UNDERSTANDING LATENT DIFFUSABILITY VIA FISHER GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models often degrade in latent spaces, yet the formal causes remain poorly understood. We quantify latent-space diffusability via the rate of change of the Minimum Mean Squared Error (MMSE) along the diffusion trajectory. Our framework decomposes this MMSE rate into contributions from Fisher Information (FI) and Fisher Information Rate (FIR). We show that isometric embeddings preserve intrinsic FI and establish quantitative intrinsic-FI bounds for a broader class of bi-Lipschitz encoders with controlled weak volume distortion, whereas FIR is governed by the interplay between encoder and data geometries. Our analysis separates four geometric contributions in local stability bounds for Gaussian-smoothed FIR: dimensional compression, tangential distortion, high-frequency encoder curvature, and curvature of data manifold. Experiments across diverse autoencoding architectures provide qualitative support for the geometric mechanisms identified by the theory and show that empirical FI and FIR track several measures of generation quality and latent-space geometry in the settings tested. We establish FI and FIR as a comprehensive analytical framework for understanding latent diffusability.

1 INTRODUCTION

Diffusion models have revolutionized generative modeling by transforming noise into structured data through a learned reverse process (Ho et al., 2020). While these models excel in high-dimensional pixel spaces, computational constraints often necessitate their application within the compressed latent spaces of pre-trained autoencoders (Rombach et al., 2022). However, a poorly understood phenomenon remains: diffusion models often degrade or fail when applied directly to the latent spaces of standard autoencoders, including Variational Autoencoders (VAEs) (Kingma & Welling, 2014; Shi et al., 2025). Skorokhodov et al. (Skorokhodov et al., 2025) refer to this mismatch between diffusion dynamics in pixel space and latent space as the *diffusability* problem.

Current explanations for this failure typically point to the "geometric distortion" introduced by standard encoders, which warp the data manifold and create regions that diverge from the Gaussian prior assumed by the diffusion process (Arvanitidis et al., 2018; Lobashev et al., 2025). In response, methods have emerged to mitigate this issue, such as geometry-preserving adaptations (Lee et al., 2025c), explicit target latent geometries (Cho et al., 2023; Sáez de Ocáriz Borde et al., 2023; Yue et al., 2026), and modified training objectives (Kouzelis et al., 2025; Yang et al., 2026; Heek et al., 2026; Baade et al., 2026) (see Appendix A for a comprehensive discussion of these methods).

Despite these empirical advances, the underlying causes of latent diffusion failure remain theoretically obscured. Crucially, the literature lacks a formal, quantitative framework for measuring how "diffusible" a latent space is *before* committing to the costly training of a full diffusion model.

We demystify latent-space diffusability through the rigorous lens of Fisher geometry (Amari, 2016), complementing recent information-geometric perspectives on diffusion trajectories (Karczewski et al., 2026). We analyze diffusability via *denoising complexity*, quantified by the rate of change of the Minimum Mean Squared Error (MMSE) along the diffusion trajectory. We show that this denoising complexity fundamentally decomposes into two terms: the Fisher Information (FI) and its rate of dissipation, which we define as the Fisher Information Rate (FIR). We establish change-of-variables bounds for intrinsic FI and localized stability bounds for ambient FIR, linking both quantities to encoder and manifold geometry. This defines model-independent population diagnostics

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and identifies local geometric conditions associated with FIR preservation. Scalable and statistically reliable pre-training estimation in high dimensions remains a separate practical challenge.

Our main contributions are as follows:

1. **Theoretical Framework:** We formalize the relationship between latent diffusability, denoising complexity (MMSE), and Fisher Information Geometry.
2. **Geometric Bounds:** We derive quantitative intrinsic-FI bounds for bi-Lipschitz encoders with controlled volume distortion (§3.1), together with sufficient local curvature and near-isometry conditions for controlling ambient Gaussian-smoothed FIR (§3.2). These results separate four geometric mechanisms: dimensional compression, tangential distortion, high-frequency encoder curvature, and intrinsic manifold curvature.
3. **Empirical Evaluation:** Experiments on toy, FFHQ, and ImageNet-based representations provide qualitative support for the geometric mechanisms identified by the theory and show that empirical FI and FIR track several measures of latent diffusion performance in the settings tested.

