Local Intrinsic Dimension Unveils Hallucinations in Diffusion Models

Bartlomiej Sobieski[1,2,3,4*†], **Matthew Tivnan**[3,4], **Michał Jan Włodarczyk**[1], **Dawid Płudowski**[1], **Pengfei Jin**[3,4], **Przemyslaw Biecek**[1,2], **Quanzheng Li**[3,4]

[1]Centre for Credible AI, Warsaw University of Technology, [2]University of Warsaw, [3]Massachusetts General Hospital, [4]Harvard Medical School

Abstract

Diffusion models are prone to generating structural hallucinations - samples that match the statistical properties of the training data yet defy underlying structural rules, resulting in anomalies like hands with more than five fingers. Recent research studied this failure mode from several viewpoints, offering partial explanations to their occurrence, such as mode interpolation. In this work, we propose a complementary perspective that treats hallucinations as instabilities on the model-induced manifold. We begin by showing that a hallucination filter based on such instabilities matches or exceeds the performance of the recently proposed temporal one. By tracing the source of these instabilities, we identify local intrinsic dimension (LID) as their primary driver and propose Intrinsic Quenching (IQ), a direct corrective mechanism that deflates it to alleviate hallucinations. IQ consistently outperforms standard hallucination reduction baselines across a wide array of benchmarks and offers a highly promising solution for enforcing anatomical consistency in downstream medical imaging tasks.

1 Introduction

Diffusion models [DMs; Sohl-Dickstein et al., 2015, Song et al., 2021, Lipman et al., 2023, Liu et al., 2023b, Albergo and Vanden-Eijnden, 2023, Song et al., 2023, Geng et al., 2025] drive modern generative modeling across computer vision and other modalities [Dhariwal and Nichol, 2021, Lou et al., 2024, Kong et al., 2021, Kotelnikov et al., 2023, Hoogeboom et al., 2022]. Recent research highlights their critical failure mode termed *hallucinations*: generated samples that successfully capture the statistical properties of the training data, but fundamentally violate its underlying physical, logical, or morphological patterns [Ramesh et al., 2021, Rombach et al., 2022].

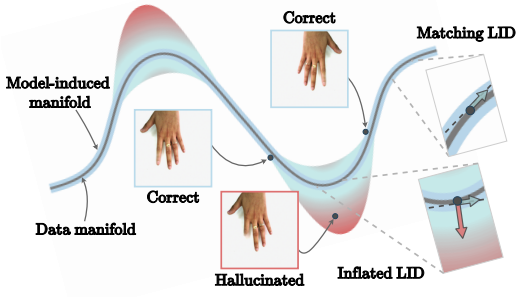

Figure 1: A diffusion model-induced manifold, approximating the data manifold, invents spurious dimensions in unstable regions, leading to structural hallucinations.

In particular, *structural* hallucinations [Kim et al., 2024] fail to recover correct object forms, generating anomalies like six-fingered hands or misaligned eyes. Lacking a formal definition, identifying them relies strictly on human perception, restricting research to empirical and correlational studies [Oorloff et al., 2025, Triaridis et al., 2025, Tian et al., 2025]. Despite initial explanatory hypotheses

*Corresponding author at b.sobieski@uw.edu.pl

†Work done during an internship at Harvard Medical School and Massachusetts General Hospital

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

[Aithal et al., 2024] and concurrent mitigation approaches [Chandran.C et al., 2025, Wewer et al., 2025], understanding their root causes and prevention remains an open question.

In this work, we propose shifting the analysis of structural hallucinations from temporal generative behavior [Aithal et al., 2024, Tian et al., 2025] to the local instability of the model-induced manifold. Specifically: **C1.** We demonstrate that a spatial hallucination filter based on local geometric instability matches or exceeds the recently proposed temporal filter [Aithal et al., 2024], identifying the inflation of local intrinsic dimension (LID) as the primary indicator. **C2.** We theoretically formulate *Intrinsic Quenching* (IQ), a thermodynamically-inspired corrective mechanism that dynamically deflates LID of the generated sample on the model-induced manifold throughout the generative process, reducing the hallucinatory behavior. **C3.** Through large-scale, human-annotated evaluations, we show that IQ outperforms all baselines in reducing structural hallucinations and offers a promising solution for enforcing anatomical consistency in downstream medical imaging tasks.

2 Background

DMs can be formulated through the framework of stochastic differential equations (SDEs):

$$\mathrm{d}\mathbf{x}_t = \mathbf{F}_t\mathbf{x}_t\mathrm{d}t + \mathbf{G}_t\mathrm{d}\mathbf{w}_t, \tag{1}$$

$$\mathrm{d}\mathbf{x}_t = [\mathbf{F}_t\mathbf{x}_t - \mathbf{G}_t\mathbf{G}_t^\top\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)]\mathrm{d}t + \mathbf{G}_t\mathrm{d}\overline{\mathbf{w}}_t, \tag{2}$$

where $\mathbf{F}_t\mathbf{x}_t$ is a linear *drift term* with $\mathbf{x}_t \in \mathbb{R}^n$ and time-dependent matrix $\mathbf{F}_t \in \mathbb{R}^{n\times n}$, $\mathbf{G}_t \in \mathbb{R}^{n\times n}$ is a matrix-valued *diffusion* coefficient, $\mathbf{w}_t \in \mathbb{R}^n$ and $\overline{\mathbf{w}}_t \in \mathbb{R}^n$ are the Wiener processes running forward and reverse in time respectively, $\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)$ is the *score* function of the time-parameterized distribution $p(\mathbf{x}_t), t \in [0,1]$, where $p(\mathbf{x}_0)$ denotes the data and $p(\mathbf{x}_1)$ a terminal distribution, typically chosen to be Gaussian. A DM, parameterized by $\boldsymbol{\theta}$, is trained to approximate the score function.

Equation (1) defines the *forward* process, which gradually corrupts data samples into Gaussian noise, whereas eq. (2) defines the *reverse* process, which generates the data through iterative denoising. The Probability Flow ODE (PF-ODE) is the deterministic counterpart of eq. (2) given by $\mathbf{F}_{\text{ODE}}(\mathbf{x}_t, t) \triangleq \frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = [\mathbf{F}_t\mathbf{x}_t - \frac{1}{2}\mathbf{G}_t\mathbf{G}_t^\top\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)]$. Substituting the $\boldsymbol{\theta}$-parameterized model's approximated score, we define the *generator* as

$$\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1) = \mathbf{x}_1 - \int_0^1 \mathbf{F}^{\boldsymbol{\theta}}_{\text{ODE}}(\mathbf{x}_t, t)\mathrm{d}t, \tag{3}$$

denoting the model-based mapping of a sample $\mathbf{x}_1 \sim p(\mathbf{x}_1)$ to the approximate data distribution $p_{\boldsymbol{\theta}}(\mathbf{x}_0)$ given by the DM. At the center of our reasoning, we rely on the following assumption.

Assumption 1. *We follow the stratified manifold hypothesis [Goresky and MacPherson, 1988], which posits that high-dimensional data resides on a disjoint collection of low-dimensional submanifolds, and assume that a DM is able to learn a statistical approximation of these strata.*

Based on assumption 1, we define $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}} = \{\mathbf{x}_0 \mid \exists_{\mathbf{x}_1\in\mathbb{R}^n}\mathbf{x}_0 = \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1)\}$, the set of all possible outputs of the generator, and refer to it as the model-induced manifold for simplicity. This construction is justified by foundational works and recent research on the geometry of DMs [Fefferman et al., 2016, Pidstrigach, 2022, Stanczuk et al., 2024].

Training DMs relies on a tractable formula for the forward kernel $p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{H}_t\mathbf{x}_0, \boldsymbol{\Sigma}_t)$, where $\mathbf{H}_t = \boldsymbol{\Phi}(t,0)$ with $\boldsymbol{\Phi}(t,s) = \exp\left(\int_s^t \mathbf{F}_u\mathrm{d}u\right)$, assuming $\mathbf{F}_t$ commutes for all t, and $\boldsymbol{\Sigma}_t = \int_0^t \boldsymbol{\Phi}(t,\tau)\mathbf{G}_\tau\mathbf{G}_\tau^\top\boldsymbol{\Phi}(t,\tau)^\top\mathrm{d}\tau$. To train a DM, one can either use $\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t \mid \mathbf{x}_0)$ as target or utilize equivalent objectives such as

$$\mathcal{L}_{\text{DSM}}(\mathbf{x}_0, t, \boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t)\|_2^2\right], \tag{4}$$

$$\mathcal{L}_{\text{ISM}}(\mathbf{x}_0, t, \boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_t\nabla_{\mathbf{x}_t}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t)\right) + \frac{1}{2}\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t)\|^2_{\boldsymbol{\Sigma}_t}\right] \tag{5}$$

where $\mathbf{x}_t = \mathbf{H}_t\mathbf{x}_0 + \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\epsilon}$, in each case marginalizing over $\mathbf{x}_0$ and t to learn the underlying score function [Hyvärinen, 2005, Vincent, 2011, Yeats et al., 2025]. We refer to $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ as the noise prediction and $\mathbf{s}_{\boldsymbol{\theta}}$ as the score prediction.

Thanks to the Tweedie's formula [Efron, 2011, Alain and Bengio, 2014], one can directly relate the score function to the mean of the posterior $p(\mathbf{x}_0 \mid \mathbf{x}_t)$ through

$$\hat{\mathbf{x}}_0(\mathbf{x}_t) \triangleq \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] = \boldsymbol{\Phi}(t,0)^{-1}\left(\mathbf{x}_t + \boldsymbol{\Sigma}_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)\right). \tag{6}$$

We denote by $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)$ the result of replacing the true score function in eq. (6) with $\mathbf{s}_{\boldsymbol{\theta}}$.

Arising from the manifold hypothesis, LID [Levina and Bickel, 2004, Bengio et al., 2013] encodes the local tangent space dimensionality, representing the valid degrees of freedom within a data stratum [Pope et al., 2021]. Historically a diagnostic tool in early deep learning research [Ma et al., 2018, Ansuini et al., 2019], LID can now be natively estimated by DMs. This capability, which justifies assumption 1, relies on $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)$ acting as an approximate orthogonal projector for small t. While early DM-based estimators were computationally heavy [Tempczyk et al., 2022, Stanczuk et al., 2024, Kamkari et al., 2024], Yeats et al. [2025] demonstrated that standard training losses (eqs. (4) and (5)) provide efficient upper bounds of LID.

3 Empirical Investigation

The initial hypothesis given by Aithal et al. [2024] suggested that DMs' hallucinations stem from interpolating between the modes of the data distribution and thus amplifying the regions of low probability. To filter out hallucinations post-generation, the authors proposed the Trajectory Variance Filter (TVF) defined as $\mathrm{TVF}(\mathbf{x}_0) = \int_{t_1}^{t_2} \left\| \hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t) - \overline{\hat{\mathbf{x}}_{0,t_1:t_2}^{\boldsymbol{\theta}}} \right\|_2^2 \mathrm{d}t$, where $\overline{\hat{\mathbf{x}}_{0,t_1:t_2}^{\boldsymbol{\theta}}}$ is the average predicted posterior mean over the $[t_1, t_2]$ interval. This represents the *temporal* view, which is used to analyze hallucinations using the intermediate properties of the reverse process. We propose to shift that perspective and instead focus on the *spatial* view by analyzing the local geometry of the DM on the induced manifold.

We define the Local Manifold Instability (LMI) of $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}$ for an initial point $\mathbf{x}_1$ as

$$\mathrm{LMI}(\mathbf{x}_1) \triangleq \mathrm{Tr}(\mathrm{Cov}_{\boldsymbol{\varepsilon}}(\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1 + \boldsymbol{\varepsilon}))) = \mathrm{Var}_{\boldsymbol{\varepsilon}}(\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1 + \boldsymbol{\varepsilon})), \tag{7}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \beta^2 \mathbf{I})$ for some small $\beta > 0$, often also referred to as *local sensitivity* [Cacuci et al., 2005]. Intuitively, LMI measures the total spatial spread of a small spherical region of noise after it is transported to the data space, providing a proxy for exploring the local geometry of a given DM.

Geometric perspective. We begin with a core question: *If structural hallucinations can be partially captured by temporal variance during the reverse process, do these failure modes also manifest as localized spatial instabilities on the model-induced manifold?* To assess it, we compare the performance of TVF and LMI as hallucination filters on a benchmark of hand images from the `11kHands` dataset [Afifi, 2019], where they are easily recognizable as hands with deformed, missing or additional fingers. We use the capable EDM model [Karras et al., 2022] trained by Tian et al. [2025] with a deterministic 40-step Euler solver to generate 128 samples. To decide whether a given sample is correct or represents a structural hallucination, we conduct a user study with independent human annotators to label the samples in a binary manner. For TVF, we collect $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)$ for each t throughout generation and then optimize t_1 and t_2 for best performance. For LMI, we estimate it with 32 perturbations per sample for a chosen β. For more details, see section B.2.

Figure 2 depicts densities for both filters divided between correct and hallucinated samples, as well as classification and separability performance metrics. Interestingly, both filters are highly effective in separating correct and hallucinated samples, with a minor advantage of LMI despite choosing the optimal time interval for TVF. While the practical applicability of LMI is limited due to its computational costs based on running the reverse process multiple times, these results provide a novel identification of hallucinations as unstable states on the model-induced manifold. We extend this experiment to other datasets in section B.3, showing that while the performance of these two filters varies, the relationship between them is preserved.

To better understand the predictive ability of LMI, we provide a proposition that connects it with the LID of the generated sample on the model-induced manifold, which we denote as $\mathrm{LID}_{\boldsymbol{\theta}}$.

Proposition 1. *Let $\mathbf{x}_0 = \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1)$ and assume a sufficiently small perturbation scale $\beta > 0$ such that a first-order linear approximation holds. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ denote the singular values of*

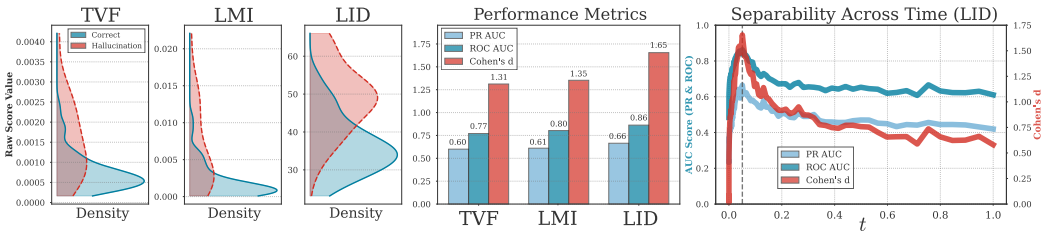

Figure 2: **Left.** Estimated densities for the values of each filter across correct and hallucinated samples. **Middle.** Separability and classification performance metrics for each filter. **Right.** Evolution of LID filter performance across the time interval. Vertical line indicates the optimal timestep. All results are reported on `11kHands` dataset with EDM.

the generator's Jacobian $\nabla_{\mathbf{x}_1}\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1)$*, where* n *is the ambient dimension. Then it holds that*

$$\mathrm{LMI}(\mathbf{x}_1) \approx \beta^2 \sum_{i=1}^{\mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0)} \sigma_i^2. \tag{8}$$

We provide the proofs and full formal statements of all theoretical results in section A.

Assumption 1 implies that $\mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0)$ measures the *effective* dimensionality (significantly non-zero singular values) of $\mathbf{x}_0$ at $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$ when considering a theoretically full-rank $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}$. Proposition 1 hence instantly reveals the reasons for the inflated LMI: it is caused by **R1.** over-estimating the true singular values (of the true stratified manifold), thus resulting in more rapid but plausible modifications under perturbation, or **R2.** the invention of spurious, off-manifold directions that influence the generated sample in an implausible or artifactual manner.

With proposition 2, we extend the result of Yeats et al. [2025] to show that the losses in eqs. (4) and (5), shown to upper-bound the true LID, also form unbiased estimators of $\mathrm{LID}_{\boldsymbol{\theta}}$.

Proposition 2. *Let* $\mathbf{x}_0 = \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1)$*. Under mild regularity conditions, it is true that* $\mathcal{L}_{DSM}(\mathbf{x}_0, t, \boldsymbol{\theta}) = \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0)$, $\mathcal{L}_{ISM}(\mathbf{x}_0, t, \boldsymbol{\theta}) = -\frac{1}{2}(n - \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0))$*, where* n *is the dimension of the ambient space and* t *is sufficiently small.*

Hence, proposition 2 provides a tool to investigate the individual influences of **R1.** and **R2.** on the inflation of LMI by estimating $\mathrm{LID}_{\boldsymbol{\theta}}$ for both correct and hallucinated samples. Crucially, it also results in a continuous relaxation of the problem of estimating $\mathrm{LID}_{\boldsymbol{\theta}}$, a theoretically discrete quantity.

