 86.2 and NED 66.1, so the KL-to-NED ordering holds and the gap (20 FID) is of the same size as under the linear schedule (15.5 in the three-seed mean of Table 2). Absolute values are not compared across schedules, since single linear-schedule runs at this seed and setting range from 66.3 (Table 8) to 85.6 (the instrumented control run).

Table 7: Cosine-schedule replicate on CIFAR-10 at $\varepsilon = 0.30$ (noise replacement, seed 42, 100K steps, 5,000-sample FID with the schedule-safe sampler).

Schedule	KL	NED
cosine	86.2	**66.1**

CIFAR-10 multi-seed protocol and per-seed values. Three divergences (KL, HD, NED) $\times$ four contamination levels ($\varepsilon \in \{0.00, 0.10, 0.20, 0.30\}$) $\times$ three seeds (42, 123, 456) = 36 runs ($\sim$108 GPU-hours) give Table 2 (main text, left). Figure 3 plots the means and standard deviations, and Figure 4 the per-seed spread. Table 8 lists the per-seed FID at $\varepsilon = 0.30$ together with the additional seeds run for the robust baselines. Per-seed variability is large for every objective; NED's advantage is clearest in the mean and at the worst seed (92.4 against 124.8 for KL). The robust baselines were run at two seeds at $\varepsilon = 0.30$, and clipped MSE at seed 123 (72.5) sits below the NED mean, so the three-seed KL-to-NED comparison is the primary evidence.

Table 8: Per-seed CIFAR-10 FID under noise replacement (100K steps). Left: $\varepsilon = 0.30$. Right: clean data ($\varepsilon = 0$) for the robust baselines, whose main-text values (Table 2, right) are single seed.

Seed	KL	HD	NED	Huber	Clipped MSE
42	66.3	56.2	55.1	94.0	90.5
123	88.1	97.1	85.0	92.4	72.5
456	124.8	122.3	92.4	–	–
mean	93.0	91.9	77.5	93.2 (2 seeds)	81.5 (2 seeds)

Seed ($\varepsilon = 0$)	Huber	Clipped MSE
42	27.8	27.7
123	33.4	27.1
456	44.5	47.3

Structured contamination: per-seed values. Table 9 lists the per-seed FID behind Table 2 (middle): three divergences $\times$ three levels $\times$ seeds 42, 123 and 456 (27 runs), with the protocol of the noise-replacement sweep and CIFAR-100 images injected in place of noise. Both robust objectives are below KL in the mean and at the worst seed at every level; the seed-level exceptions are small (NED is above KL on seed 42 at all three levels, by at most 1.7 FID, and HD on seed 456 at $\varepsilon = 0.10$, by 1.0). HD and NED are not separated: HD has the lower mean at $\varepsilon \in \{0.10, 0.20\}$ and NED at $\varepsilon = 0.30$, the differences (0.7, 5.4 and 3.4) are comparable to the seed-to-seed standard deviations (3.2 to 10.5), and the lower of the two changes with the seed (HD on every run of seed 42, NED on four of the six runs of seeds 123 and 456).

Data-cleaning baselines. Contamination may also be handled by preprocessing rather than by the training objective. Table 10 compares two baselines at $\varepsilon = 0.30$ (seed 42) under the paper's FID protocol: oracle removal, which trains KL on the 70% of samples known to be clean from the true contamination indicators (no practical method has this), and a practical loss-rank filter that ranks all training samples by their mean m_t under a model warmed up for 10K steps on the contaminated set and drops the top 30% (it is given the true ε, which favors the baseline). The results form a ladder: KL on the contaminated set 66.3; HD 56.2 and NED 55.1; filter then KL 47.9; oracle removal then KL 33.7; filter then NED 32.7; KL on fully clean data 26.8. Under noise replacement the filter's precision and recall are both 1.00, and in this best case it is strong, removing roughly half of the

Table 9: Per-seed CIFAR-10 FID under structured out-of-distribution contamination (CIFAR-100 images injected, 100K steps). The means and standard deviations are those of Table 2 (middle).