2 Background

Score-based Diffusion Models. Continuous-time score-based models typically transform a data distribution into a noise prior through a forward diffusion process. Let μ be a Borel probability measure on $\mathbb{R}^D$ concentrated on a low-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$. We consider a diffusion process parameterized by the noise variance $\tau := \sigma^2(t)$, where a clean sample $\mathbf{x}_0 \sim \mu$ is corrupted by standard Gaussian noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, yielding the observation $\mathbf{x} = \mathbf{x}_0 + \sqrt{\tau}\mathbf{n}$ (Song et al., 2021). The marginal density of these noisy observations, known as the heat-smoothed density $p_\tau(\mathbf{x})$, and the corresponding probability flow ODE evolve as:

$$\frac{\partial p_\tau}{\partial \tau} = \frac{1}{2}\Delta_{\mathbf{x}} p_\tau \quad \text{and} \quad d\mathbf{x} = -\frac{1}{2}\nabla_{\mathbf{x}} \log p_\tau(\mathbf{x}) d\tau. \tag{1}$$

Generating samples requires a neural approximation of the *score function*

$$\mathbf{s}_\tau(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\tau(\mathbf{x}). \tag{2}$$

Manifold geometry complicates learning $\mathbf{s}_\tau$ via high-frequency variations in normal directions.

Latent Diffusion and the Geometry Gap. To reduce computational costs, diffusion often operates within the compressed latent space $\mathcal{Z}$ of a pre-trained autoencoder (Rombach et al., 2022). However, standard VAEs regularize $\mathcal{Z}$ toward a Gaussian prior using Kullback-Leibler divergence, often inducing "manifold crinkling" (Arvanitidis et al., 2018). This phenomenon forces distant points on the data manifold $\mathcal{X}$ into adjacent latent coordinates to satisfy the prior, causing Euclidean distances in $\mathcal{Z}$ to misrepresent the intrinsic data geometry. This disparity creates a fundamental "geometry gap." Recent architectures mitigate these distortions by regularizing latent structure: Geometry Preserving Encoder/Decoder (GPE) aligns latent and data-space Euclidean geometries (Lee et al., 2025c), while spherical and manifold autoencoders employ non-Euclidean latent spaces to capture complex topologies (Sáez de Ocáriz Borde et al., 2023; Yue et al., 2026).

The Diffusability Issue. We use the population denoising resistance in (6) as an information-theoretic notion of diffusability. In this framework, FIR measures the average local variation of the true population score and its contribution to the rate of change of MMSE. By itself, FIR does not constitute a theorem about approximation error, optimization difficulty, or sample complexity for a finite neural score model.

Spectral Artifacts in Latent Spaces. Compressed latent spaces often exacerbate these irregularities. Spectral analyses show autoencoders introduce high-frequency components or singularities, impeding coarse-to-fine synthesis (Skorokhodov et al., 2025). Even with smoother encodings (Lee et al., 2025a;b), diffusion remains unstable if the encoded manifold in $\mathcal{Z}$ lacks structural regularity.

Limitations of Metric Regularization. Recent techniques constrain the encoder Jacobian to enforce local isometry (Chen et al., 2020; Palma et al., 2025). While empirically addressing the geometry gap, they lack a formal information-geometric analysis of diffusability. To address this gap, we leverage Fisher geometry to establish a rigorous link between encoder geometry and denoising complexity, explaining why certain latent spaces remain difficult to diffuse despite geometric-preservation properties.

Information-Geometric Formalism. We unify the analysis of both data and latent spaces by considering a generic Borel probability measure μ on $\mathbb{R}^k$. This encompasses both the pixel-space setting, where $k = D$ and μ is supported on $\mathcal{M}$, and the latent-space setting, where $k = d$ and μ is the pushforward measure under the encoder E.

For $k \geq 1$ and $\tau > 0$, write

$$\gamma_\tau^k := \mathcal{N}(\mathbf{0}, \tau I_k).$$

When the ambient dimension is clear, we write simply γ_τ. The heat-smoothed measure

$$\mu_\tau := \mu * \gamma_\tau^k$$

admits a probability density $p_\tau(\mathbf{x})$ with score $\mathbf{s}_\tau(\mathbf{x})$ defined in (2). Whenever μ, τ, and k are clear from context, we may omit the superscript (k) and the explicit argument μ_τ from our notation.