Under the same experimental setup, we compare the performance of $\mathrm{LID}_{\boldsymbol{\theta}}$ as a hallucination filter with TVF and LMI using $\mathcal{L}_{\mathrm{DSM}}$-based estimator (eq. (4)) with 32 noise samples for approximating the expectation. Figure 2 (right) first shows how this performance depends on t, revealing large variability and the importance of choosing a small enough timestep, following the assumption of proposition 2. Crucially, fig. 2 (middle) shows that $\mathrm{LID}_{\boldsymbol{\theta}}$ significantly outperforms both LMI and TVF using the optimal t, providing the best separability and classification performance with higher values indicating hallucinations.

From a geometric perspective, this result reveals the primary origin of the local instabilities; hallucinations occur when the DM overestimates the complexity of the generated samples by inflating their $\mathrm{LID}_{\boldsymbol{\theta}}$. This result suggests that hallucinations appear when the model-induced manifold possesses unnecessary directions of expansion, which represent illogical degrees of freedom, such as a varying number of fingers or the possibility of more than one thumb (**R2.**). This also explains the relatively worse performance of LMI, which is simultaneously inflated when the variability along entirely valid directions increases (**R1.**).

4 Theoretical Foundation

While the empirical evidence justifies a hallucination filter based on $\mathrm{LID}_{\boldsymbol{\theta}}$, filters themselves remain only a partial solution to the hallucination problem, as they inevitably require generating multiple invalid samples. In what follows, we propose a corrective mechanism for the sampling process inspired by the evidence that one should deflate $\mathrm{LID}_{\boldsymbol{\theta}}$ for particular samples to avoid hallucinations.

Theorem 1. *Let* $\mathbf{x}_0 = \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1)$*. Assume that* $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1)$ *is decomposed as* $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}} = \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{\leq\tau} \circ \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{>\tau}$ *for sufficiently small time* $\tau > 0$*, where* $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{>\tau} : \mathbf{x}_1 \mapsto \mathbf{x}_\tau$ *and* $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{\leq\tau} : \mathbf{x}_\tau \mapsto \mathbf{x}_0$*. Let the singular values*

of the macroscopic flow Jacobian $\nabla_{\mathbf{x}_1}\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{>\tau}(\mathbf{x}_1)$ *be monotonically ordered as* $\sigma_1^{>\tau} \geq \sigma_2^{>\tau} \geq \cdots \geq \sigma_n^{>\tau} \geq 0$*. For sufficiently small* $\beta > 0$*, it is true that*

$$\mathrm{LMI}(\mathbf{x}_1) \lessapprox \beta^2 \sum_{i=1}^{\mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau))} (\sigma_i^{>\tau})^2. \tag{9}$$

The contribution of theorem 1 is two-fold. First, it preserves the connection between local instability and the sample's $\mathrm{LID}_{\boldsymbol{\theta}}$, while only requiring the posterior mean $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau)$ for some small $\tau > 0$ instead of the final $\mathbf{x}_0$. Second, it replaces the global singular values from proposition 1 with a spectral bottleneck: a summation where the macroscopic variance ($t > \tau$) is strictly truncated by the intrinsic dimensionality of the terminal projection ($t \leq \tau$). Therefore, actively steering the trajectory to minimize $\mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau))$ dynamically drops variance terms from this summation, provably decreasing the upper bound of LMI for the resulting sample. We formalize the implied sampling process below.

Theorem 2. *Assume* $t \leq \tau$*, where* $\tau > 0$ *is sufficiently small. Replacing the standard trained score function* $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t)$ *in the reverse process (eq.* (2)*) with*

$$\tilde{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{x}_t) = \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t) - \lambda_t \nabla_{\mathbf{x}_t}\mathcal{E}(\mathbf{x}_t), \tag{10}$$

where the energy $\mathcal{E}(\mathbf{x}_t) = \mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t))$*, is equivalent to sampling from the ideal Boltzmann distribution over the true terminal states,* $p_t^{\boldsymbol{\theta},\lambda_t}(\mathbf{x}_t) \propto p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)\mathbb{E}_{p^{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_t)}[\exp(-\lambda_t \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0))]$.

Theorem 2 guarantees that, for small enough t, correcting the reverse process with the gradient of $\mathrm{LID}_{\boldsymbol{\theta}}$ for the posterior mean $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)$ moves the resulting sample along $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$ towards terminal states of lower $\mathrm{LID}_{\boldsymbol{\theta}}$, *i.e.*, a lower-dimensional stratum. We obtain the gradient through $\nabla_{\mathbf{x}_t}\mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)) \approx \nabla_{\mathbf{x}_t}\mathcal{L}_{\mathrm{DSM}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t), t, \boldsymbol{\theta})$, where the approximation comes from estimating the DSM expectation with k i.i.d noise samples.

Probabilistic perspective. The resulting correction represents a projected gradient descent scheme on $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$. However, utilizing the $\mathcal{L}_{\mathrm{DSM}}$ estimator does not immediately reveal the probabilistic effect of that correction on the generated sample $\mathbf{x}_0$. To address that, we provide corollary 1 showing that one may interchangeably use either the $\mathcal{L}_{\mathrm{DSM}}$- or $\mathcal{L}_{\mathrm{ISM}}$-based estimator.

Corollary 1. *The gradients* $\nabla_{\mathbf{x}_t}\mathcal{L}_{DSM}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t), t, \boldsymbol{\theta})$ *and* $\nabla_{\mathbf{x}_t}\mathcal{L}_{ISM}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t), t, \boldsymbol{\theta})$ *are exactly collinear and* $\nabla_{\mathbf{x}_t}\mathcal{L}_{DSM}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t), t, \boldsymbol{\theta}) = 2\nabla_{\mathbf{x}_t}\mathcal{L}_{ISM}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t), t, \boldsymbol{\theta})$.

This collinearity allows us to reason about the probabilistic effect of the stable and computationally cheap $\mathcal{L}_{\mathrm{DSM}}$ gradient through the theoretical lens of $\mathcal{L}_{\mathrm{ISM}}$.

Corollary 2. *Assuming* $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t) = \nabla_{\mathbf{x}_t}\log p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)$*, eq.* (5) *can be equivalently reformulated as* $\mathcal{L}_{ISM}(\mathbf{x}_0, t, \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_t\nabla_{\mathbf{x}_t}^2\log p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)\right) + \frac{1}{2}\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t)\right\|_{\boldsymbol{\Sigma}_t}^2\right]$*. Thus, as* $t \to 0$*, minimizing the energy from eq.* (10) *steers towards stationary points (*$\nabla_{\mathbf{x}_t}\log p_t^{\boldsymbol{\theta}}(\mathbf{x}_t) \to \mathbf{0}$*) with maximal negative curvature (*$\mathrm{Tr}\left(\boldsymbol{\Sigma}_t\nabla_{\mathbf{x}_t}^2\log p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)\right) \ll 0$*) of the model-induced probability distribution* $p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)$*, inducing a mode-seeking behavior.*

By mitigating mode interpolations, corollary 2 directly connects our optimization scheme with the work of Aithal et al. [2024]. As the proposed correction essentially aims at eliminating unstable, low-probability states by 'cooling' them and shifting towards stable local maxima, we term our approach as *Intrinsic Quenching* (IQ).

From theory to practice. Both the theoretical assumptions and the separability performance from fig. 2 indicate that IQ should be only used for small times $t < \tau$. We thus apply it solely within a narrow interval $[t_1, t_2]$ of the reverse process, where t_2 is small. This focused intervention is also empirically justified when inspecting the energy gradients $\nabla_{\mathbf{x}_t}\mathcal{E}(\mathbf{x}_t)$ (fig. 3). By overlaying them over the final generated image, we discover a smooth transition from global, structural attributes through localized, highly-specific features correlated with the hallucinatory content, up to dispersed but cohesive states. This observation naturally aligns with multiple prior studies that decompose the diffusion process into semantically interpretable phases [Park et al., 2023, Sclocchi et al., 2025, Wang and Pehlevan, 2025]. In this case, however, $\mathrm{LID}_{\boldsymbol{\theta}}$ provides an inherent attribution map [Lundberg and Lee, 2017, Sundararajan et al., 2017] that allows quantifying such effects, as it indicates pixels with the highest contribution to the model's error at time t.

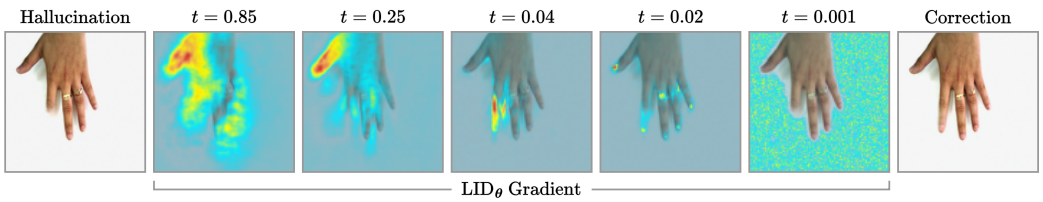

Figure 3: *Hallucination* depicts the initial unconditionally generated sample by EDM on `11kHands`. For a set of manually selected timesteps, $\text{LID}_{\boldsymbol{\theta}}$ gradients are visualized to highlight their coarse-to-fine transition and perceptual connection to hallucinated content. *Correction* displays the sample generated with IQ, the proposed corrective mechanism applied in the $[0.02, 0.05]$ interval.

To ensure that the modified updates behave stably, we compute the λ_t scale dynamically for each t by ensuring that the magnitude of the energy term projected into data space ($\left\|\lambda_t \boldsymbol{\Phi}(t,0)^{-1} \boldsymbol{\Sigma}_t \nabla_{\mathbf{x}_t} \mathcal{E}(\mathbf{x}_t)\right\|_2$) always equates to a fixed ratio λ of the natural update. For example, in the EDM framework, where the network is trained directly to predict $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)$ (and $\boldsymbol{\Phi}(t,0) = \mathbf{I}$), we use $\lambda_t = \lambda \frac{\left\|\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t) - \mathbf{x}_t\right\|_2}{\left\|\boldsymbol{\Sigma}_t \nabla_{\mathbf{x}_t} \mathcal{E}(\mathbf{x}_t)\right\|_2 + \epsilon}$, where ϵ ensures numerical stability.

The DM's score function naturally interacts with the energy gradient from IQ; the former pushes the sample towards $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$ with decreasing t, while the latter shifts the sample along the manifold to regions of lower $\text{LID}_{\boldsymbol{\theta}}$, suppressing the most volatile directions. To ensure that this intervention selectively targets hallucinations rather than uniformly restricting all samples, we employ $\text{LID}_{\boldsymbol{\theta}}$ as a dynamic, mid-generation filter. Specifically, we disable the correction ($\lambda_t = 0$) for stable samples satisfying $\mathcal{E}(\mathbf{x}_t) < q_t$, where q_t is a time-dependent threshold. To calibrate q_t, we unconditionally generate a reference set of samples from the DM, precompute $\mathcal{E}(\mathbf{x}_t)$ for each $\mathbf{x}_t$ across the active interval $[t_1, t_2]$, and set q_t to the q-th percentile of these empirical energy values at each timestep. We provide the pseudocode of IQ in algorithm 1.

5 Related Works

Hallucination reduction. Observed since early natural image DMs [Ramesh et al., 2021, Rombach et al., 2022], hallucinations were initially hypothesized by Aithal et al. [2024] as effects of mode interpolations, exacerbating errors during iterative DM training. Subsequent mitigations [Fu et al., 2025, Cho et al., 2025, Bhosale et al., 2026, Chandran.C et al., 2025] include Adaptive Attention Modulation [AAM; Oorloff et al., 2025], which optimizes attention temperatures [Vaswani et al., 2017] via an anomaly detector [Roth et al., 2022], requiring the original training data. Dynamic Guidance [DG; Triaridis et al., 2025] applies Classifier Guidance [CG; Dhariwal and Nichol, 2021] toward the most probable class at each step, necessitating an independently trained noisy classifier. More recently, Tian et al. [2025] (RODS) formulate DMs as a continuation method [Allgower and Georg, 2012], mitigating hallucinations by intervening intrinsically upon detecting vector field instabilities.

Task-specific methods. Because deformed hands are a prominent failure mode, several methods aim to correct or filter them specifically [Narasimhaswamy et al., 2024, Lu et al., 2024, Shi et al., 2026]. Similarly, Lu et al. [2025] identify local generation bias causing visual text failures. Hallucinations pose critical risks in image translation and reconstruction by affecting real-world decisions. Addressing this, Kim et al. [2024] propose local diffusion for conditional translation, partitioning in- and out-of-distribution regions. In reconstruction, several works introduce quantitative hallucination scores [Tivnan et al., 2024, Ren et al., 2025, Ku et al., 2025]. Distinctively, Cao et al. [2025] employ temporal score analysis, akin to Aithal et al. [2024] and Tian et al. [2025], to eliminate related out-of-distribution artifacts.

6 Experiments

Datasets. We perform a large-scale quantitative comparison of IQ with recent methods for hallucination reduction by unifying several existing benchmarks. As a tractable toy problem, we use the `GaussianGrid` dataset proposed by Aithal et al. [2024], comprising a 2D mixture of uniformly spaced Gaussians in a form of rotated grid with 25 modes. For semantic structural hallucinations, we

Table 1: Quantitative comparison of all sampling methods across six datasets. We highlight the primary performance indicators, based on human evaluation, regarding perceived image quality (UP) and hallucination ratio (HR), reported as averages across annotators. For 95% CIs, see table 6.

	11kHands					FFHQ					AFHQV2				
Method	FID ↓	IV ↑	DSV ↑	UP ↑	HR ↓	FID ↓	IV ↑	DSV ↑	UP ↑	HR ↓	FID ↓	IV ↑	DSV ↑	UP ↑	HR ↓
Baseline	16.3	0.032	0.15	39.8	29.3	13.5	0.067	0.40	45.3	8.2	16.7	0.054	0.48	41.8	6.9
DG	16.2	0.033	0.15	39.5	29.7	13.8	0.070	0.40	44.1	9.8	16.7	0.054	0.48	40.2	8.6
AAM	16.4	0.033	0.15	40.6	29.3	13.5	0.068	0.40	45.7	10.2	17.6	0.055	0.49	42.2	14.3
RODS^{CAS}	15.8	0.033	0.15	40.2	25.8	13.6	0.066	0.40	45.3	7.7	16.9	0.054	0.48	42.2	6.9
RODS^{SAS}	16.2	0.033	0.15	41.0	29.7	13.6	0.066	0.40	45.3	8.2	16.8	0.054	0.48	42.2	6.6
IQ	16.6	0.030	0.13	68.0	9.0	13.9	0.065	0.39	46.1	4.2	17.0	0.054	0.48	42.6	5.9

	MNIST				SimpleShapes				GaussianGrid	
Method	FID ↓	IV ↑	DSV ↑	HR ↓	FID ↓	IV ↑	DSV ↑	HR ↓	MMD ↓	HR ↓
Baseline	32.3	0.050	0.13	37.3	27.5	0.031	0.11	25.8	0.010	20.2
DG	32.1	0.050	0.14	23.6	27.8	0.031	0.11	27.3	0.080	11.1
AAM	32.3	0.049	0.13	34.8	27.2	0.025	0.06	55.5	-	-
RODS^{CAS}	35.4	0.049	0.13	41.8	28.1	0.031	0.10	30.5	0.016	10.4
RODS^{SAS}	34.8	0.047	0.13	55.3	27.8	0.031	0.11	28.9	0.016	19.4
IQ	31.8	0.051	0.14	10.2	23.3	0.033	0.13	9.4	0.016	8.9

evaluate on SimpleShapes and MNIST for synthetic images following Aithal et al. [2024], Triaridis et al. [2025], Oorloff et al. [2025], and animal faces [AFHQV2, 64×64 resolution; Choi et al., 2020], human faces [FFHQ, 64×64 resolution Karras et al., 2019] and hand images [11kHands, 128×128 resolution; Afifi, 2019] entirely following Tian et al. [2025] for natural images.

Models. For GaussianGrid, we train a 3-layer MLP with hidden dimension and time embedding of size 256 based on DDPM [Ho et al., 2020] with a linear schedule and 1000 steps. For SimpleShapes, we use a UNet [Ronneberger et al., 2015] with 3 convolutional blocks and an attention layer for both encoder and decoder, train it using 1000-step DDPM with squared cosine scheduler [Nichol and Dhariwal, 2021] and use 250-step sampling for inference. For MNIST, we use a pretrained UNet[3] from HuggingFace [von Platen et al., 2022] also based on 1000-step DDPM and 250-step inference. For AFHQV2 and FFHQ, we use unconditional VE EDM checkpoints from the original paper [Karras et al., 2022] and sample with a deterministic 40-step Euler solver, following Tian et al. [2025]. For 11kHands, we reuse the VE EDM network pretrained by Tian et al. [2025], also with 40-step sampling for inference.