	$\varepsilon = 0.10$			$\varepsilon = 0.20$			$\varepsilon = 0.30$		
Seed	KL	HD	NED	KL	HD	NED	KL	HD	NED
42	32.2	27.2	33.9	36.6	30.9	36.8	38.9	37.4	40.2
123	51.9	31.9	31.0	42.0	30.4	32.1	58.8	40.2	34.0
456	46.4	47.4	43.8	53.2	39.7	48.5	50.7	45.4	38.5
mean	43.5	35.5	36.2	43.9	33.7	39.1	49.5	41.0	37.6
std	10.1	10.5	6.7	8.5	5.2	8.4	10.0	4.1	3.2

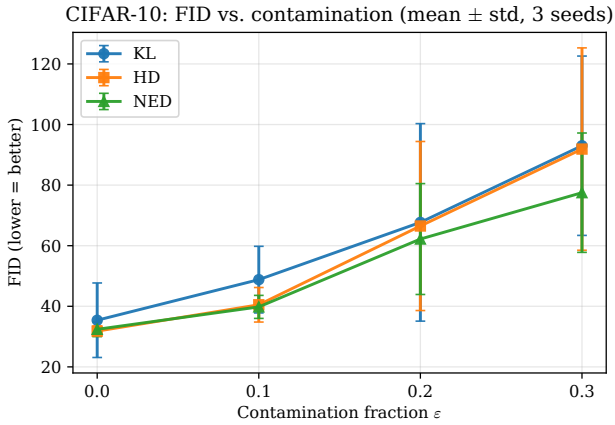

Figure 3: CIFAR-10 FID vs. contamination level under noise replacement (mean $\pm$ std over 3 seeds). NED degrades least and maintains lower variance than KL, especially under contamination. All divergences are comparable on clean data, consistent with the first-order equivalence (Theorem 2.1(a)).

contamination penalty, at the cost of a warmup model, a ranking pass, a second training run, and knowledge of ε, none of which the one-line objective change needs; filtering followed by NED matches the oracle, so cleaning and the robust objectives compose rather than compete. Filtering presupposes a detectability margin, which the structured regime removes: there the same procedure runs at precision and recall 0.31 (chance is 0.30), filtering discards mostly clean images, and filtering then KL yields FID 77.9 against 38.9 for KL alone, while HD and NED (Table 2, middle) have no such failure mode.

Table 10: Data-cleaning baselines on CIFAR-10 at $\varepsilon = 0.30$ (seed 42, 100K steps).

Regime	Training set and objective	FID
noise replacement	contaminated set, KL	66.3
noise replacement	contaminated set, HD / NED	56.2 / 55.1
noise replacement	loss-rank filter (precision, recall 1.00), then KL	47.9
noise replacement	oracle removal, then KL	33.7
noise replacement	loss-rank filter, then NED	32.7
noise replacement	fully clean data, KL	26.8
structured OOD	contaminated set, KL	38.9
structured OOD	loss-rank filter (precision, recall 0.31), then KL	77.9
structured OOD	contaminated set, HD / NED	37.4 / 40.2

CIFAR-100 results. Table 11 reports single-seed CIFAR-100 FID scores under noise replacement. The pattern is consistent with CIFAR-10: NED achieves the best FID under heavy contamination (75.3 vs. 98.0 for KL), while all divergences are comparable on clean data.

Table 11: CIFAR-100 FID (noise replacement, single seed, 100K steps).

ε	KL	HD	NED
0.00	**35.0**	35.8	35.6
0.10	46.4	48.0	**44.7**
0.20	64.3	**53.9**	56.9
0.30	98.0	99.4	**75.3**

NED quadrature ablation. Since $h'_{\mathrm{NED}}(m)$ has no closed form, it is approximated by Gauss–Legendre quadrature over the lognormal variable U_m (Lemma 2.1). Table 12 varies the number of quadrature nodes on clean CIFAR-10.

Table 12: NED quadrature resolution on clean CIFAR-10 (single seed, 100K steps).

Quadrature nodes	FID
16	**25.7**
32	30.0
64	29.7
128	29.7

Performance is broadly stable across quadrature resolutions; the variation (FID 25.7–30.0) is within the cross-seed noise observed in Table 2. All experiments in this paper use 64 nodes as a conservative default to ensure numerical stability across all mismatch levels.

Additional figures. Figure 4 shows the per-seed spread behind Table 2, Figure 5 paired samples at $\varepsilon = 0.30$, and Figure 6 the FID degradation of the robust baselines.

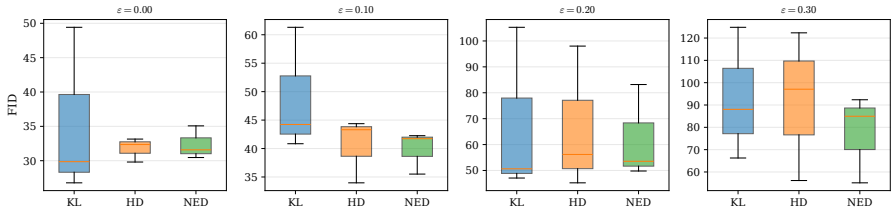

Figure 4: Box plots of CIFAR-10 FID across 3 seeds at each contamination level (noise replacement). HD and NED exhibit tighter interquartile ranges than KL, especially on clean data, indicating that bounded-influence objectives stabilize training.