The *Fisher Information* (FI) of μ_τ is its square-averaged score and the *Fisher Information Rate* (FIR) is the dissipation rate of FI. That is,

$$\mathcal{I}^{(k)} \equiv \mathcal{I}^{(k)}(\mu_\tau) := \mathbb{E}_{p_\tau}\left[\|\mathbf{s}_\tau(\mathbf{x})\|^2\right] \quad \text{and} \quad \mathcal{R} \equiv \mathcal{R}^{(k)}(\mu_\tau) := -\frac{d}{d\tau}\mathcal{I}^{(k)}(\mu_\tau). \tag{3}$$

Standing Regularity Assumptions. For identities involving MMSE, we assume that the clean measure has finite second moment:

$$\int_{\mathbb{R}^k} \|x\|^2 \, d\mu(x) < \infty.$$

For the heat-semigroup differentiation and integration-by-parts identities below, we assume that the displayed quantities are finite and that differentiation under the integral sign and cutoff-based integrations by parts are justified. These conditions hold for the compactly supported and Gaussian measures used in our theoretical and numerical examples.

The next proposition, proved in Appendix B.1, establishes a representation of FIR, which emerges in information theory (Costa, 1985), diffusion semigroups (Bakry & Émery, 1985), and optimal transport (Villani, 2009). Let $\mathbf{H}_\tau = \nabla_{\mathbf{x}}^2 \log p_\tau(\mathbf{x})$.

Proposition 2.1 (Hessian representation of FIR)**.** *For every $\tau > 0$,*

$$\mathcal{R}^{(k)}(\mu_\tau) = \mathbb{E}_{p_\tau}\left[\|\nabla_{\mathbf{x}}\mathbf{s}_\tau(\mathbf{x})\|_F^2\right] = \mathbb{E}_{p_\tau}\left[\mathrm{Tr}(\mathbf{H}_\tau^2)\right]. \tag{4}$$

Computational Efficiency. While the FIR involves the Frobenius norm of the score Jacobian, it remains computationally tractable for high-dimensional neural networks. Using the Hutchinson trace estimator (Hutchinson, 1989), we approximate the norm as $\mathcal{R}(\tau) = \mathbb{E}_{\mathbf{x},\mathbf{v}}[\|\nabla_{\mathbf{x}}\mathbf{s}_\tau(\mathbf{x})\mathbf{v}\|^2]$, where $\mathbf{v}$ is a random noise vector. The term $\nabla_{\mathbf{x}}\mathbf{s}_\tau(\mathbf{x})\mathbf{v}$ is evaluated as a Jacobian-vector product (JVP) in a single forward-mode pass, bypassing the need to explicitly compute or store the full Jacobian matrix (see Appendix C).

3 Analysis of Diffusability via Fisher Information

To formalize "diffusability," we analyze the sensitivity of the global Minimum Mean Squared Error (MMSE) to τ. Following the setting of §2, where $k \in \{D, d\}$, the error in estimating the clean signal $\mathbf{x}_0$ from its noisy observation $\mathbf{x}$ at variance τ is defined as $\mathrm{MMSE}^{(k)}(\mu_\tau) = \mathbb{E}[\|\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0|\mathbf{x}]\|^2]$.

The I-MMSE identity (Guo et al., 2005), proved in Appendix B.2, implies that the optimal denoising performance is fundamentally constrained by this information content:

$$\mathrm{MMSE}^{(k)}(\mu_\tau) = \tau k - \tau^2 \mathcal{I}^{(k)}(\mu_\tau). \tag{5}$$

Differentiating (5) and applying the definition of FIR yields the decomposition:

$$\underbrace{\frac{d}{d\tau}\mathrm{MMSE}^{(k)}(\mu_\tau)}_{\text{Denoising Resistance}} = \underbrace{k - 2\tau\mathcal{I}^{(k)}(\mu_\tau)}_{\text{Intrinsic Noise Gain}} + \underbrace{\tau^2\mathcal{R}^{(k)}(\mu_\tau)}_{\text{Geometric Complexity Penalty}}. \tag{6}$$

This decomposition separates two contributions to population denoising resistance: the *Intrinsic Noise Gain*, which describes the baseline dimensional and information-dependent contribution, and the *Geometric Complexity Penalty*, which is governed by the variation of the population score.