Baselines. We compare IQ against recent methods for hallucination reduction: DG [Triaridis et al., 2025], AAM [Oorloff et al., 2025] and RODS [Tian et al., 2025]. Importantly, DG requires labeled training data of the DM of interest and assumes access to a pretrained noisy classifier for CG. Similarly, AAM also assumes access to training data to obtain the anomaly detection model and is limited to architectures with attention layers. In this context, RODS and IQ remain the only methods with no additional assumptions about the DM of interest. Our comparisons also include the standard unmodified sampling as a default baseline. For details regarding the unification of methods and adaptation to our codebase, see section B.1.

Metrics. On GaussianGrid, we measure hallucination ratio (HR) as the percentage of samples exceeding 3 standard deviations from the closest mode [Aithal et al., 2024], alongside Maximum Mean Discrepancy (MMD) for distributional similarity. For images, we evaluate 2048 samples per method using proxy metrics: Fréchet Inception Distance [FID; Heusel et al., 2017] for quality, Inception Variance [IV; Miao et al., 2024] and DreamSim Variance [DSV; Domingo-Enrich et al., 2025] for diversity [Tian et al., 2025].

Because automated evaluations struggle with semantic perception, we quantify HR on image datasets via a blinded user study (section B.2). Annotators first complete a *calibration phase* viewing 128 true images to grasp dataset variability. In the subsequent *labeling phase*, they identify structural hallucinations in a blinded manner on outputs generated by competing methods from identical latent points $\mathbf{x}_1$. Furthermore, feature-based metrics notoriously misalign with human perception, often favoring texture over structure [Geirhos et al., 2018], a flaw amplified in DMs where human

[3] https://huggingface.co/1aurent/ddpm-mnist

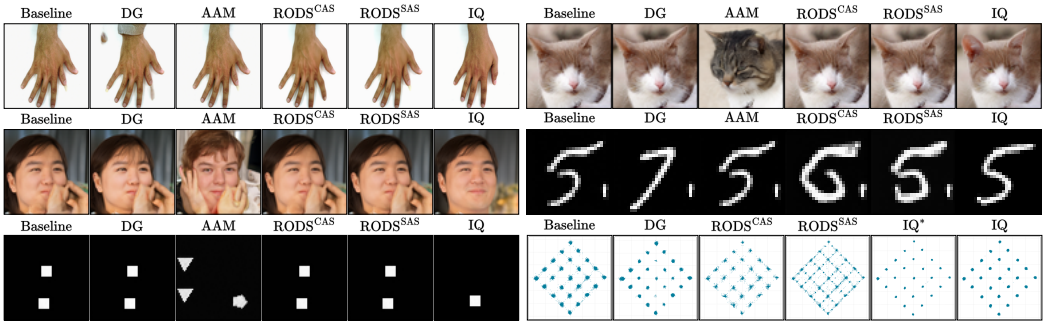

Figure 4: Qualitative comparison of all sampling methods across six datasets. AAM is omitted on `GaussianGrid` due to incompatible network architecture. IQ* denotes a variant of IQ without filtering ($q = 0$). For high-resolution results on `GaussianGrid`, see fig. 17.

preference remains the ultimate ground truth [Stein et al., 2023, Jayasumana et al., 2024]. Addressing this, we expand our study to evaluate User Preference (UP) on natural images. For each latent point, annotators select the highest-quality generations, with UP reported as the overall selection frequency per method.

Results. Quantitative and qualitative results are summarized in table 1 and fig. 4. On `GaussianGrid`, IQ confirms its theoretical validity: it minimally deviates from the baseline while maximally reducing hallucinations by eliminating mode interpolations. Figure 4 also highlights the role of dynamic filtering; without it, IQ over-squeezes samples into modes, whereas filtering halts correction appropriately. Crucially, IQ succeeds even when true data LID equals the ambient dimension, confirming our probabilistic interpretation. Conversely, DG achieves comparable HR but samples unevenly due to classifier bias, while RODS^{SAS} fails to meaningfully reduce hallucinations despite matching RODS^{CAS} in MMD.

On `SimpleShapes`, IQ reduces over 60% of hallucinations. We hypothesize the concurrent improvement in feature-based metrics stems from the specific nature of these failures: generating extra shapes creates natural outliers that inflate FID. Since these anomalies dominate baseline generation, their removal naturally increases valid diversity. This is corroborated on `MNIST`, where FID and diversity strongly correlate with HR. Despite extensive tuning, baselines struggle on these simple sets. This validates our LID formulation: an optimal DM must bound its factors of variation, and while IQ directly suppresses unnecessary degrees of freedom, other methods cannot effectively enforce this constraint.

On natural images, where human-centric UP and HR are the primary performance indicators, differences in feature-based metrics shrink. Crucially, on `11kHands`, IQ reduces hallucinations by almost 70% and achieves almost 70% UP, vastly outperforming baselines. The accompanying minor regressions in FID and diversity corroborate findings [Stein et al., 2023, Jayasumana et al., 2024] that such metrics fail to accurately capture human preference. On `FFHQ` and `AFHQV2`, performance gaps narrow. While users preferred IQ slightly more, overlapping confidence intervals (CIs) indicate no statistically significant UP difference, which annotators also attributed to the datasets' lower resolutions. Nonetheless, IQ successfully cuts hallucinations by around 50% on `FFHQ` and 15% on the highly challenging `AFHQV2`, where IQ's HR CI overlaps with RODS and baseline sampling despite a better average.

Hallucinations in medical diagnosis. Generative hallucinations pose severe risks in real-world decision-making, particularly in medical imaging where models restore high-frequency details to enable low-dose radiation [Wang et al., 2025]. To demonstrate IQ's practical capabilities, we evaluate hallucination reduction in low-dose computed tomography (LDCT) reconstruction.

We formulate the inverse problem $\mathbf{y} = \mathbf{A}\mathbf{x} + \sigma\boldsymbol{\epsilon}$, where $\mathbf{x}$ represents ground truth `RSNA` brain CT scans [Anouk Stein et al., 2019], $\mathbf{A}$ is the forward projection matrix (X-ray line integrals), $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{y}$ is the measurement sinogram. By zeroing the singular values of $\mathbf{A}$ below a threshold ξ and adding noise, we simulate low-dose sparse-view CT. We employ System-Embedded Diffusion Bridges [SDB; Sobieski et al., 2026], which solve linear Gaussian problems by embedding measurement parameters into the diffusion coefficients: $\mathbf{H}_t\mathbf{x} = \mathbf{A}^+\mathbf{A}\mathbf{x} + \alpha_t(\mathbf{I} - \mathbf{A}^+\mathbf{A})\mathbf{x}$ and $\boldsymbol{\Sigma}_t = \gamma_t\mathbf{A}^+\boldsymbol{\Sigma}{\mathbf{A}^+}^\top + \beta_t(\mathbf{I} - \mathbf{A}^+\mathbf{A})$, utilizing the Moore-Penrose pseudoinverse $\mathbf{A}^+$. Because SDB maps measurements $\mathbf{y}$ to reconstructions $\mathbf{x}$ via a matrix-valued diffusion process, it allows for direct

Table 2: Quantitative comparison of all sampling methods applied to SDB on the RSNA dataset. We include standard reconstruction metrics regarding perception and distortion. We highlight mAP and mROC, performance measures of the ResNet50 observer, as indicators of reconstruction errors and hallucinations, where higher values suggest improved anatomical consistency.

Method	FID ↓	IV ↑	DSV ↑	PSNR ↑	SSIM ↑	LPIPS ↓	mAP ↑	mROC ↑
Baseline	33.3	0.066	0.264	33.54	0.898	0.0387	0.27	0.85
DG	35.6	0.066	0.264	33.55	0.899	0.0388	0.31	0.86
AAM	35.6	0.066	0.264	33.54	0.898	0.0388	0.29	0.86
RODS^{CAS}	33.3	0.067	0.264	33.54	0.898	0.0388	0.29	0.86
RODS^{SAS}	33.3	0.066	0.263	33.55	0.898	0.0387	0.29	0.86
IQ	33.3	0.066	0.264	33.54	0.901	0.0388	0.31	0.86

integration of the evaluated hallucination reduction methods. Throughout this experiment, we strictly adhere to the original SDB experimental setup on RSNA [Sobieski et al., 2026].

Reducing misdiagnosis. In this context, structural hallucinations manifest as mischaracterizations of the patient condition, *e.g.*, fabricated synthetic pathology. To quantify diagnostic safety, we trained a ResNet50 observer to detect five hemorrhage types using 17,948 ground-truth RSNA slices. On a 512-slice validation split, the observer achieves 0.38 macro-averaged multi-label Average Precision (mAP) and 0.91 macro-averaged multi-label Area Under the ROC (mROC). When evaluated on baseline SDB reconstructions [Sobieski et al., 2026], performance drops to 0.27 mAP and 0.85 mROC, reflecting quality degradation and generative hallucinations.

We adapt all baseline mitigations to the SDB framework, tracking image quality and diversity (FID, IV, DSV), distortion (PSNR, SSIM, LPIPS) [Zhang et al., 2018, Blau and Michaeli, 2018], and diagnostic safety (mAP, mROC) (table 2). Because SDB strictly preserves the measurement range space, image and reconstruction quality metric variations are heavily constrained to the null space and remain mostly flat, although DG and AAM noticeably degrade FID. Diagnostic performance, however, reveals a sharp distinction. While all methods marginally raise mROC to 0.86, IQ and DG increase mAP by roughly 15% (reaching 0.31). Crucially, DG requires an independently trained classifier with direct access to ground-truth hemorrhage labels, granting it an inherent advantage. Despite operating entirely zero-shot, IQ matches DG's diagnostic gains while strictly preserving reconstruction quality. This underscores IQ's universal applicability and viability for broader inverse problem frameworks [Chung et al., 2023, Liu et al., 2023a, Luo et al., 2023, Garber and Tirer, 2024, Yue et al., 2024, Zhou et al., 2024].

Mitigating hallucinations in foundation models. Prior work on unsupervised hallucination in DMs [Oorloff et al., 2025, Triaridis et al., 2025, Tian et al., 2025] limited the evaluation to models trained on small-scale, homogeneous data, where hallucinations are more easily identifiable. A natural next step are DM-based foundation models (FMs) [Rombach et al., 2022] trained on large-scale heterogeneous data, where hallucination mitigation offers direct practical returns. However, the high complexity of FMs training data makes evaluation much more difficult, as hallucinations may appear within varying contexts and in more diversified forms, which is why current literature lacks appropriate benchmarks.

To at least partially address this gap, we perform a small-scale evaluation on Stable Diffusion XL [SDXL; Podell et al., 2024]. We first manually search for a detailed prompt that results in SDXL generating realistic hand images, while also producing clear hallucinations (*e.g.*, additional/missing fingers, nonanatomic shapes). We then reimplement both IQ and RODS (as the most direct competitor) to work directly in SDXL's latent space. After an initial hyperparameter search, both methods and unconditional generation were evaluated by 3 human annotators on randomly chosen set of 64 generated images. Table 3 provides initial evidence that IQ is once again able to mitigate hallucinations most effectively, reducing the baseline rate by around 30%, with RODS being able to reduce only at most 13%. Annotators reported no visible drop in quality across the entire comparison.

Table 3: Quantitative small-scale comparison of IQ, RODS and unconditional generation on SDXL.

Method	HR ↓
Baseline	77.1
RODS^{SAS}	74.7
RODS^{CAS}	67.5
IQ	53.1

This study shows both the potential of transferring IQ to large-scale models and highlights the need for further development, as many hallucinations remain. We attribute this gap primarily to

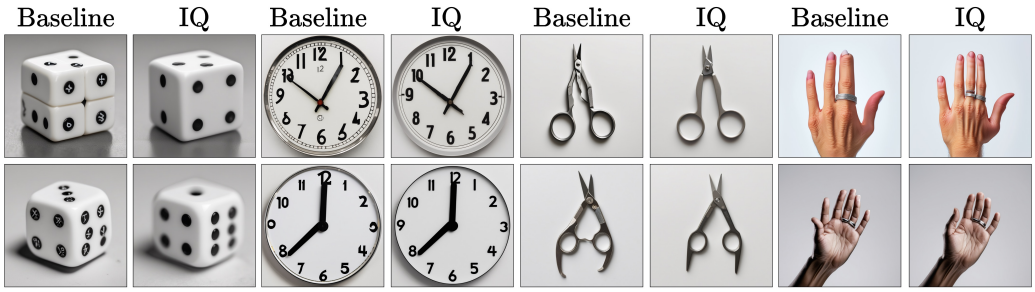

Figure 5: Qualitative comparison of hallucinated samples and their corresponding IQ corrections on SDXL. For each pair, a different setting of λ, t_1 and t_2 was used, with other hyperparameters fixed.

the heterogeneity of the data. As the model-induced manifold gets more complex and diverse, we hypothesize that hallucinations no longer occupy a single range of superfluous dimensions, but instead make it highly dependent on the surrounding data and their own complexity. For example, the same value of $\text{LID}_{\boldsymbol{\theta}}$ might both indicate a valid but complex generated sample (like a hand with a sophisticated pattern in the background) or an actual hallucination (additional finger in a hand placed in plain environment) due to high heterogeneity.

This line of reasoning is supported by our initial exploration into how IQ transfers across different hallucination types in SDXL. Figure 5 displays pairs of originally hallucinated samples and their corresponding IQ corrections, where for each pair a different hyperparameter configuration was used. Crucially, we were not able to identify a single setting that worked universally across all cases, but were able to easily tune λ, t_1 and t_2 for each case individually. These hyperparameters translate directly into how the underlying geometry is affected, highlighting that it differs across different data subdomains, but is influenced in the same way – the spurious dimensions are stripped, which corrects initially wrong samples. It is the heterogeneity of this data that forces per-case tuning, showcasing an important future work direction – once an objective fingerprint of hallucinations across different subdomains is found, it can be used to decide automatically on where (t_1, t_2) and how heavily (λ) a method such as IQ should be applied.

Ablation studies. Due to limited space, we continue the experimental evaluation in section B, providing ablation studies for each component of IQ (section B.8), runtime comparison (section B.9), qualitative examples (section B.14) and more.

7 Discussion and Limitations

Our work serves as a direct counterpart to Ross et al. [2025], who demonstrate how the collapse of LID values provides a direct indicator of *memorized* samples. These complementary findings offer empirical evidence for an intuitive understanding of LID: it serves as a proxy for the *creativity* of a DM, where hallucinations (the model being overly creative) and memorization (a lack of creativity) span the full generative spectrum.

Evaluation persists as the primary limitation of IQ and other hallucination reduction methods. While human annotators generally agree when identifying structural failures, disagreements remain common, highlighting the subjective nature of the task. Constructing fully objective, large-scale benchmarks with automated evaluation would streamline the development of more effective methods. Regarding the observed drops in DSV and IV, proxy metrics for diversity, in datasets like `11kHands` and `FFHQ`, our analysis in section B.6 indicates that they are caused by the removal of hallucinated samples, which artificially inflate these metrics; on the subset of correct samples, the diversity of IQ matches all baselines.

Furthermore, methods *intrinsic* to the model, such as RODS and IQ, still induce a moderate computational overhead. Discovering cheaper estimators for LID would lead to immediate practical improvements for IQ. Finally, while unconditional generation is the cornerstone of generative modeling, hallucinations pose the highest risk in conditional settings involving real-world decision-making and diagnosis [Antun et al., 2020], highlighting a critical direction for future investigation.

Acknowledgments and Disclosure of Funding

Work on this project is financially supported by the Polish National Science Centre PRELUDIUM BIS grant No. 2023/50/O/ST6/00301, the Foundation for Polish Science (FNP) grant 'Centre for Credible AI' No. FENG.02.01-IP.05-0058/24.

The computational resources for this work were provided by the Laboratory of Bioinformatics and Computational Genomics and the High Performance Computing Center of the Faculty of Mathematics and Information Science, Warsaw University of Technology. We also gratefully acknowledge Poland's High-performance Infrastructure PLGrid ACC Cyfronet AGH for providing computer facilities and support within computational grant no. PLG/2025/018330.

References

Mahmoud Afifi. 11k hands: gender recognition and biometric identification using a large dataset of hand images. In *Multimedia Tools and Applications*, 2019.

Sumukh K Aithal, Pratyush Maini, Zachary Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In *Advances in neural information processing systems*, 2024.

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. In *The Journal of Machine Learning Research*. JMLR, 2014.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023.

Eugene L Allgower and Kurt Georg. *Numerical continuation methods: an introduction*. Springer Science & Business Media, 2012.

MD Anouk Stein, Carol Wu, Chris Carr, George Shih, Jayashree Kalpathy-Cramer, Julia Elliott, kalpathy, Luciano Prevedello, MD Marc Kohli, Matt Lungren, Phil Culliton, Robyn Ball, and Safwan Halabi MD. Rsna intracranial hemorrhage detection. https://kaggle.com/competitions/rsna-intracranial-hemorrhage-detection, 2019. Kaggle.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. In *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 2020.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. In *IEEE transactions on pattern analysis and machine intelligence*. IEEE, 2013.