K Class-conditional MNIST

The construction extends to class conditioning without modification (Appendix A). We ran class-conditional MNIST with the noise-replacement protocol; contaminated images keep their labels, so the corruption sits inside the class-conditional target, the harder case. The pipeline is that of Section 3 at 50K steps, single seed. Because FID does not test whether generation follows the conditioning signal, the primary metric is conditional fidelity: the accuracy with which a frozen classifier (98.9% on real test digits) assigns the conditioning class to 500 generated samples per class. Table 13 reports it together with FID (5,000 samples against the clean test split).

Trained clean, KL, HD and NED reach 97.3/97.7/97.5%; at $\varepsilon = 0.30$ all three fall to 67.5/68.4/66.8%, 0.70 times the clean values, and FID shows the same pattern. Under this severe contamination no objective prevents the small MNIST model from learning the noise mode, and the mismatch audit shows why this is a boundary case rather than a counterexample: contaminated samples stay at moderate absolute mismatch (median per-image mean mismatch over the noise levels $t \in \{100, 300, 500, 700, 900\}$ of 2.3 against 0.17 clean, a large relative separation of 13.5 at small absolute scale), where HD retains weight $e^{-2.3/4} \approx 0.57$. On a common reference

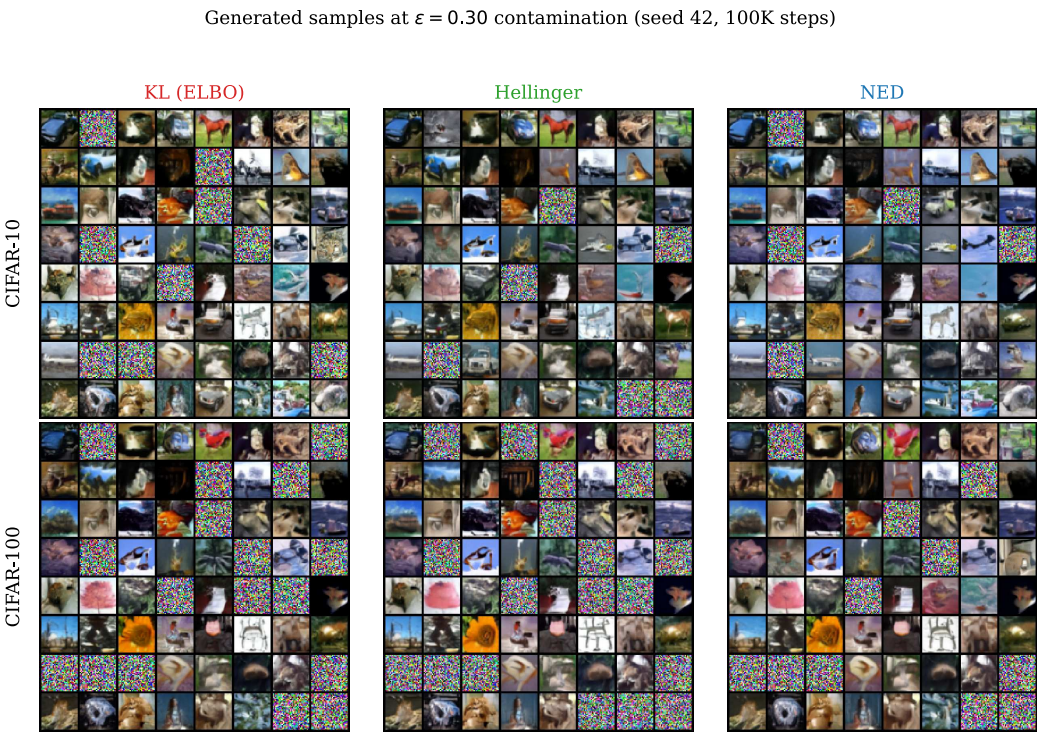

Figure 5: Generated samples at ε=0.30 contamination (noise replacement, seed 42, 100K steps). **Top:** CIFAR-10. **Bottom:** CIFAR-100. The comparison is deliberately paired: the training script draws the sample-grid latents once at initialization, after seeding and before the training loop, the three runs share the seed and the code path up to that draw, and the DDIM sampler is deterministic ($\eta = 0$), so the three grids start from the same latent tensor and closely related models produce matching layouts. The remaining differences between columns, such as the noise-like artifacts in the KL column that are absent under HD and NED, reflect the training objective and its stated protocol (HD includes the 5,000-step MSE warmup). Since this is one paired seed, the figure is illustrative; the quantitative comparisons rest on the tables across seeds (Table 2).