A latent representation can alter either contribution. In §3.1, we study how deterministic geometric encoders affect intrinsic FI and the Intrinsic Noise Gain. In §3.2, we derive sufficient local geometric conditions for controlling the Geometric Complexity Penalty. We describe the essence of several theoretical results in the main text, while their complete formal statements, assumptions, and proofs are provided in the corresponding appendix sections referenced below.

3.1 Intrinsic Fisher Information under Geometric Encoders

We first study how intrinsic Fisher information changes when a measure on the data manifold is pushed forward deterministically by an encoder. The result isolates two effects: tangential metric distortion and variation of the encoder's tangential volume Jacobian. This intrinsic comparison is distinct from the ambient Gaussian-smoothed data and latent channels considered later.

Intrinsic FI under bi-Lipschitz encoders. Proposition B.1, stated and proved in Appendix B.3, considers smooth m-dimensional manifolds $\mathcal{M} \subset \mathbb{R}^D$ and $\mathcal{N} \subset \mathbb{R}^d$, together with a bi-Lipschitz homeomorphism $E : \mathcal{M} \to \mathcal{N}$. Its lower and upper intrinsic Lipschitz constants are denoted by c and C. Let $\mathcal{I}_{\mathcal{M}}$ be the intrinsic FI of a measure ν on $\mathcal{M}$, and let $\mathcal{I}_{\mathcal{N}}$ be that of its pushforward $E_{\#}\nu$. Define the tangential Jacobian $\mathcal{J}_E(x) = \sqrt{\det(DE(x)^\top DE(x))}$, its tangential gradient $g_E = \nabla_{\mathcal{M}} \log \mathcal{J}_E$, and the Jacobian-variation term $\widetilde{\varepsilon}^2 = \int_{\mathcal{M}} \|g_E(x)\|^2 \, d\nu(x)$.

Under the weak regularity conditions given in the appendix, the pushforward has finite intrinsic FI, and for every auxiliary parameter $\theta \in (0, 1)$,

$$\frac{1-\theta}{C^2}\mathcal{I}_{\mathcal{M}} - \frac{(\theta^{-1}-1)\widetilde{\varepsilon}^2}{C^2} \le \mathcal{I}_{\mathcal{N}} \le \frac{1+\theta}{c^2}\mathcal{I}_{\mathcal{M}} + \frac{(\theta^{-1}+1)\widetilde{\varepsilon}^2}{c^2}.$$

Thus, c and C control tangential metric distortion, while $\widetilde{\varepsilon}$ controls variation of the tangential volume Jacobian. In particular, an isometric encoder preserves intrinsic FI exactly.

Proposition B.1 concerns intrinsic FI under deterministic pushforward and does not directly compare the ambient Gaussian-smoothed data and latent channels, since nonlinear encoding generally does not commute with Gaussian smoothing.

For piecewise-smooth encoders, the proposition requires $\log \mathcal{J}_E$ to have no jump across activation interfaces. Thus, Leaky ReLU with slope parameter $\alpha > 0$ is bi-Lipschitz but, for $\alpha \neq 1$, is not covered directly by the proposition. Gaussian smoothing nevertheless gives finite FI for any latent probability measure $\mu_Z := E_{\#}\nu$, with $\mathcal{I}^{(d)}(\mu_{Z,\tau}) \le d/\tau$ for every $\tau > 0$. The full discussion of piecewise-smooth encoders and Gaussian regularization is provided in Appendix B.3. Exact smoothed-density and FI calculations for a Gaussian toy model with coordinatewise ReLU and Leaky-ReLU encoders are given in Appendix E.1.1.

Corollary B.4 further isolates ambient dimensional compression. On flat supports, preserving tangential FI changes the Intrinsic Noise Gain by exactly $D - d$. On curved supports, $D - d$ is instead the flat tangent-space baseline, with additional curvature-dependent corrections.

3.2 Fisher Information Rate and Second-Order Distortion

The dominant factor in diffusion difficulty at higher noise levels is the *Geometric Complexity Penalty* governed by the FIR, $\mathcal{R}^{(k)}(\mu_\tau)$. High-frequency manifold crinkling or structural irregularities inflate the score Jacobian $\nabla_{\mathbf{x}}\mathbf{s}_\tau$, causing denoising errors to compound quadratically as indicated by the factor τ^2 in (6). Therefore, effective diffusion requires keeping this penalty low.