Mahesh Bhosale, Naresh Kumar Devulapally, Abdul Wasi, Chau Pham, Vishnu Suresh Lokhande, and David Doermann. Variance-guided score regularization for hallucination mitigation in diffusion models, 2026. URL https://openreview.net/forum?id=nY4nULFzDP.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018.

Dan G Cacuci, Mihaela Ionescu-Bujor, and Ionel Michael Navon. *Sensitivity and uncertainty analysis, volume II: applications to large-scale systems*. CRC press, 2005.

Yu Cao, Zengqun Zhao, Ioannis Patras, and Shaogang Gong. Temporal score analysis for understanding and correcting diffusion artifacts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7707–7716, 2025.

Barath Chandran.C, Srinivas Anumasa, and Dianbo Liu. Laplacian score sharpening for mitigating hallucination in diffusion models. In *arXiv*, 2025.

Hyunmin Cho, Donghoon Ahn, Susung Hong, Jee Eun Kim, Seungryong Kim, and Kyong Hwan Jin. Tag: Tangential amplifying guidance for hallucination-resistant diffusion sampling. In *arXiv*, 2025.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in neural information processing systems*, 2021.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=xQBRrtQM8u`.

Bradley Efron. Tweedie's formula and selection bias. In *Journal of the American Statistical Association*. Taylor & Francis, 2011.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. In *Journal of the American Mathematical Society*, 2016.

Shuai Fu, Jian Zhou, Qi Chen, Huang Jing, Huy Anh Nguyen, Xiaohan Liu, Zhixiong Zeng, Lin Ma, Quanshi Zhang, and Qi Wu. Counting hallucinations in diffusion models. In *arXiv*, 2025.

Tomer Garber and Tom Tirer. Image restoration by denoising diffusion models with iteratively preconditioned guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mark Goresky and Robert MacPherson. Stratified morse theory. In *Stratified Morse Theory*. Springer, 1988.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, 2020.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. In *Journal of Machine Learning Research*, 2005.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9307–9315, 2024.

Hamidreza Kamkari, Brendan L Ross, Rasa Hosseinzadeh, Jesse C Cresswell, and Gabriel Loaiza-Ganem. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in neural information processing systems*, 2022.

Seunghoi Kim, Chen Jin, Tom Diethe, Matteo Figini, Henry FJ Tregidgo, Asher Mullokandov, Philip Teare, and Daniel C Alexander. Tackling structural hallucination in image translation with local diffusion. In *European Conference on Computer Vision*. Springer, 2024.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=a-xFK8Ymz5J`.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*. PMLR, 2023.

Alice Ku, Matthew Tivnan, and Dufan Wu. Hallucination analysis of score-based diffusion models for ct denoising through spatial frequency decomposition. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2025. doi: 10.1109/ISBI60581.2025.10981211.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, 2004.

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

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I2SB: Image-to-Image Schrödinger Bridge. In *International Conference on Machine Learning*, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023b.

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

Rui Lu, Runzhe Wang, Kaifeng Lyu, Xitai Jiang, Gao Huang, and Mengdi Wang. Towards understanding text hallucination of diffusion models via local generation bias. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=SKW10XJlAI`.

Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7085–7093, 2024.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. In *International Conference on Machine Learning*, 2023.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024.

Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Handiffuser: Text-to-image generation with realistic hand appearances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2468–2479, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Trevine Oorloff, Yaser Yacoob, and Abhinav Shrivastava. Mitigating hallucinations in diffusion models through adaptive attention modulation. In *arXiv*, 2025.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. In *Advances in Neural Information Processing Systems*, volume 36, pages 24129–24142, 2023.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems*, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*. PMLR, 2021.

Weiming Ren, Raghav Goyal, Zhiming Hu, Tristan Ty Aumentado-Armstrong, Iqbal Mohomed, and Alex Levinshtein. Hallucination score: Towards mitigating hallucinations in generative image super-resolution. In *arXiv*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Brendan Leigh Ross, Hamidreza Kamkari, Tongzi Wu, Rasa Hosseinzadeh, Zhaoyan Liu, George Stein, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. A geometric framework for understanding memorization in generative models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aZ1gNJu8w0.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. In *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 2025.

Chuancheng Shi, Shiming Guo, Ke Shui, Yixiang Chen, and Fei Shen. Sgmhand: Structure-guided modulation for structure-aware hand inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Bartlomiej Sobieski, Matthew Tivnan, Yuang Wang, Siyeop yoon, Pengfei Jin, Dufan Wu, Quanzheng Li, and Przemyslaw Biecek. System-embedded diffusion bridge models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=cipx3rwfWp.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

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*, 2021.

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

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

Piotr Tempczyk, Rafał Michaluk, Lukasz Garncarek, Przemysław Spurek, Jacek Tabor, and Adam Golinski. Lidl: Local intrinsic dimension estimation using approximate likelihood. In *International Conference on Machine Learning*, pages 21205–21231. PMLR, 2022.

Yiqi Tian, Pengfei Jin, Mingze Yuan, Na Li, Bo Zeng, and Quanzheng Li. RODS: Robust optimization inspired diffusion sampling for detecting and reducing hallucination in generative models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=fhuqIxoPcr`.

Matthew Tivnan, Siyeop Yoon, Zhennong Chen, Xiang Li, Dufan Wu, and Quanzheng Li. Hallucination index: An image quality metric for generative reconstruction models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 449–458. Springer, 2024.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Kostas Triaridis, Alexandros Graikos, Aggelina Chatziagapi, Grigorios G Chrysos, and Dimitris Samaras. Mitigating diffusion model hallucinations with dynamic guidance. In *arXiv*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017.

Pascal Vincent. A connection between score matching and denoising autoencoders. In *Neural computation*. MIT Press, 2011.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Shanshan Wang, Xuanru Zhou, Cheng Li, Shuqiang Wang, Ye Li, Tao Tan, and Hairong Zheng. Generative artificial intelligence in medical imaging: Foundations, progress, and clinical translation. In *Research*, volume 8, page 1029. AAAS, 2025.

Christopher Wewer, Bart Pogodzinski, Bernt Schiele, and Jan Eric Lenssen. Spatial reasoning with denoising models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Eric Yeats, Aaron Jacobson, Darryl Hannan, Yiran Jia, Timothy Doster, Henry Kvinge, and Scott Mahan. A connection between score matching and local intrinsic dimension. In *NeurIPS 2025 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2025.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Conghan Yue, Zhengwei Peng, Junlong Ma, Shiyan Du, Pengxu Wei, and Dongyu Zhang. Image restoration through generalized ornstein-uhlenbeck bridge. In *International Conference on Machine Learning*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *International Conference on Learning Representations*, 2024.

Appendix

A Theoretical results **17**

A.1 Proposition 1 17

A.2 Proposition 2 18

A.3 Theorem 1 19

A.4 Theorem 2 21

B Extended experiments **22**

B.1 Experimental setup 22

B.2 User study 22

B.3 Filter performance 22

B.4 LID performance curves 23

B.5 Extended results 24

B.6 Disentangling diversity and hallucinations 24

B.7 Hallucination reduction in flow-based models 25

B.8 Ablation studies 26

B.9 Runtime comparison 28

B.10 Hyperparameters 29

B.11 SDXL evaluation details 29

B.12 Pseudocode 30

B.13 Resources 30

B.14 Qualitative examples 30

A Theoretical results

A.1 Proposition 1

Proposition 1. *Let $\mathcal{G}_{\boldsymbol{\theta}}$ be the deterministic generative mapping of a diffusion model from the latent noise space to the induced data manifold $\mathcal{M}_{\boldsymbol{\theta}}$, evaluated at an initial state $\mathbf{x}_1$. Assume the perturbation scale $\beta > 0$ is sufficiently small such that a first-order approximation holds. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ denote the singular values of the generator's Jacobian $\nabla_{\mathbf{x}_1}\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_1)$. Let $\mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0)$ denote the local intrinsic dimensionality of the generated data point $\mathbf{x}_0 = \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_1)$ with respect to the induced manifold $\mathcal{M}_{\boldsymbol{\theta}}$. The Local Manifold Instability (*LMI*) is approximately equal to the scaled sum of its top $\lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor$ squared singular values:*

$$\mathrm{LMI}(\mathbf{x}_1) \approx \beta^2 \sum_{i=1}^{\lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor} \sigma_i^2. \tag{11}$$

Proof. By definition, the Local Manifold Instability (LMI) at an initial state $\mathbf{x}_1$ is given by $\mathrm{LMI}(\mathbf{x}_1) \triangleq \mathrm{Tr}(\mathrm{Cov}_{\boldsymbol{\epsilon}}(\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_1 + \boldsymbol{\epsilon})))$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \beta^2\mathbf{I})$ for a perturbation scale $\beta > 0$. Applying a first-order Taylor expansion to the generative mapping $\mathcal{G}_{\boldsymbol{\theta}}$ around $\mathbf{x}_1$ yields:

$$\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_1 + \boldsymbol{\epsilon}) = \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_1) + \nabla_{\mathbf{x}_1}\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_1)\boldsymbol{\epsilon} + \mathcal{O}(\|\boldsymbol{\epsilon}\|_2^2). \tag{12}$$

Substituting this expansion into the covariance expression and recognizing that $\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_1)$ is a deterministic constant with respect to $\boldsymbol{\epsilon}$, it vanishes from the computation. Letting $\mathbf{J} = \nabla_{\mathbf{x}_1}\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_1)$, we obtain:

$$\mathrm{Cov}_{\boldsymbol{\epsilon}}(\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}_1 + \boldsymbol{\epsilon})) = \mathbf{J}\,\mathrm{Cov}_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})\mathbf{J}^\top + \mathcal{O}(\beta^3) = \beta^2\mathbf{J}\mathbf{J}^\top + \mathcal{O}(\beta^3). \tag{13}$$

Applying the trace operator to calculate the total variance (LMI):

$$\mathrm{LMI}(\mathbf{x}_1) = \mathrm{Tr}(\beta^2 \mathbf{J}\mathbf{J}^\top) + \mathcal{O}(\beta^3) = \beta^2 \left\|\mathbf{J}\right\|_F^2 + \mathcal{O}(\beta^3). \tag{14}$$

For a sufficiently small noise scale β, the linear regime dominates and the higher-order remainder $\mathcal{O}(\beta^3)$ can be safely neglected, yielding the first-order approximation:

$$\mathrm{LMI}(\mathbf{x}_1) \approx \beta^2 \left\|\mathbf{J}\right\|_F^2. \tag{15}$$

Our next objective is to relate this squared Frobenius norm to the effective spectral properties of the mapping. The squared Frobenius norm of any matrix is algebraically equivalent to the sum of all its squared singular values:

$$\left\|\mathbf{J}\right\|_F^2 = \sum_{i=1}^{n} \sigma_i^2. \tag{16}$$

We address the geometric dimensionality of the generation process. The generator $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}$ continuously maps the n-dimensional ambient noise space into a lower-dimensional induced data manifold $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$. The local intrinsic dimensionality at the specific generated data point $\mathbf{x}_0$, denoted as $\mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0)$, represents the effective number of independent, structurally significant directions in which the model's output varies.

While deep neural networks practically possess full-rank Jacobians (meaning all $\sigma_i > 0$), assumption 1 implies that the mapping highly contracts the space along off-manifold directions. Consequently, the singular values corresponding to the remaining $n - \lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor$ directions are infinitesimally small, representing negligible off-manifold numerical noise rather than structurally meaningful degrees of freedom. By treating these tail singular values as negligible ($\sigma_i \approx 0$ for $i > \lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor$), we can approximate the full spectral sum by truncating it at the effective dimensionality:

$$\sum_{i=1}^{n} \sigma_i^2 \approx \sum_{i=1}^{\lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor} \sigma_i^2. \tag{17}$$

Finally, substituting this spectral truncation back into eq. (15) yields the completed theoretical approximation:

$$\mathrm{LMI}(\mathbf{x}_1) \approx \beta^2 \sum_{i=1}^{\lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor} \sigma_i^2. \tag{18}$$

This demonstrates that, under a localized linear regime, the spatial instability of the generator is closely approximated by the variance of the injected noise scaled by the structurally significant spectral mass of the local manifold. □

A.2 Proposition 2

Proposition 2. *Let $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$ be the $\mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0)$-dimensional manifold induced by a fully trained diffusion model parameterized by $\boldsymbol{\theta}$, embedded in an n-dimensional ambient space, with probability density $p_{\boldsymbol{\theta}}(\mathbf{x}_0)$. Assume $p_{\boldsymbol{\theta}}(\mathbf{x}_0)$ is locally uniform on $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$ with negligible curvature at the limit $t \to 0$. Let the forward process be defined by an arbitrary drift matrix $\mathbf{H}_t$ and a strictly positive-definite covariance matrix $\boldsymbol{\Sigma}_t$. As $t \to 0$, the expected DSM loss evaluates exactly to the local intrinsic dimension $\mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0)$, and the expected ISM loss evaluates exactly to $-\frac{1}{2}(n - \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0))$.*

Proof. Let the forward diffusion state be $\mathbf{x}_t = \mathbf{H}_t \mathbf{x}_0 + \boldsymbol{\Sigma}_t^{\frac{1}{2}} \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{x}_0 \in \boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$. By Tweedie's formula applied to the model's induced distribution, the predicted score function $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t)$ defines the model's posterior mean estimate $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)$ via:

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t) = -\boldsymbol{\Sigma}_t^{-1}(\mathbf{x}_t - \mathbf{H}_t \hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)). \tag{19}$$

The predicted noise is defined as $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t) \triangleq -\boldsymbol{\Sigma}_t^{\frac{1}{2}} \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t)$. Substituting the predicted score yields:

$$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t) = \boldsymbol{\Sigma}_t^{-\frac{1}{2}}(\mathbf{x}_t - \mathbf{H}_t \hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)). \tag{20}$$

To evaluate the DSM expectation $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t)\|_2^2\right]$, we express the target forward noise as $\boldsymbol{\epsilon} = \boldsymbol{\Sigma}_t^{-\frac{1}{2}}(\mathbf{x}_t - \mathbf{H}_t\mathbf{x}_0)$. The residual error inside the norm becomes:

$$\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t) = \boldsymbol{\Sigma}_t^{-\frac{1}{2}}\mathbf{H}_t(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)-\mathbf{x}_0). \tag{21}$$

As $t \to 0$, the model's posterior mean $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)$ acts as the optimal projection of the noisy state back onto the locally flat model-induced manifold $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$. Because the forward process introduces spatial drift $\mathbf{H}_t$ and covariance $\boldsymbol{\Sigma}_t$, the projected state minimizes the Mahalanobis distance $\|\mathbf{x}_t - \mathbf{H}_t\mathbf{x}\|^2_{\boldsymbol{\Sigma}_t^{-1}}$ for $\mathbf{x} \in \boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$.

Let $\mathcal{T}_{\mathbf{x}_0}\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$ be the tangent space of the model's manifold at $\mathbf{x}_0$. The projection onto this mapped tangent space $\mathbf{H}_t\mathcal{T}_{\mathbf{x}_0}\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$ under the Mahalanobis inner product $\langle \mathbf{u}, \mathbf{v}\rangle_{\boldsymbol{\Sigma}_t^{-1}} = \mathbf{u}^\top\boldsymbol{\Sigma}_t^{-1}\mathbf{v}$ is governed by an oblique projection matrix $\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}$. Therefore, the deviation of the model's posterior mean mapped through the drift is exactly the oblique projection of the forward perturbation:

$$\mathbf{H}_t(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)-\mathbf{x}_0) \approx \mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}(\mathbf{x}_t-\mathbf{H}_t\mathbf{x}_0) = \mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\epsilon}. \tag{22}$$

Substituting this geometric relation back into eq. (21), the DSM expectation becomes:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_{\text{DSM}}(\mathbf{x}_0,t,\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\Sigma}_t^{-\frac{1}{2}}\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\epsilon}\right\|_2^2\right]. \tag{23}$$

Define the transformed operator $\mathbf{M}_{\mathcal{T}} \triangleq \boldsymbol{\Sigma}_t^{-\frac{1}{2}}\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}\boldsymbol{\Sigma}_t^{\frac{1}{2}}$. This operator represents the tangent space projection mapped into the whitened isotropic noise space. We prove $\mathbf{M}_{\mathcal{T}}$ is a standard Euclidean orthogonal projector by checking symmetry. By definition, $\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}$ is self-adjoint with respect to the Mahalanobis inner product, meaning $(\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t})^\top\boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_t^{-1}\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}$. Thus:

$$\mathbf{M}_{\mathcal{T}}^\top = \boldsymbol{\Sigma}_t^{\frac{1}{2}}(\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t})^\top\boldsymbol{\Sigma}_t^{-\frac{1}{2}} = \boldsymbol{\Sigma}_t^{\frac{1}{2}}(\boldsymbol{\Sigma}_t^{-1}\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}\boldsymbol{\Sigma}_t)\boldsymbol{\Sigma}_t^{-\frac{1}{2}} = \boldsymbol{\Sigma}_t^{-\frac{1}{2}}\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}\boldsymbol{\Sigma}_t^{\frac{1}{2}} = \mathbf{M}_{\mathcal{T}}. \tag{24}$$

Because $\mathbf{M}_{\mathcal{T}}$ is symmetric ($\mathbf{M}_{\mathcal{T}}^\top = \mathbf{M}_{\mathcal{T}}$) and idempotent ($\mathbf{M}_{\mathcal{T}}^2 = \mathbf{M}_{\mathcal{T}}$), it is a valid Euclidean orthogonal projector. Its rank is strictly preserved under similarity transformations and the non-singular linear mapping $\mathbf{H}_t$, meaning $\text{rank}(\mathbf{M}_{\mathcal{T}}) = \text{rank}(\mathbf{P}_{\mathcal{T}}^{\boldsymbol{\Sigma}_t,\mathbf{H}_t}) = \text{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0)$.