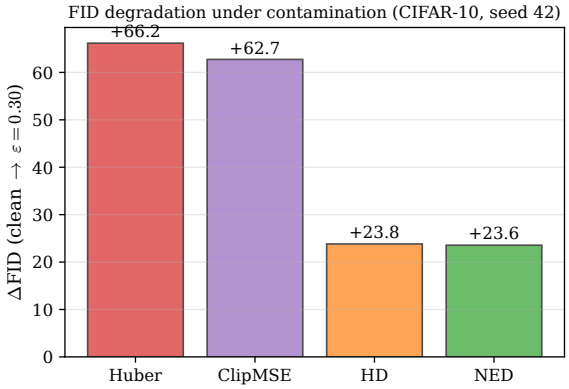

Figure 6: FID degradation (ΔFID from clean to ε=0.30, noise replacement) for standard robust losses vs. divergence-induced objectives on CIFAR-10 (seed 42). NED degrades least (+23.5), while Huber and clipped MSE degrade substantially.

Table 13: Class-conditional MNIST (noise replacement, 50K steps, single seed). Conditional accuracy of a frozen classifier on 500 generated samples per class, and FID.

	conditional accuracy (%)			FID		
ε	KL	HD	NED	KL	HD	NED
0.00	97.3	97.7	97.5	6.8	5.7	5.9
0.30	67.5	68.4	66.8	111.3	112.3	119.7

model (the KL run at $\varepsilon = 0.30$), the leakage-to-retention ratio computed on the per-image mismatch averaged over these noise levels is 0.59 (HD) and 0.28 (NED), ordered as Proposition 2.1 predicts but with no useful suppression margin for any member. On the same grid and at matched training depth, the CIFAR-10 noise-replacement checkpoint of the warmup audit below has medians 0.63 and 5.40, HD weight 0.26 at the contaminated median, and ratios 0.30 (HD) and 0.13 (NED); at the per-step level the contaminated medians at $t = 99$ are 8.1 on MNIST against 22.5 on CIFAR-10 (Appendix I), a scale gap in line with the pixel counts (784 against 3072). Robustness is governed by mismatch separation and scale, not by the nominal contamination fraction; the calibration rule of Appendix D.1 diagnoses this case from the per-step quantiles before training (0.6 clean and 8.1 contaminated at $t = 99$: the separation passes and the scale fails).

A prospective warmup audit. The same audit works before committing to a full training run. On SVHN at $\varepsilon = 0.30$, a 10K-step warmup model gives clean and contaminated medians of the per-image mean mismatch (over the same five noise levels as above) of 0.29 and 5.45 (separation 18.5), HD weight 0.26 at the contaminated median, and leakage-to-retention 0.28 (HD) and 0.098 (NED). A CIFAR-10 checkpoint at the same depth and on the same grid measures nearly identically (medians 0.63 and 5.40, separation 8.6, HD weight 0.26), and CIFAR-10 under noise replacement is the separated regime in which the robust members win (Section 3). By this matched comparison SVHN patterns with the separated CIFAR-10 regime, not with the MNIST boundary case; a warmup screen is a prediction; the full SVHN models were not trained.