The FIR is mathematically governed by the magnitude of the score Jacobian, equivalently the Hessian of the log-density. At the level of deterministic intrinsic pushforwards, differentiating the score change-of-variables identity produces terms involving both the encoder's second derivatives and variation of its volume Jacobian; we record this calculation in Appendix B.6. This intrinsic calculation motivates the need for second-order encoder control, but it is distinct from the comparison of ambient Gaussian-smoothed laws. The latter comparison is established directly by Theorem B.8 and

Corollary B.11. Nonsmooth activation interfaces fall outside the C^2 theorem and can generate large small-noise score curvature after Gaussian smoothing, as illustrated by the ReLU and Leaky-ReLU experiments. Because this injected curvature directly distorts the natural geometric complexity of the data, we need a metric to quantify it. To formalize this, we define the relative deviation between data and latent space FIRs.

Definition 3.1 (FIR Deviation)**.** Let the intrinsic dimension of μ on $\mathbb{R}^D$ be m, with $m \le d \le D$, and let $\mu_Z := E_{\#}\mu$ be its pushforward on $\mathbb{R}^d$. Let μ_τ and $\mu_{Z,\tau}$ denote the corresponding Gaussian-smoothed measures.

The FIR deviation is

$$\mathcal{D}_{\mathcal{R}}(\tau) := \left| 1 - \frac{\mathcal{R}^{(D)}(\mu_\tau)}{\mathcal{R}^{(d)}(\mu_{Z,\tau})} \right|.$$

When $m < d \le D$, we additionally define the dimension-normalized FIR deviation by

$$\mathcal{D}^{\mathrm{sc}}_{\mathcal{R}}(\tau) := \left| 1 - \left(\frac{d-m}{D-m} \right) \frac{\mathcal{R}^{(D)}(\mu_\tau)}{\mathcal{R}^{(d)}(\mu_{Z,\tau})} \right|.$$

The unscaled deviation compares the full ambient FIRs and remains meaningful in the equal-dimensional case $m = d = D$. On flat m-dimensional supports with $m < d \le D$, Lemma B.6 shows that the factor $(d-m)/(D-m)$ compensates for the universal normal-direction FIR contribution. On curved supports, it should instead be interpreted as a leading codimension normalization: curvature-dependent corrections remain and are quantified locally by Corollary B.11.

3.3 Stability Theorems on Manifolds of Positive Reach

To analyze FIR stability, we first establish a flat, linear baseline, and then extend the comparison locally to nonlinear encoders and curved manifolds.

We quantify the curvature of $\mathcal{M}$ by its reach $\rho_{\mathcal{M}} := \mathrm{reach}(\mathcal{M})$, defined as the supremum of all radii $r > 0$ such that every point within distance r of $\mathcal{M}$ has a unique nearest point on $\mathcal{M}$.

Assumption 3.2 (**Flat Support and Linear Mapping**)**.** Let $M \subset \mathbb{R}^D$ be an m-dimensional linear subspace with an isometry $\mathbf{U} : \mathbb{R}^m \to M$. Let ρ be a probability measure on $\mathbb{R}^m$ satisfying $\mathrm{supp}(\rho) \subseteq B_m(\mathbf{0}, R_\rho)$ for some $R_\rho < \infty$. Let $\mu := (\mathbf{U}_{\#}\rho) \otimes \delta_{\mathbf{0}}$ be the corresponding probability measure on $\mathbb{R}^D$, where $\delta_{\mathbf{0}}$ is the Dirac measure at $\mathbf{0}$ in the orthogonal complement of M. Let $m \le d \le D$, and fix an isometry $\mathbf{V} : \mathbb{R}^m \to \mathbb{R}^d$. Assume the encoder $E : \mathbb{R}^D \to \mathbb{R}^d$ is linear on M and satisfies $E(\mathbf{U}\mathbf{x}) = \mathbf{V}(\mathbf{A}\mathbf{x})$ for an invertible matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$.