Applying the standard identity for the expected squared norm of an orthogonally projected standard Gaussian, we evaluate the trace:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_{\text{DSM}}(\mathbf{x}_0,t,\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\boldsymbol{\epsilon}^\top\mathbf{M}_{\mathcal{T}}^\top\mathbf{M}_{\mathcal{T}}\boldsymbol{\epsilon}\right] = \text{Tr}(\mathbf{M}_{\mathcal{T}}) = \text{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0). \tag{25}$$

Finally, by Stein's Lemma, the generalized score matching objectives satisfy the exact algebraic identity $\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_{\text{DSM}}(\mathbf{x}_0,t,\boldsymbol{\theta})] = n + 2\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_{\text{ISM}}(\mathbf{x}_0,t,\boldsymbol{\theta})]$. Rearranging this relationship and substituting the evaluated DSM expectation directly yields:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\mathcal{L}_{\text{ISM}}(\mathbf{x}_0,t,\boldsymbol{\theta})] = -\frac{1}{2}(n - \text{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0)), \tag{26}$$

proving geometric and algebraic consistency for arbitrary drift and covariance schedules exclusively within the model-induced geometry. □

A.3 Theorem 1

We include a more general formulation of theorem 1 that is not limited to isotropic noise as $t \to 0$. This allows us to provide theoretical guarantees also for frameworks like SDB.

Theorem 1. *Let $\mathbf{x}_0 = \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1)$. Assume that $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1)$ is decomposed as $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}} = \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{\leq\tau} \circ \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{>\tau}$ for sufficiently small time $\tau > 0$, where $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{>\tau} : \mathbf{x}_1 \mapsto \mathbf{x}_\tau$ and $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{\leq\tau} : \mathbf{x}_\tau \mapsto \mathbf{x}_0$. Let the singular values of the macroscopic flow Jacobian $\nabla_{\mathbf{x}_1}\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{>\tau}(\mathbf{x}_1)$ be monotonically ordered as $\sigma_1^{>\tau} \geq \sigma_2^{>\tau} \geq \cdots \geq \sigma_n^{>\tau} \geq 0$. Let $K = \left\|\nabla_{\mathbf{x}_\tau}\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{\leq\tau}(\mathbf{x}_\tau)\right\|_2^2 \geq 1$ represent the squared spectral norm of the terminal oblique projection. For sufficiently small $\beta > 0$, it is true that*

$$\text{LMI}(\mathbf{x}_1) \lessapprox K\beta^2 \sum_{i=1}^{\lfloor \text{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau))\rfloor} (\sigma_i^{>\tau})^2. \tag{27}$$

Proof. Applying the multivariate chain rule to the decomposed generative flow $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}} = \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{\leq\tau} \circ \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{>\tau}$, the total generator Jacobian $\mathbf{J} = \nabla_{\mathbf{x}_1} \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}(\mathbf{x}_1)$ is the matrix product of the stage-wise Jacobians:

$$\mathbf{J} = \nabla_{\mathbf{x}_\tau} \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{\leq\tau}(\mathbf{x}_\tau) \nabla_{\mathbf{x}_1} \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{>\tau}(\mathbf{x}_1) = \mathbf{J}_{\leq\tau} \mathbf{J}_{>\tau}. \tag{28}$$

Recall from proposition 1 the first-order approximation for LMI for sufficiently small β, which evaluates to the scaled squared Frobenius norm of the total Jacobian: $\mathrm{LMI}(\mathbf{x}_1) \approx \beta^2 \left\|\mathbf{J}\right\|_F^2$. Substituting the chain rule expansion into this norm yields:

$$\left\|\mathbf{J}\right\|_F^2 = \left\|\mathbf{J}_{\leq\tau} \mathbf{J}_{>\tau}\right\|_F^2. \tag{29}$$

For sufficiently small τ, the terminal integration step $\boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{\leq\tau}$ behaves as a projection onto the induced manifold $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$. In the general case of anisotropic terminal noise (such as the embedded covariance structure in SDB), this operation acts as an *oblique* projector rather than a standard Euclidean orthogonal projector.

Consequently, its Jacobian $\mathbf{J}_{\leq\tau}$ possesses exactly $\lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor$ non-zero singular values. We denote these terminal singular values as $\sigma_i^{\leq\tau}$ and bound them by the squared spectral norm of the oblique projector: $(\sigma_i^{\leq\tau})^2 \leq \left\|\mathbf{J}_{\leq\tau}\right\|_2^2 \triangleq K$. The remaining $n - \lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor$ singular values are strictly equal to 0 (annihilating variance along normal directions).

To decouple the two generative regimes, we apply the singular value product inequality (von Neumann trace inequality) to bound the Frobenius norm of the product. The sum of the squared singular values of the product is bounded by the sum of the products of their individual squared singular values:

$$\left\|\mathbf{J}_{\leq\tau} \mathbf{J}_{>\tau}\right\|_F^2 \leq \sum_{i=1}^{n} (\sigma_i^{\leq\tau})^2 (\sigma_i^{>\tau})^2. \tag{30}$$

Substituting the bounded spectrum of the terminal oblique projector ($\sigma_i^{\leq\tau} \leq \sqrt{K}$ for $i \leq \lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor$ and 0 otherwise) acts as a strict algebraic cutoff for the summation, scaled by the projection constant K:

$$\sum_{i=1}^{n} (\sigma_i^{\leq\tau})^2 (\sigma_i^{>\tau})^2 \leq K \sum_{i=1}^{\lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor} (\sigma_i^{>\tau})^2. \tag{31}$$

Substituting this strict upper bound back into the LMI approximation provides an intermediate theoretical guarantee:

$$\mathrm{LMI}(\mathbf{x}_1) \lessapprox K\beta^2 \sum_{i=1}^{\lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0) \rfloor} (\sigma_i^{>\tau})^2. \tag{32}$$

Next, we address the topological equivalence of evaluating the intrinsic dimensionality at the unobserved terminal state $\mathbf{x}_0$ versus the predicted posterior mean $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau)$. At any intermediate time $\tau > 0$, the state $\mathbf{x}_\tau$ is corrupted by forward process variance. Consequently, $\mathbf{x}_\tau$ has full support in the n-dimensional ambient space, meaning $\mathbf{x}_\tau \notin \boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$.

To evaluate the topological complexity of the target manifold from $\mathbf{x}_\tau$, we compute the Tweedie estimate derived from the predicted score function, $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau)$. By the fundamental theorem of estimation, this conditional expectation acts as the Minimum Mean Square Error (MMSE) estimator. Because $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$ generally possesses curvature, the expectation of points on the manifold strictly resides within the convex hull of $\boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$, meaning $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau) \notin \boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}$ in the general case.

However, we establish asymptotic equivalence between this estimation and the terminal generative mapping. The terminal state is obtained by integrating the Probability Flow ODE from τ to 0: $\mathbf{x}_0 = \boldsymbol{\mathcal{G}}_{\boldsymbol{\theta}}^{\leq\tau}(\mathbf{x}_\tau)$. As the integration interval $\tau \to 0$, the distribution concentrates locally, the exact ODE integration converges to the analytic score-based projection, and the Euclidean distance to the manifold vanishes:

$$\lim_{\tau \to 0} \mathrm{dist}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau), \boldsymbol{\mathcal{M}}_{\boldsymbol{\theta}}) = 0. \tag{33}$$

This implies that for a sufficiently small τ, the Tweedie estimate $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau)$ and the true terminal state $\mathbf{x}_0$ reside within the exact same infinitesimal local neighborhood.

Under the assumption that $\mathcal{M}_{\boldsymbol{\theta}}$ is a smooth manifold, the local intrinsic dimensionality $\mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x})$ is a locally constant topological invariant. Because $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau)$ approaches $\mathbf{x}_0$ sufficiently close, their geometric dimensionalities strictly align:

$$\mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau)) = \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0). \tag{34}$$

Substituting this topological equivalence into the intermediate bound (eq. (32)) yields the completed operational approximation:

$$\mathrm{LMI}(\mathbf{x}_1) \lessapprox K\beta^2 \sum_{i=1}^{\lfloor \mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_\tau)) \rfloor} (\sigma_i^{>\tau})^2. \tag{35}$$

□

A.4 Theorem 2

Theorem 2. *Let $p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)$ be the model-induced marginal density at time $t \leq \tau$. Assume the network parameterizes the exact score of its induced distribution, such that $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \log p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)$. We define the ideal geometrically-guided distribution by re-weighting the marginal over the true terminal states: $p_t^{\boldsymbol{\theta},\lambda_t}(\mathbf{x}_t) \propto p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)\mathbb{E}_{p^{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_t)}[\exp(-\lambda_t \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0))]$. Assume τ is sufficiently small such that the posterior variance $\sigma_t^2 \to 0$. Under these conditions, guiding the reverse diffusion with the modified score $\tilde{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{x}_t) = \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t) - \lambda_t \nabla_{\mathbf{x}_t} \mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t))$ is mathematically equivalent to sampling from the ideal distribution $p_t^{\boldsymbol{\theta},\lambda_t}(\mathbf{x}_t)$, shifting probability mass toward lower-dimensional target strata.*

Proof. We define the ideal guided target density by re-weighting the model-induced marginal $p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)$ with the expected terminal energy:

$$p_t^{\boldsymbol{\theta},\lambda_t}(\mathbf{x}_t) = \frac{1}{Z_t^{\boldsymbol{\theta}}} p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)\mathbb{E}_{p^{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_t)}\left[\exp(-\lambda_t \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0))\right], \tag{36}$$

where $Z_t^{\boldsymbol{\theta}}$ is the partition function.

Taking the logarithm and differentiating with respect to $\mathbf{x}_t$ yields the ideal modified score:

$$\nabla_{\mathbf{x}_t} \log p_t^{\boldsymbol{\theta},\lambda_t}(\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \log p_t^{\boldsymbol{\theta}}(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log \mathbb{E}_{p^{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_t)}\left[\exp(-\lambda_t \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0))\right]. \tag{37}$$

To render this tractable, we analyze the domain $t \leq \tau$. As the noise scale diminishes ($\sigma_t^2 \to 0$), the model's posterior distribution $p^{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_t)$ concentrates into a Dirac delta distribution centered at the Tweedie estimate $\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)$. Consequently, the log-expectation of the energy converges exactly to the energy of the expectation:

$$\lim_{\sigma_t^2 \to 0} \log \mathbb{E}_{p^{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_t)}\left[\exp(-\lambda_t \mathrm{LID}_{\boldsymbol{\theta}}(\mathbf{x}_0))\right] = -\lambda_t \mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)). \tag{38}$$

While the theoretical intrinsic dimension of a smooth manifold is a locally constant integer (which would yield a trivial zero gradient), our operational definition of $\mathrm{LID}_{\boldsymbol{\theta}}$ relies on the generalized score-matching estimator $\mathcal{L}_{\mathrm{DSM}}(\mathbf{x}, t, \boldsymbol{\theta})$. Because this estimator evaluates the model's spatial error, it is a continuous, non-linear, and differentiable function of its inputs, parameterized by the neural network $\boldsymbol{\theta}$. Therefore, applying the spatial gradient operator yields a well-defined, non-zero vector field:

$$\nabla_{\mathbf{x}_t} \mathcal{E}(\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)) \approx \nabla_{\mathbf{x}_t} \mathcal{L}_{\mathrm{DSM}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t), t, \boldsymbol{\theta}). \tag{39}$$

Substituting this limit and the assumed model score function ($\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \log p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)$) into the target gradient yields the operational guidance rule:

$$\tilde{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{x}_t) = \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t) - \lambda_t \nabla_{\mathbf{x}_t} \mathrm{LID}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_0^{\boldsymbol{\theta}}(\mathbf{x}_t)). \tag{40}$$

By substituting $\tilde{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{x}_t)$ into the reverse probability flow ODE, the generated trajectory simulates sampling from the ideal terminal-weighted distribution $p_t^{\boldsymbol{\theta},\lambda_t}(\mathbf{x}_t)$. Because this guided distribution is multiplicatively bounded by the original prior $p_t^{\boldsymbol{\theta}}(\mathbf{x}_t)$, the integration strictly preserves the topological validity dictated by the model-induced distribution while explicitly minimizing the local intrinsic dimension of the expected final sample via the continuous gradient signal. □

B Extended experiments

B.1 Experimental setup

As the domain of hallucination reduction currently lacks a single unified evaluation scheme, we reimplement all of the methods within our codebase. For DG, we train 6 noisy classifiers from scratch, one for each dataset, using ResNets of different sizes adapted to the complexity of the data for image datasets and a 3-layer MLP for `GaussianGrid`. For AAM, we adapt the temperature gradient descent scheme for all DMs and train a separate PatchCore anomaly detection model [Roth et al., 2022] for each dataset. To find optimal hyperparameters, we perform a large grid search for AAM, DG and RODS on all datasets, except `11kHands`, `FFHQ` and `AFHQV2` for RODS, where we use the originally proposed ones [Tian et al., 2025]. We report the hyperparameters of all methods in section B.10.

B.2 User study

For each of the five image datasets, the user study consisted of two subsequent phases: the *calibration* phase and the *labeling* phase. The former involved presenting 128 images from the original dataset in batches, aiming to calibrate the user's perception of structures and correlations that appear in the true data distribution. Although this phase might initially seem redundant, fig. 6 presents several examples of hand images from the `11kHands` dataset that contain atypical hand and finger placements, highlighting the diversity and complex relationships that should not be labeled as hallucinations. Hence, this phase ensures that annotators are aware of what a structural hallucination is in each considered case.

The labeling phase involved presenting the annotator with a grid of images, each generated with a different sampling method from the same starting point $\mathbf{x}_1$ of the diffusion process. The placement of all methods was fixed across grids, and the study was blinded; *i.e.*, the annotators were not given the names of the methods, which were replaced by `Image n` captions, with `n` being the method index (see fig. 6).

The user study relied on two independent annotators, both of whom were machine learning researchers. The task was first briefly described to them to ensure that the term *structural hallucinations* was understood correctly. At each iteration of the labeling phase, the annotators first had to indicate the indices of images that contained a structural hallucination (which could also be none) and then (for natural image datasets) the indices of images that were of the highest quality according to their own perception.

Because the labeling process was highly time-consuming, the study utilized a different number of images for each dataset, chosen a priori based on our visual assessment and the subjective advantage of IQ. This approach ensured that statistically significant results would be obtained (with a high probability) using a minimal number of images to reduce the burden on the annotators, which could have otherwise negatively impacted the study. The numbers of starting points (*i.e.*, the number of grids presented to each annotator) were as follows: 128 for `11kHands`, 1024 for `AFHQV2` and `FFHQ` each, 512 for `MNIST` and `SimpleShapes` each. Each annotator spent approximately 8 hours in total (across two sessions) labeling all of the provided samples.

At last, we note that the initial empirical investigation (section 3) uses the labels provided by one of the annotators on the `11kHands` dataset.

B.3 Filter performance

We compare three filters (TVF, LMI, and LID) in terms of separability and classification performance of correct and hallucinated samples in table 4. For hyperparameters of each filter, see table 5. LID consistently outperforms both TVF and LMI in almost all cases. LMI outperforms TVF in all cases, except for PR AUC and Cohen's d on `SimpleShapes`, and outperforms LID in some cases on `AFHQV2` and `MNIST`.