L Limitations and Broader Impact

Limitations. The local conditional objective is a training surrogate motivated by the path-space upper bound. For general (non-KL) f-divergences it is not itself a proven upper bound on the marginal divergence $D_G(g\|f_\theta)$; Proposition 2.2 gives one-sided control of the path and marginal divergences by the local terms for generators comparable to Hellinger, but the bound is far from tight at trained losses (Appendix E), and the two-sided gap between the surrogate and the marginal divergence remains an open question. The asymptotic product-to-sum reduction (Theorem 2.3) is first-order in the local mismatch and holds inside an auxiliary construction with a pairwise uncorrelatedness assumption; our audit (Appendix H) shows that trained models are far from its small-mismatch regime, so it is an interpretation of the surrogate, not a description of the trained models, and finite-mismatch corrections remain an open direction. The robustness claims are empirically supported by the contamination experiments but are not accompanied by formal contamination-risk bounds or breakdown-point guarantees; Proposition 2.1 compares influence profiles under a monotone likelihood ratio premise and expresses no preference when the clean and contaminated mismatch laws overlap. The multi-seed CIFAR-10 results show substantial cross-seed variance, particularly at high contamination; the structured out-of-distribution results use three seeds, the CIFAR-100, data-cleaning and MNIST results are single seed, and the robust baselines were run at one to three seeds with untuned thresholds. All experiments use 32×32 images, a 35M-parameter model and a short training budget; higher-resolution validation is the first item of future work. The framework assumes Gaussian reverse kernels (DDPM variational posteriors) with fixed matched covariance and a fixed noise schedule; other parameterizations prescribe their own w_t (Section 2.2), but only one other schedule was tested (a single-seed cosine replicate, Appendix J) and no other parameterization, and extending the theory to non-Gaussian or learnable-covariance settings would broaden applicability. Under the Gaussian reverse-kernel structure, the framework extends naturally to flow matching by applying h_G to the velocity-field residual $\|v_\theta - v^*\|^2$; empirical validation of this extension is left to future work. Finally, the results suggest a broader design space of monotone, locally linear, saturating residual transformations h beyond those induced by classical f-divergences

(Theorem 2.1(f) shows that the standard Huber and clipping baselines are not members of the induced class); exploring that space, and integrating the induced objectives into ambient-type training or distillation from robustly trained teachers (Section 4), are directions for future work.

Broader impact. Diffusion models are widely used for image and audio generation, and improvements to their training objectives may amplify both beneficial applications (e.g., scientific visualization, creative tools) and potential misuse (e.g., deepfakes, disinformation). Our robust training framework may partially mitigate data-poisoning attacks by downweighting corrupted samples, but it does not address misuse of the generated content itself. We encourage practitioners to pair improved training methods with appropriate content safeguards.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the contributions (lognormal characterization and local conditional objective, classification of the induced objectives, a comparison principle for divergence selection, one-sided finite-horizon control of the path divergence, and the empirical validation), which are established in Sections 2–3; the scope of the theoretical results (asymptotic reduction, one-sided bound) is stated in Section 2.4 and Appendix L.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 and Appendix L discuss limitations, including the surrogate nature of the objective for non-KL divergences, the looseness of the finite-horizon bound at trained losses, the measured distance of trained models from the asymptotic regime (Appendix H), the single-seed status of several experiments, and the restriction to Gaussian matched-covariance reverse kernels at 32×32 resolution.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All results state their assumptions explicitly (including the $G''(0) = 1$ normalization, growth, tail integrability, comparability and moment conditions). Complete proofs are provided in Appendices A–G; the product-to-sum result (Theorem 2.3, Corollary 2.1) is asymptotic and relies on explicit moment, orthogonality, and tail conditions listed in Appendices F and G.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 and Appendix J detail both contamination protocols, the network architecture, training hyperparameters (including the HD warmup), seeds, and evaluation metrics. CIFAR-10 main results and the structured contamination regime use 3 seeds; CIFAR-100, the data-cleaning baselines, the ablations, the cosine replicate and MNIST use a single seed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code to reproduce all experiments (training scripts, sweep scripts, baseline comparisons, audit scripts) along with pre-computed result files is provided as supplementary material. CIFAR-10, CIFAR-100, MNIST and SVHN are publicly available via torchvision. The README provides exact commands and expected outputs for each table in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix J for the full experimental protocol including optimizer, learning rate, batch size, and contamination procedure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: CIFAR-10 main results and the structured contamination results report mean $\pm$ sample standard deviation over 3 seeds, with per-seed values in Appendix J. CIFAR-100, data-cleaning, ablation, cosine-replicate and MNIST results are single-seed and are stated as such.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix J specifies the GPU type, Adam optimizer at $\mathrm{lr} = 2\times10^{-4}$, approximate training time per run ($\sim$3h), and the compute of the main sweeps ($\sim$150 GPU-hours across all phases of the original submission); the additional runs reported in the appendices (structured contamination, cleaning baselines, extra seeds, MNIST, audits) have the same per-run cost.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The research is theoretical and methodological in nature and conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix L discusses broader impacts including potential misuse of generative models and the mitigation value of robust training.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This is a theoretical/methodological contribution; no pre-trained models or datasets are released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: CIFAR-10 and CIFAR-100 [Krizhevsky, 2009] are cited; MNIST and SVHN are standard public datasets accessed through torchvision. No other external assets are used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or pre-trained models are released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as a component of the core methodology; they were used only for editing (grammar, spelling, word choice).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.