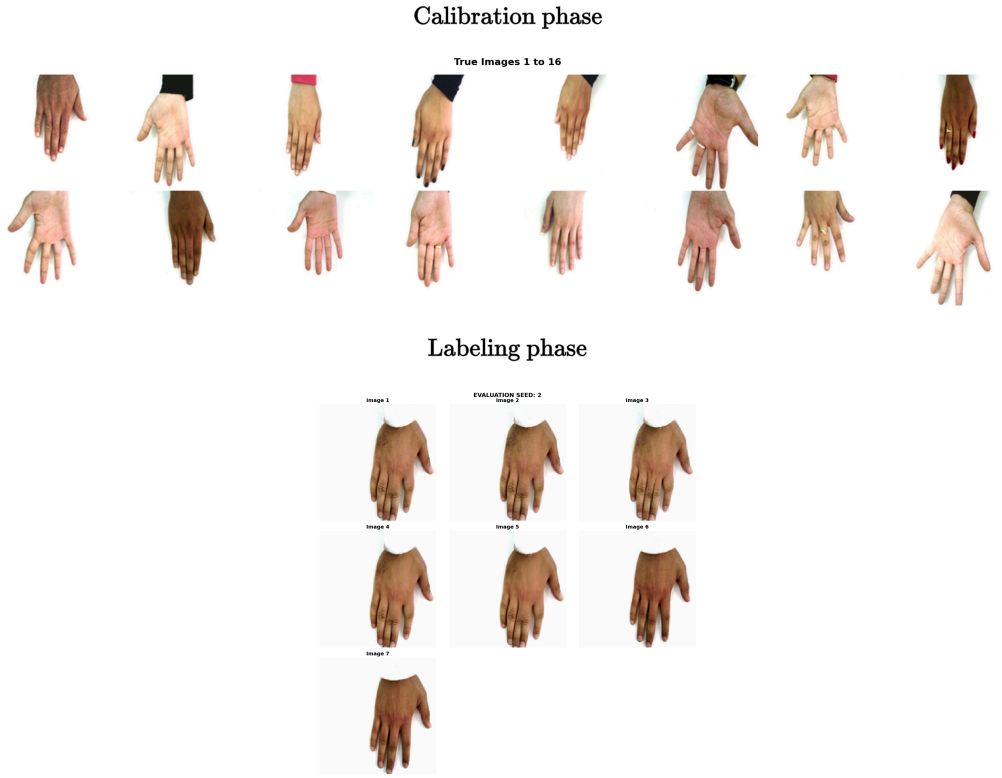

Figure 6: Examples of the user study interface from the `11kHands` dataset, depicting calibration and labeling phases.

Table 4: Extended comparison of three filters (TVF, LMI, and LID) in terms of separability and classification performance on natural and synthetic datasets.

	`AFHQV2`			`FFHQ`		
Method	PR AUC ↑	ROC AUC ↑	Cohen's d ↑	PR AUC ↑	ROC AUC ↑	Cohen's d ↑
TVF	0.12	0.55	0.18	0.16	0.53	0.17
LMI	0.16	0.66	0.31	0.17	0.60	0.17
LID	0.22	0.57	0.41	0.20	0.60	0.43

	`MNIST`			`SimpleShapes`		
Method	PR AUC ↑	ROC AUC ↑	Cohen's d ↑	PR AUC ↑	ROC AUC ↑	Cohen's d ↑
TVF	0.58	0.60	0.40	0.67	0.58	0.34
LMI	0.62	0.67	0.57	0.60	0.59	0.23
LID	0.63	0.66	0.42	0.67	0.61	0.34

B.4 LID performance curves

We visualize how the classification and separability performance of the LID filter changes across time on all image datasets in fig. 7. Crucially, the optimal timestep is very small in all cases, highlighting the importance of the theoretical assumptions.

Table 5: Summary of hyperparameters for the three evaluated filters (LID, LMI, and TVF) across natural and synthetic datasets. Time variables (t, t_1, t_2) are normalized to the $[0, 1]$ interval as decimal fractions.

Dataset	LID	LMI	TVF	
	t	β	t_1	t_2
`11kHands`	0.05	2.0	0.05	0.125
`FFHQ`	0.00625	2.0	0.05	0.1
`AFHQV2`	0.0125	2.0	0.9	0.95
`MNIST`	0.001	0.1	0.155	0.158
`SimpleShapes`	0.37	5.0	0.063	0.222

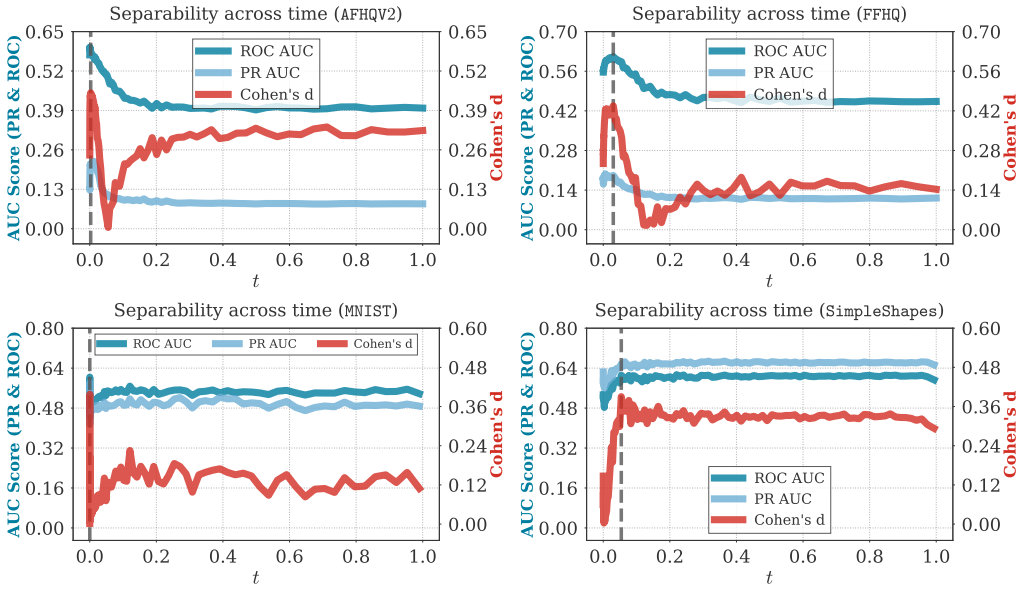

Figure 7: Performance curves of the LID filter across diffusion timesteps. The optimal timestep for classification and separability metrics is indicated by a vertical line in each case.

B.5 Extended results

We provide the average HR and UP in all considered cases, together with 95% CIs, in table 6. These results are based on the annotations of two independent experts (as mentioned in section B.2) using varying sample sizes for each dataset (for reasons previously described): 128 for `11kHands`, 1024 for `AFHQV2` and `FFHQ` each, and 512 for `MNIST` and `SimpleShapes` each. The results for `GaussianGrid` are based on 16384 samples for each method.

We also note that we generate the `SimpleShapes` dataset procedurally by using the implementation of Aithal et al. [2024], which divides a black 64×64 plane into three vertical bands of equal width and then independently inserts a triangle (leftmost band), a square (middle band) and a rhomb (rightmost band), each with 50% probability, into a random position within its band. Due to this procedural approach, we define hallucinations as improper placement of an object (*e.g.*, a triangle in the middle band) or the generation of more than 3 objects.

B.6 Disentangling diversity and hallucinations

Because IQ exhibits drops in proxy diversity metrics on several datasets (*e.g.*, `11kHands`), a natural question is whether these drops stem solely from the removal of invalid hallucinated samples, or whether they also limit the diversity of valid outputs. To investigate this, we recompute both diversity metrics on a subset of samples limited purely to the correct non-hallucinated cases, based on the annotators' labels from section B.2. To eliminate the effect of sample size, we limit the number of

Table 6: Detailed 95% confidence intervals for User Preference (UP) and Hallucination Ratio (HR) evaluated on high-resolution and synthetic datasets. Values are formatted as Mean [Lower Bound, Upper Bound].

	11kHands		FFHQ		AFHQV2	
Method	UP (%) ↑	HR (%) ↓	UP (%) ↑	HR (%) ↓	UP (%) ↑	HR (%) ↓
Baseline	39.8 [34.8, 44.9]	29.3 [24.1, 35.1]	45.3 [42.8, 47.9]	8.2 [7.0, 9.4]	41.8 [38.5, 45.0]	6.9 [5.9, 8.1]
DG	39.5 [34.6, 44.3]	29.7 [24.4, 35.6]	44.1 [41.3, 47.0]	9.8 [8.6, 11.1]	40.2 [36.4, 44.0]	8.6 [7.5, 9.9]
AAM	40.6 [35.7, 45.5]	29.3 [24.1, 35.1]	45.7 [43.2, 48.2]	10.2 [9.0, 11.6]	42.2 [39.0, 45.4]	14.3 [12.9, 15.9]
RODS-CAS	40.2 [35.3, 45.1]	25.8 [20.8, 31.5]	45.3 [42.8, 47.9]	7.7 [6.6, 8.9]	42.2 [39.0, 45.4]	6.9 [5.9, 8.1]
RODS-SAS	41.0 [35.8, 46.2]	29.7 [24.4, 35.6]	45.3 [42.8, 47.9]	8.2 [7.0, 9.4]	42.2 [39.0, 45.4]	6.6 [5.6, 7.7]
IQ	68.0 [63.1, 72.8]	9.0 [6.1, 13.1]	46.1 [43.7, 48.4]	4.2 [3.4, 5.1]	42.6 [39.5, 45.7]	5.9 [4.9, 7.0]

	MNIST	SimpleShapes	GaussianGrid
Method	HR (%) ↓	HR (%) ↓	HR (%) ↓
Baseline	37.3 [33.2, 41.6]	25.8 [19.0, 34.0]	20.2 [19.5, 20.8]
DG	23.6 [20.2, 27.5]	27.3 [20.4, 35.6]	11.1 [10.6, 11.6]
AAM	34.8 [30.8, 39.0]	55.5 [46.8, 63.8]	-
RODS-CAS	41.8 [37.6, 46.1]	30.5 [23.2, 38.9]	10.4 [9.9, 10.9]
RODS-SAS	55.3 [50.9, 59.5]	28.9 [21.8, 37.3]	19.4 [18.8, 20.0]
IQ	10.2 [7.8, 13.1]	9.4 [5.4, 15.7]	8.9 [8.4, 9.3]

samples of each method to the minimum number of non-hallucinated samples across all methods. Table 7 presents the obtained results on 11kHands up to 3 significant digits, with UP and HR included for reference. The proxy metrics for diversity are essentially matched across all approaches on correct samples, differing only at the 3rd significant digit. This equivalent performance is obtained with IQ simultaneously achieving much greater hallucination mitigation in comparison to other methods. These results strongly support the claim that hallucinations are indeed responsible for artificially inflating proxy diversity metrics, while IQ preserves the diversity of valid outputs.

Table 7: Comparison of proxy diversity metrics computed on the subset of correct, non-hallucinated 11kHands samples, with UP and HR reported for reference. To eliminate the effect of sample size, each method is limited to the minimum number of non-hallucinated samples across all methods. Diversity values are reported up to 3 significant digits.

Method	IV ↑	DSV ↑	UP ↑	HR ↓
True data	0.0353	0.156	-	-
Baseline	0.0309	0.111	39.8	29.3
DG	0.0309	0.115	39.5	29.7
AAM	0.0308	0.112	40.6	29.3
RODS^{CAS}	0.0305	0.118	40.2	25.8
RODS^{SAS}	0.0305	0.115	41.0	29.7
IQ	0.0304	0.114	68.0	9.0

B.7 Hallucination reduction in flow-based models

As flow-based models gain prevalence, a natural question is whether the proposed LID-based analysis and the IQ correction extend beyond diffusion models. From a theoretical standpoint, diffusion- and flow-based models can be considered equivalent thanks to the PF-ODE formulation and a bijective mapping between the score and the velocity originally used in Flow Matching [FM; Lipman et al., 2023]. Therefore, to apply IQ with the exact formulation from our work, one has to simply translate the velocity obtained from an FM model to the score, apply the steps outlined in section 4, and translate back to the velocity domain. To demonstrate the feasibility of this combination in practice, we train an Optimal Transport Conditional Flow Matching [OT-CFM; Tong et al., 2024] model with a standard 2D UNet architecture [Ronneberger et al., 2015] on 11kHands for 24 A100 GPU hours,

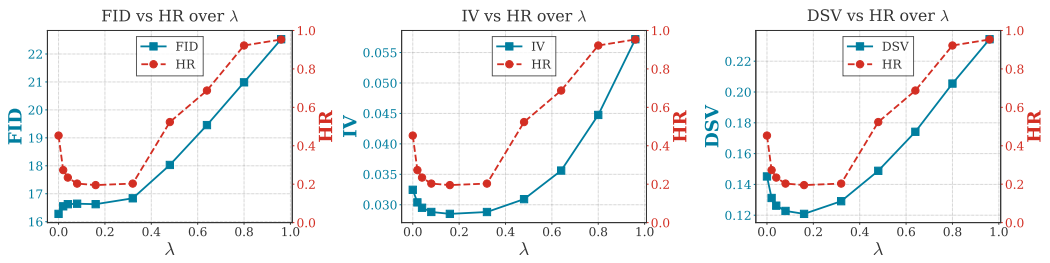

Figure 8: Ablation of the adaptive step size magnitude λ on the `11kHands` dataset.

extend its sampling scheme to allow for the application of both IQ and RODS in both variants, and conduct the evaluation using the exact setup from section 6 (*e.g.*, 2048 samples for proxy metrics, the blinded human study for UP and HR) with a 100-step Euler sampling scheme.

Table 8 presents the results. We observe that the considered OT-CFM model hallucinates slightly more often by default than the EDM model considered in our work, possibly due to the shorter training time. Interestingly, despite extensive hyperparameter tuning, RODS fails to mitigate hallucinations almost entirely, repeating the pattern observed for EDM, while only partially improving the proxy metrics, highlighting their inability to properly align with human perception and underscoring the difficulty of mitigating hallucinations in this setting. On the other hand, IQ significantly reduces hallucinations and achieves the highest UP, confirming that humans prefer its samples. Moreover, IQ exhibits the same minor shifts in diversity metrics as for EDM, which we explain in section B.6 by hallucinations artificially amplifying these scores. Overall, the results obtained for OT-CFM mirror the pattern observed for EDM, highlighting the strong performance and robustness of IQ across different frameworks.

Table 8: Quantitative comparison of all sampling methods on `11kHands` using an OT-CFM model with a 100-step Euler sampling scheme.

Method	FID ↓	IV ↑	DSV ↑	UP ↑	HR ↓
Baseline	18.8	0.026	0.13	58.4	43.8
RODS^{SAS}	17.9	0.026	0.13	57.9	41.4
RODS^{CAS}	18.1	0.026	0.13	58.4	43.0
IQ	18.9	0.023	0.11	71.2	7.8

B.8 Ablation studies

In the following subsections, we provide an ablation study of each hyperparameter of IQ on the `11kHands` dataset, where hallucinations are easiest to judge subjectively. For each hyperparameter combination, we manually labeled 128 samples, resulting in a total of 4096 labeled cases, and reported feature-based metrics using 2048 samples. Because the original annotators did not take part in these studies, the HR differs from the main results in table 1. In each study, we freeze all of the hyperparameters and vary only the one of interest. We use the settings from table 16 as the optimal configuration.

We note that while feature-based metrics are not the optimal choice for evaluating image quality when hallucinations are present, they serve as a useful proxy; large collapses or expansions in these metrics indicate an inevitable change in quality. Crucially, in the following experiments, we also observe that diversity metrics (IV and DSV) consistently inflate (suggesting *improvements*) even when FID grows exponentially and HR achieves almost 100%. This observation further highlights the misalignment between feature-based metrics (for diversity in this case) and human perception.

B.8.1 Ablating λ

Figure 8 provides the results of varying the adaptive step size magnitude λ across the $[0, 1]$ range. Across all metrics, we observe that $[0.05, 0.2]$ is its optimal range, where HR plateaus around 0.3, while other metrics start to grow exponentially.

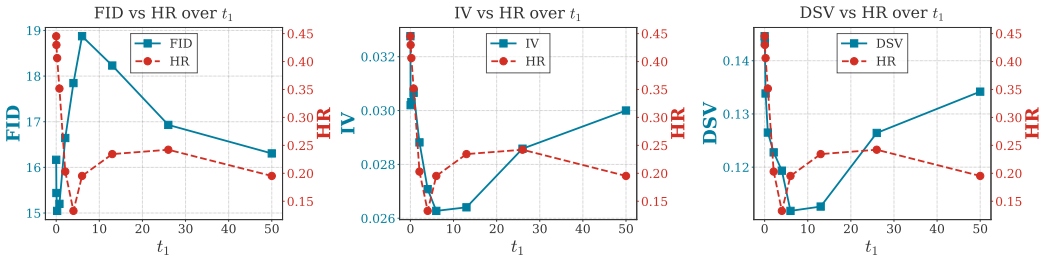

Figure 9: Ablation of the time interval bounds for IQ on the `11kHands` dataset. For each t_1, its corresponding t_2 was chosen so that IQ is applied at exactly 4 consecutive timesteps.

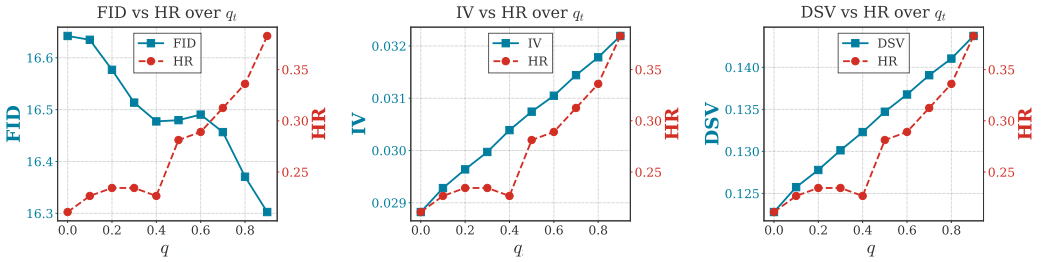

Figure 10: Ablation of the filtering parameter q on the `11kHands` dataset.

B.8.2 Ablating time interval

To ablate the choice of the time interval for IQ, we pick t_1 from an approximately log-uniform range of the EDM σ scale and choose the corresponding t_2 so that guidance is applied at exactly 4 consecutive timesteps. Figure 9 reveals a sharp transition from high HR values into a clear, optimal point, after which the hallucinations once again increase, confirming the theoretical results. Moreover, a clear trade-off between feature-based metrics and HR is observed, once again showing that FID, as well as other metrics, improve when hallucinations occur frequently, and start to worsen once they decrease.

B.8.3 Ablating filtering

To ablate the filtering parameter q, we sweep it across the $[0, 1]$ range, where $q = 0$ implies that IQ is applied unconditionally across all samples, while $q = 1$ means that it is never applied. Figure 10 visualizes the resulting dependency between q, feature-based metrics, and HR, leading to two conclusions. First, there is an evident correlation between q and the metrics (negative for FID, positive for IV and DSV). Second, there appears to be an optimal choice of q at around 0.4, where the HR achieves satisfactory values while not significantly altering feature-based metrics. This once again provides evidence that IQ might be trading the inevitable cost in feature-based metrics to decrease hallucinations.

The effect of filtering is also clear when analyzing the `GaussianGrid` results (fig. 17). With $q = 0$ (depicted as IQ*), all of the generated samples squeeze at the Gaussian modes, almost entirely zeroing out the variance of each Gaussian component. By applying filtering (IQ), this effect is eliminated and the original variance is almost matched.

B.8.4 Ablating constant scaling

To assess the validity of adaptive step size λ_t, we replace it with constant scaling and sweep its values between 0.0 and 10.0 (based on the range of values we observed from the adaptive approach). Figure 11 visualizes the results with an additional dimension for the presence of artifacts in the generated samples, *i.e.*, oversaturated colors and severe structural distortions in more extreme cases. The results confirm the instability of constant scaling, showing that within the optimal range of the trade-off between HR and other metrics, artifacts start to appear. This behavior is not observed when using the adaptive size (fig. 8), confirming its validity.

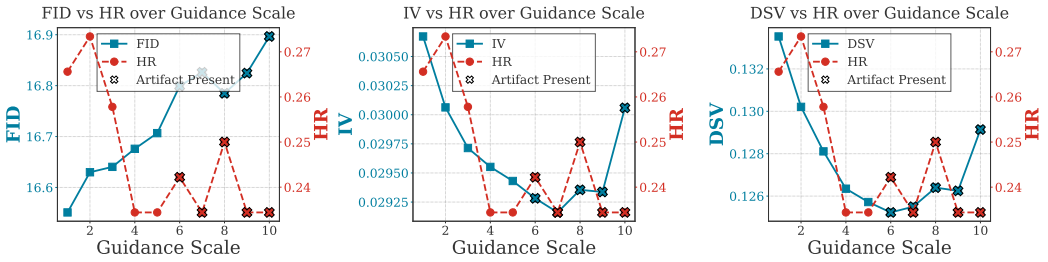

Figure 11: Ablation of constant scaling on the `11kHands` dataset.

B.8.5 Ablating the number of noise samples

The last hyperparameter of IQ is k, the number of i.i.d noise samples used to estimate the DSM expectation and, in turn, $\text{LID}_{\boldsymbol{\theta}}$ (see section B.10). In contrast to the previous studies, here we reuse the exact evaluation from section 6 (2048 samples for proxy metrics, the blinded human study for UP and HR), hence the results for $k = 32$ match table 1. Table 9 presents the results of sweeping $k \in \{2, 8, 32\}$ on `11kHands`, together with the mean runtime. Decreasing k 4-fold has almost no effect on the performance of IQ, while lowering the runtime by almost $3\times$. Furthermore, $k = 2$ still preserves most of IQ's effectiveness, mitigating over 50% of hallucinations with a runtime almost $5\times$ lower than for the default setting. These results suggest that k does not have to account for highly precise gradient estimation, but rather functions as a dial for the efficiency-budget trade-off. We leverage this observation when benchmarking large-scale models in section B.9.

Table 9: Ablation of the number k of i.i.d noise samples used to estimate $\text{LID}_{\boldsymbol{\theta}}$ on the `11kHands` dataset. Runtime is reported as the mean runtime (in seconds), alongside the 95% confidence interval.

Method	FID ↓	IV ↑	DSV ↑	UP ↑	HR ↓	Runtime (s)
Baseline	16.3	0.032	0.15	39.8	29.3	2.84 ± 0.003
IQ ($k = 2$)	16.4	0.031	0.14	61.3	13.5	3.90 ± 0.004
IQ ($k = 8$)	16.5	0.031	0.13	66.5	9.8	6.93 ± 0.010
IQ ($k = 32$)	16.6	0.030	0.13	68.0	9.0	19.01 ± 0.010

B.9 Runtime comparison

We compare the average sampling runtime of IQ with default unconditional sampling and both RODS variants as the only unconstrained methods for hallucination reduction. The comparison is performed on the most practical and high-resolution `RSNA` dataset, with results presented in table 10. While all methods add a significant overhead to baseline sampling, IQ remains around 20% faster than both RODS^{CAS} and RODS^{SAS}. These results confirm the superiority of IQ but also point toward an important research direction of reducing the computational cost while preserving the method's performance.

Table 10: Runtime comparison of unconstrained methods on the `RSNA` dataset. Values are reported as the mean sampling loop time (in seconds) over 16 batches of size 8, alongside the 95% confidence interval.

Method	Runtime (s)
Baseline	11.49 ± 0.63
IQ	49.83 ± 0.40
RODS^{CAS}	63.86 ± 2.27
RODS^{SAS}	60.33 ± 2.37

To further assess the practical applicability of IQ, we implement the inference pipelines of Stable Diffusion XL (SDXL) [Podell et al., 2024] and PixelDiT [Yu et al., 2026] (ImageNet, 512×512

resolution) and adapt both IQ and RODS to their sampling procedures, using the same hyperparameters as for `11kHands` (for IQ, this also means that the time window contains the same number of timesteps) and varying k, following the analysis in section B.8. Table 11 presents the mean runtime with 95% CIs for all methods on both models. Crucially, in terms of runtime, IQ scales much better to both text-to-image SDXL and high-resolution PixelDiT than RODS. Even in the default setting ($k = 32$), IQ remains faster than both RODS variants by a visible margin (*e.g.*, a $6.7\times$ overhead on SDXL compared to $7.8\times$ and $9.8\times$ for RODS^{SAS} and RODS^{CAS}), and decreasing k to 2 reduces the overhead to $1.5\times$ on SDXL and $1.3\times$ on PixelDiT.

Table 11: Runtime comparison of unconstrained methods on large-scale models. All methods use the hyperparameters from `11kHands` (for IQ, the time window contains the same number of timesteps). Values are reported as the mean runtime (in seconds) alongside the 95% confidence interval.

	Runtime (s)	
Method	SDXL	PixelDiT
Baseline	31.931 ± 0.014	7.88 ± 0.003
IQ ($k = 2$)	48.633 ± 0.045	10.25 ± 0.010
IQ ($k = 8$)	81.274 ± 0.017	14.45 ± 0.003
IQ ($k = 32$)	215.031 ± 0.053	31.42 ± 0.005
RODS^{SAS}	249.459 ± 1.501	41.78 ± 0.015
RODS^{CAS}	311.964 ± 2.464	41.82 ± 0.015

B.10 Hyperparameters

We include the hyperparameters used for DG in table 12, AAM in table 13, RODS^{CAS} in table 14, RODS^{SAS} in table 15, and IQ in table 16. These were obtained by running a grid search over approximately 500 configurations across all datasets for each baseline method.

For IQ, we emphasize that k, the number of i.i.d noise samples used to estimate $\text{LID}_{\boldsymbol{\theta}}$, is kept constant at $k = 32$ across all datasets in the main evaluation. We observed that this value universally ensures the stability of the estimator; section B.8 additionally shows that lowering k preserves most of the mitigation strength while substantially reducing the runtime. To compute q_t, we utilize 2048 reference samples for each dataset.

For `RSNA`, IQ is applied between $t = 0.7$ and $t = 0.75$, which corresponds to low noise levels within the SDB framework that utilizes an inverted U-shape noise schedule. For `SimpleShapes`, we followed Aithal et al. [2024], which noted that the typical behavior observed in DMs shifts to larger timesteps and hence apply IQ at relatively higher times. However, for timesteps smaller than the indicated interval, the image is already fully formed and hence guidance has no practical effect.

Table 12: Hyperparameters for Dynamic Guidance (DG). The names of hyperparameters follow their respective names in our implementation. For EDM datasets, `dg_start` and `dg_end` are normalized to the $[0, 1]$ interval by dividing the σ boundaries by $\sigma_{max} = 80$. For DDPM datasets, they are normalized by dividing the timesteps by $T_{max} = 1000$. For SDB, timesteps natively reside in the $[0, 1]$ interval.

Parameter	`11kHands`	`FFHQ`	`AFHQV2`	`MNIST`	`SimpleShapes`	`GaussianGrid`	`RSNA`
`dg_scale`	15.0	10.0	1.0	10.0	10.0	10.0	10^5
`dg_start`	0.125	1.0	0.125	0.5	0.25	0.25	0.7
`dg_end`	0.00125	0.0125	0.0125	0.25	0.0	0.0	0.4

B.11 SDXL evaluation details

We provide the complete details of the small-scale SDXL evaluation from section 6, including the structured prompt, the statistical analysis, and the hyperparameters. The prompt enforces predictable hand positioning, enabling straightforward visual inspection of structural hallucinations (*e.g.*, missing or additional fingers, deformed joints). While these errors mirror cases from `11kHands`, their

Table 13: Hyperparameters for Adaptive Attention Modulation (AAM). The names of hyperparameters follow their respective names in our implementation.

Parameter	11kHands	FFHQ	AFHQV2	MNIST	SimpleShapes	RSNA
γ	2.5	1.5	2.5	2.5	2.5	2.0
β	1.25	1.3	1.3	0.1	0.3	0.01
N	15	15	15	15	5	15
η	0.1	0.05	0.05	0.1	0.05	0.01

Table 14: Hyperparameters for RODS^{CAS}. The names of hyperparameters follow their respective names in our implementation.

Parameter	11kHands	FFHQ	AFHQV2	MNIST	SimpleShapes	GaussianGrid	RSNA
perturb_scale	30.0	8.0	1.0	5.0	1.0	0.1	0.1
cor_threshold	0.014	0.09	0.1	0.005	0.005	0.0001	0.002
num_adv_steps	1	1	1	1	1	1	1

scope is naturally broader given SDXL's generative capacity, making mitigation significantly more challenging:

```
A hyper-realistic macro photograph of a single human left hand, palm
facing the camera, held flat against a plain white background, with
all five fingers spread wide apart and clearly separated:  one thumb,
one index finger, one middle finger, one ring finger, and one little
finger, five fingers in total.  A plain silver band ring is worn
on the ring finger.  Soft even studio lighting, sharp focus, high
contrast, 8k resolution, highly detailed skin texture.
```

Following an initial hyperparameter grid search, the evaluation is conducted on 64 generated samples labeled case-by-case by 3 independent annotators. For IQ, the hyperparameters comprise the quenching magnitude $\lambda = 1$, the time window $[0.76, 0.81]$, expressed in SDXL's normalized scale, and no filtering ($q = 0$), with $k = 16$ set to maintain a reasonable time budget. The same grid search procedure, including the tuning of both variants, is applied to RODS. To evaluate statistical significance, we fit a Generalized Estimating Equation (GEE) logistic regression model (binomial family, logit link, exchangeable working correlation grouped by the model's generation seeds). Based on Wald's test on model coefficients, we obtain a Holm-Bonferroni-adjusted p-value of 0.00059, controlling for multiple testing across the evaluated hyperparameter grid, confirming the high statistical robustness of the reported mitigation. Annotators reported no visible degradation in image quality, suggesting that in models like SDXL, IQ primarily acts by shifting generation toward regions with deflated LID to stabilize structural output. Naturally, this initial result does not claim that IQ is the ultimate solution for hallucination mitigation in such models, but offers a promising base for potential future improvements.

B.12 Pseudocode

We provide detailed pseudocode of IQ in algorithm 1.

B.13 Resources

All of the experiments were performed on NVIDIA A100 40GB and H100 80GB GPUs, with runtime (table 10) reported on the former. The entire project required around 1000 GPU hours in total (this includes any preliminary or grid search experiment). The GPUs were part of NVIDIA DGX nodes.

B.14 Qualitative examples

We include more qualitative comparisons on 11kHands (fig. 12), AFHQV2 (fig. 14), FFHQ (fig. 13), MNIST (fig. 15), SimpleShapes (fig. 16), high-resolution version of GaussianGrid results (fig. 17),

Table 15: Hyperparameters for RODS^{SAS}. The names of hyperparameters follow their respective names in our implementation.

Parameter	11kHands	FFHQ	AFHQV2	MNIST	SimpleShapes	GaussianGrid	RSNA
perturb_scale	30.0	8.0	1.0	5.0	20.0	0.1	0.1
cor_threshold	0.014	0.09	0.1	0.01	0.05	0.0001	0.002
num_adv_steps	1	1	1	1	1	1	1

Table 16: Hyperparameters for IQ. The names of hyperparameters follow their respective names in our implementation. For EDM datasets, t_1 and t_2 are normalized to the $[0, 1]$ interval by dividing the guidance σ boundaries by $\sigma_{max} = 80$. For DDPM datasets, they are normalized by dividing the timesteps by $T_{max} = 1000$. For SDB, timesteps natively reside in the $[0, 1]$ interval.

Parameter	11kHands	FFHQ	AFHQV2	MNIST	SimpleShapes	GaussianGrid	RSNA
t_1	0.025	0.025	0.0088	0.24	0.48	0.0	0.70
t_2	0.0625	0.0375	0.025	0.40	0.96	0.72	0.75
λ	0.08	0.035	0.022	0.5	0.08	0.9	0.8
q	0.4	0.05	0.2	0.0	0.05	0.65	0.0
k	32	32	32	32	32	32	32

ground truth scans and their reconstructions from RSNA (fig. 18) and the corresponding qualitative comparison (fig. 19).

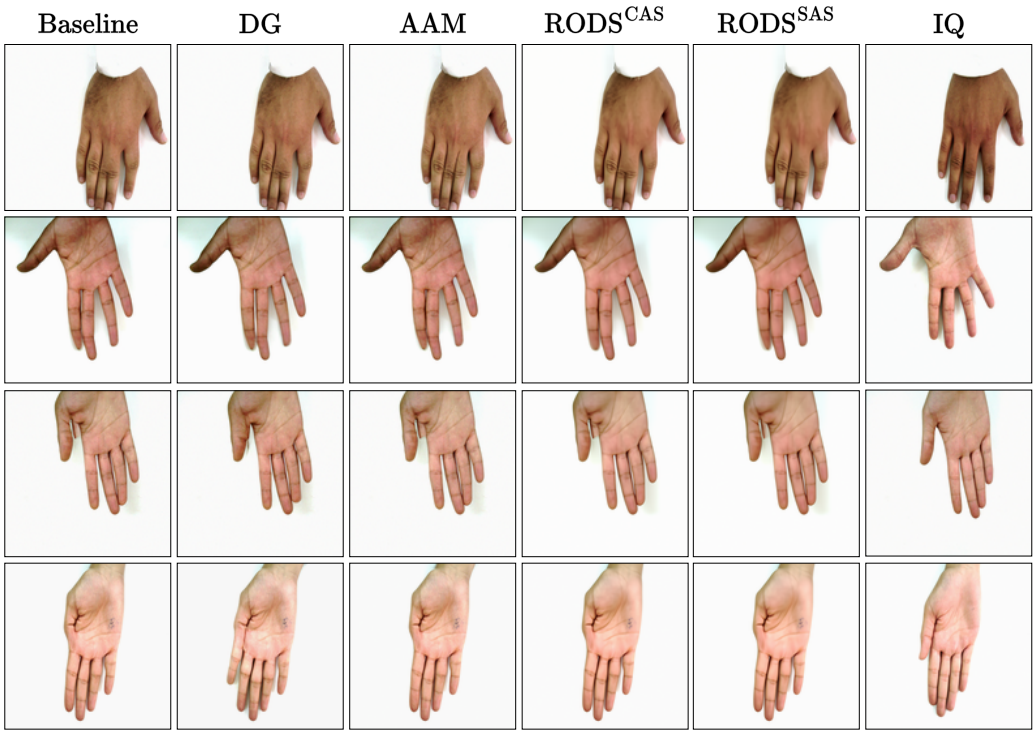

Figure 12: Qualitative comparison of all sampling methods on `11kHands`.



Figure 13: Qualitative comparison of all sampling methods on `FFHQ`.



Figure 14: Qualitative comparison of all sampling methods on `AFHQV2`.

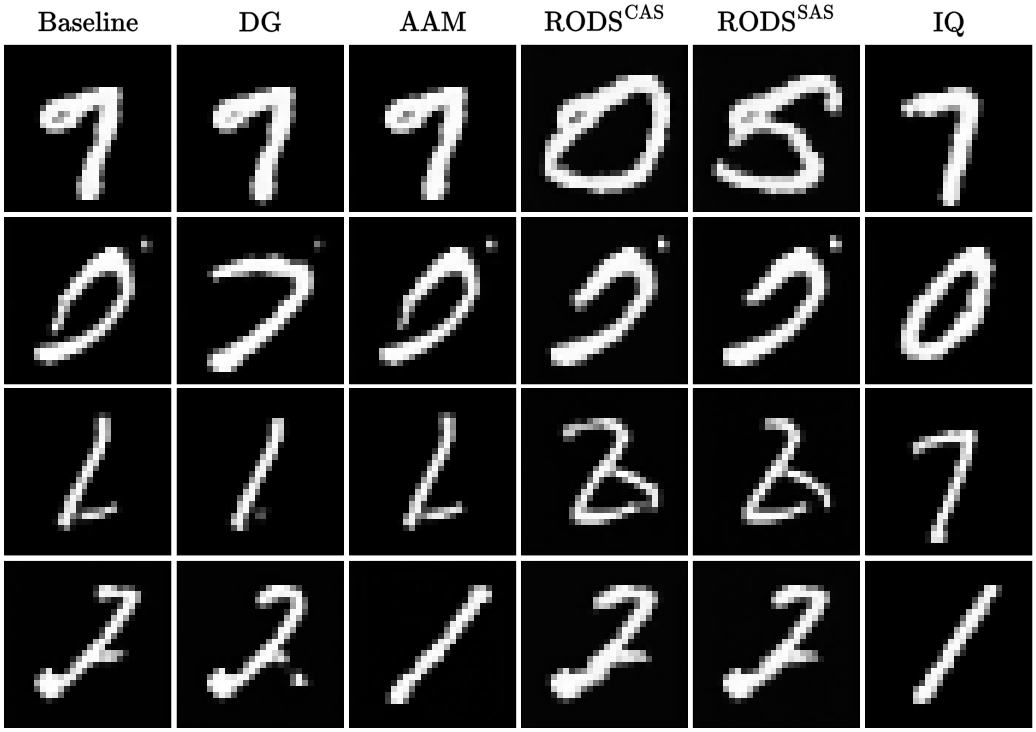

Figure 15: Qualitative comparison of all sampling methods on `MNIST`.

Algorithm 1 Intrinsic Quenching (IQ): sampling step.
Note: autograd returns the gradient of the energy function w.r.t the input.

```
# net: EDM denoiser predicting clean data x_0
# x: noisy input batch at time t
# lambda: target magnitude ratio for guidance
# q: target quantile for filtration (e.g., 0.4)

# Apply Intrinsic Quenching only within the window [t_1, t_2]
if t < t_1 or t > t_2:
    return stopgrad(net(x, t))

x.requires_grad_(True)

# Predict posterior mean x_0_hat and evaluate DSM energy (LID)
x_0_hat = net(x, t)
LID = dsm_loss(x_0_hat, x, t)

# Compute energy gradient w.r.t the input state x
grad = autograd(LID.sum(), x)

# Project guidance into data space (Sigma_t = t^2 * I for EDM)
raw_guidance = (t ** 2) * grad

# Calculate adaptive scale to maintain constant magnitude ratio
nat_update = x_0_hat - x
scale = (lambda * norm(nat_update)) / (norm(raw_guidance) + 1e-8)

# Filtration strategy
# q_t is the q-th quantile of the baseline LID distribution at time t
q_t = quantile(baseline_LID(t), q)
mask = where(LID >= q_t, 1.0, 0.0)

# Apply masked adaptive EDM update
guided_x_0_hat = x_0_hat - (mask * scale * raw_guidance)

return stopgrad(guided_x_0_hat)
```

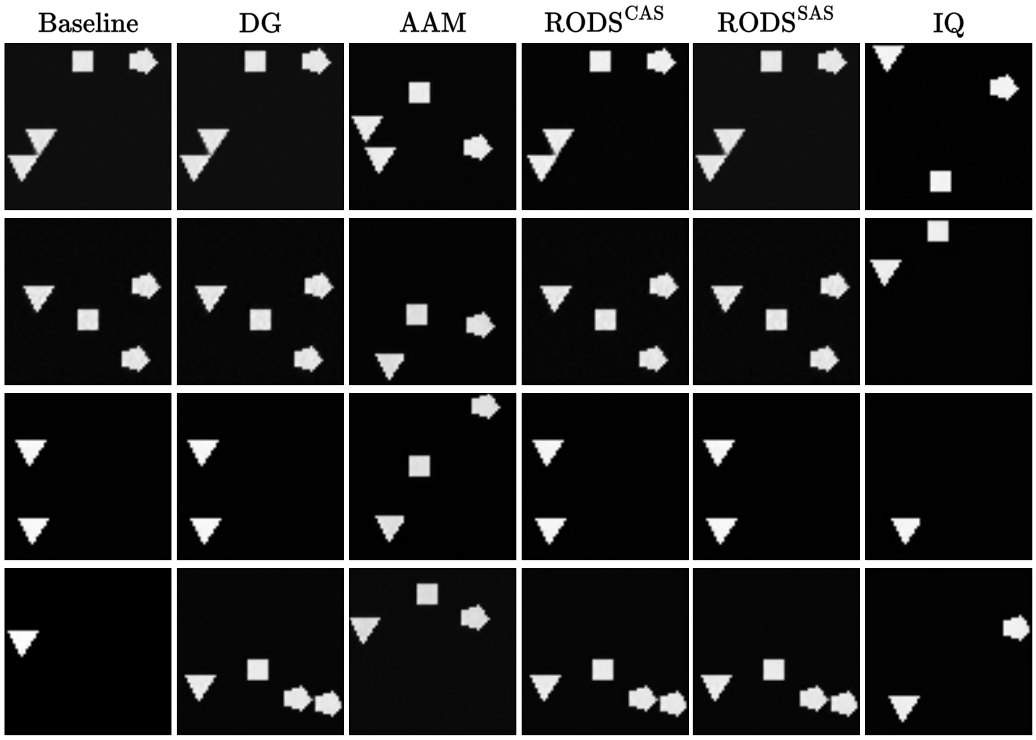

Figure 16: Qualitative comparison of all sampling methods on `SimpleShapes`.

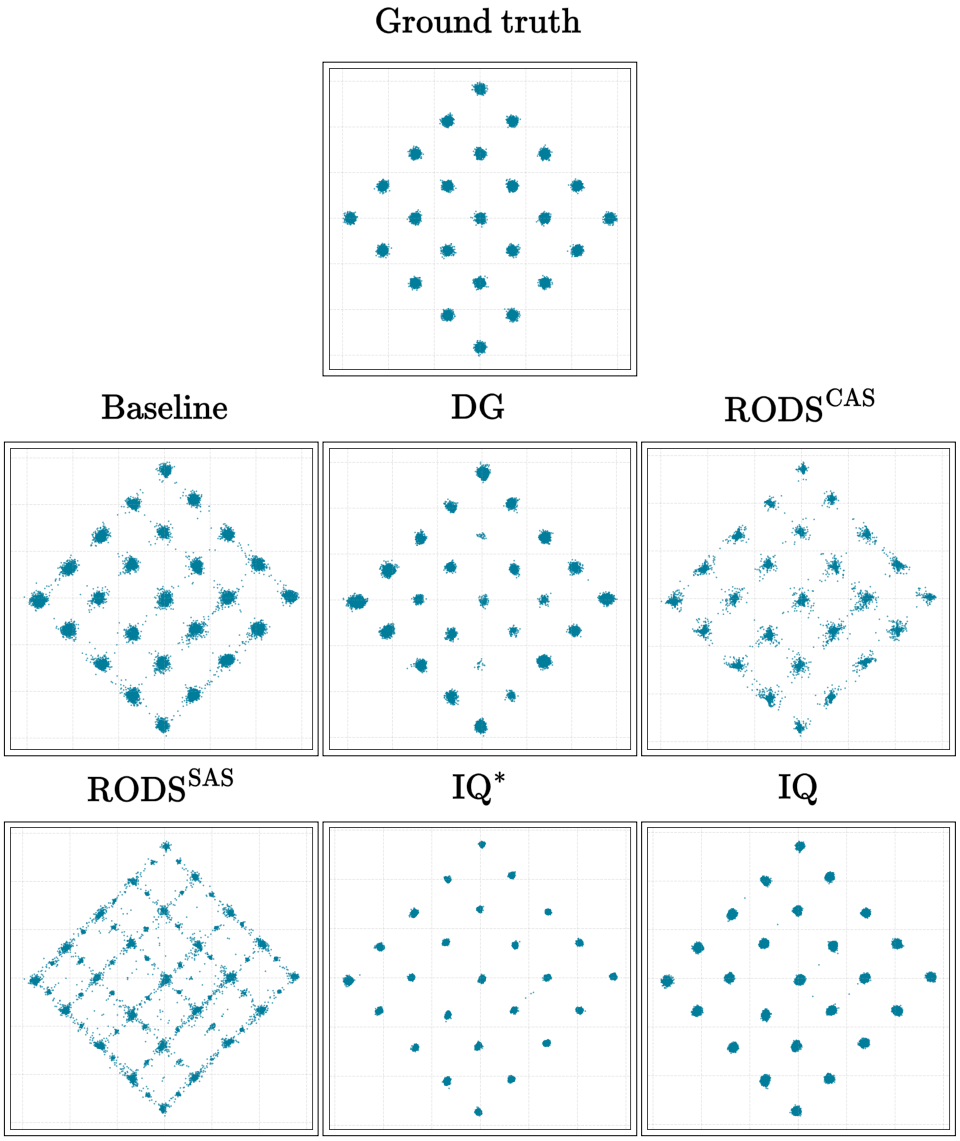

Figure 17: High-resolution visualization of samples generated by each method on the `GaussianGrid` dataset. Each plot depicts 16384 samples.

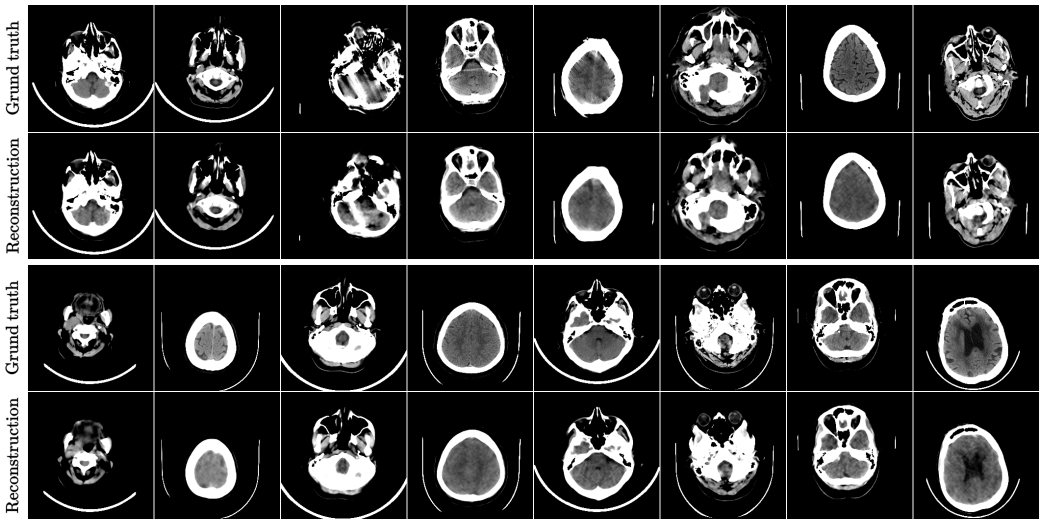

Figure 18: Ground truth brain CT scans from the RSNA dataset and their corresponding reconstructions obtained with SDB using 100-step sampling.

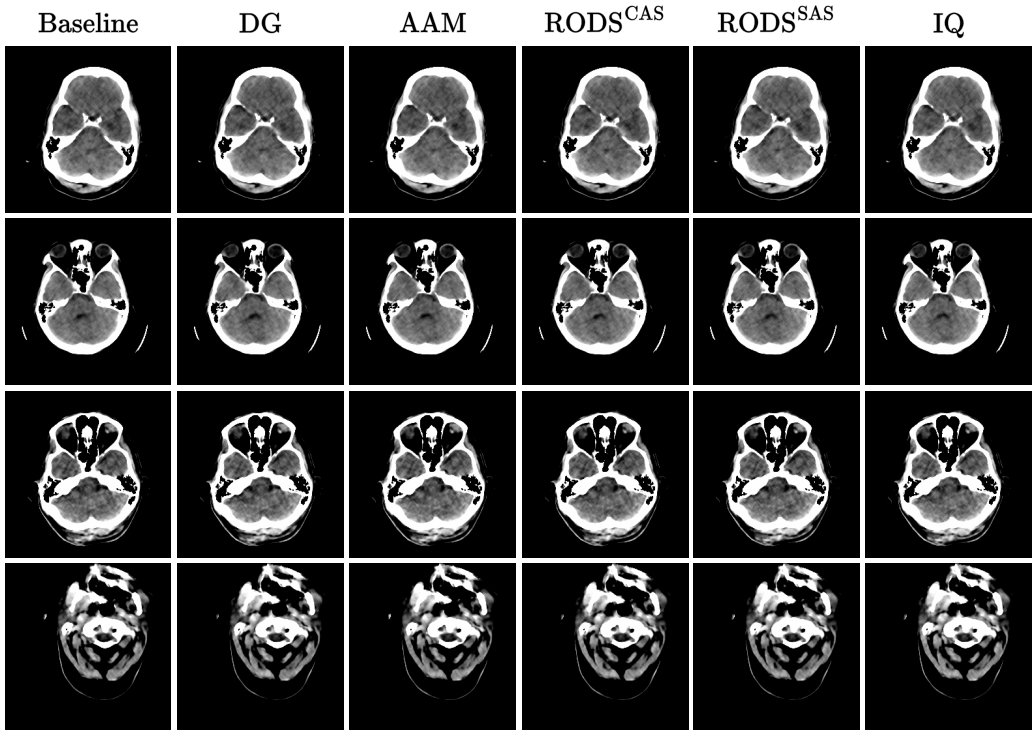

Figure 19: Qualitative comparison of all sampling methods on RSNA.

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: Our contributions include the initial empirical investigation (section 3), the theoretical results with proofs (informal statements in section 4, full formal statements with proofs in section A) and the empirical results in section 6 and section B.

   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: The final section discusses the limitations regarding the limited scope of proxy metrics for image quality and diversity, pointing to an unknown theoretical relationship, and the computational overhead of the proposed method.

   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: These are provided in section 4 and section A.

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: The paper provides a detailed explanation of the proposed method, high-level overview of each baseline, hyperparameters of all methods and a fully reproducible codebase.

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. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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 is included as supplementary material.

   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: The necessary details are either provided in the main text, the Appendix or in the attached codebase.

   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: Error bars are reported for the human-based evaluation. Proxy metrics (such as FID) are reported based on a single run, following standard practice and due to high computational demands.

   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: Resources are reported in section B.13.

   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 conducted research fully conforms with the 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: Removing structural hallucinations in diffusion models has positive societal impacts; it allows for a smoother, more reliable interaction with image generation models, increasing trustworthiness, and may positively affect scenarios where diffusion models influence a decision-making process, *e.g.*, in medical image reconstruction. Regarding negative impacts, failure of IQ or similar methods in reducing hallucinations leads to increased computational overhead, which inflates the carbon footprint.

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: The paper does not pose such risks.

    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: The paper properly cites each source of existing data that was used throughout the research process and is required for reproduction. For ported chunks of code, the original repository is mentioned. The supplementary material does not provide any of the cited resources directly; it redirects the reader to the proper source.

    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: The paper's contributions are not focused on providing new assets. The details required for reproducing the results are provided within the attached codebase.

    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: [Yes]

    Justification: The user study is described in detail (section B.2). The human subjects were compensated in a way that fully aligns with their country's law.

    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: [Yes]

Justification: The IRB approval for the conducted user study was not required, as it was based on risk-free protocol and anonymous, publicly available datasets. The human subjects were fully informed about the structure of the user study and gave informed consent regarding the participation.

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 were used solely for editing and formatting purposes.

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.