Theorem 3.3 (**Linear Stability**)**.** *Under Assumption 3.2, let $\mu_Z := E_{\#}\mu$. Assume that $\mathbf{A}$ satisfies the near-isometry condition $\|\mathbf{A}^\top\mathbf{A} - \mathbf{I}_m\|_2 \le \delta$ for some $\delta \in (0,1)$. Then there exists a constant $C_m > 0$, depending only on m, such that, for every $\tau \ge R_\rho^2$, the following statements hold:*

- *If $m = d = D$, then $\mathcal{D}_{\mathcal{R}}(\tau) \le C_m\delta$.*
- *If $m < d \le D$, then $\mathcal{D}^{\mathrm{sc}}_{\mathcal{R}}(\tau) \le C_m \left(\frac{D-d}{D-m} + (d-m)\delta \right)$.*

The condition $\tau \ge R_\rho^2$ uniformly controls the clean tangential signal and its near-isometric perturbation, allowing Lemma B.9 to yield a constant depending only on m. Compact support is sufficient but not necessary for linear FIR stability. The unbounded Gaussian distributions used in the toy experiments are handled separately using the exact formulas in Appendix E.

Local nonlinear and curved stability of FIR. To extend the linear analysis to curved manifolds and nonlinear encoders, we localize the data distribution to $\mathcal{M} \cap B(x_0, r\sqrt{\tau})$, the neighborhood visible at diffusion scale $\sqrt{\tau}$. A canonical construction of this localized distribution is given in Definition B.7 in Appendix B.9. The result below applies more generally to any probability measure supported on this neighborhood.

Under this localization, Theorem B.8, stated and proved in Appendix B.9, considers a C^2 encoder $E : \mathbb{R}^D \to \mathbb{R}^d$ on an m-dimensional C^2 manifold $\mathcal{M} \subset \mathbb{R}^D$ with positive reach $\rho_{\mathcal{M}}$. Define $A_{x_0} := DE(x_0)|_{T_{x_0}\mathcal{M}}$ and the tangent-affine approximation $F_{x_0}(x) := E(x_0) + A_{x_0}P_{T_{x_0}\mathcal{M}}(x - x_0)$. The local tangential distortion is $\delta_{x_0} := \|A_{x_0}^\top A_{x_0} - I_m\|_{\mathrm{op}} \le \delta_* < 1$.

Let $\lambda^{(r)}_{x_0,\tau}$ denote the localized measure, and define $P_{x_0,\tau} := (E_{\#}\lambda^{(r)}_{x_0,\tau}) * \gamma_\tau$, $Q_{x_0,\tau} := ((F_{x_0})_{\#}\lambda^{(r)}_{x_0,\tau}) * \gamma_\tau$, and the local affine residual $\epsilon_{x_0,\tau} := \sup_{x\in\operatorname{supp}\lambda^{(r)}_{x_0,\tau}} \|E(x) - F_{x_0}(x)\|$. If $\epsilon_{x_0,\tau} \le a_0\sqrt{\tau}$, then $|1 - \mathcal{R}(P_{x_0,\tau})/\mathcal{R}(Q_{x_0,\tau})| \le (C/b_{x_0})\epsilon_{x_0,\tau}/\sqrt{\tau}$, where $C = C(d, r, \delta_*, a_0)$, and $b_{x_0} := \max\{d - m, d/[1 + r^2(1 + \delta_*)]^2\}$.

The theorem further relates the $\epsilon_{x_0,\tau}$ to the geometry of the encoder and data manifold. When the encoder differential and second derivative are locally bounded by J and H, respectively, $\epsilon_{x_0,\tau}$ is of order $r^2\tau(H + J/\rho_{\mathcal{M}})$. Consequently, whenever $\frac{r^2\sqrt{\tau}}{2}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right) \le a_0$,

$$\left|1 - \frac{\mathcal{R}(P_{x_0,\tau})}{\mathcal{R}(Q_{x_0,\tau})}\right| \le \frac{Cr^2}{2b_{x_0}}\sqrt{\tau}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right).$$

Combined local data-versus-latent bound. Corollary B.11, stated and proved in Appendix B.9, combines encoder nonlinearity and manifold curvature to compare localized data-space and latent-space FIRs directly. Let $P^X_{x_0,\tau}$ denote the Gaussian-smoothed law of the centered local data $x - x_0$ in $\mathbb{R}^D$, and let $P^Z_{x_0,\tau}$ denote the Gaussian-smoothed law of the centered encoded data $E(x) - E(x_0)$ in $\mathbb{R}^d$. Under the local near-isometry condition $\delta_{x_0} \le \delta_* < 1$ and the diffusion-scale assumptions detailed in the appendix, there exist positive constants C_A, C_E, and C_M such that

$$\left|\tau^2\left[\mathcal{R}(P^X_{x_0,\tau}) - \mathcal{R}(P^Z_{x_0,\tau})\right] - (D - d)\right| \le C_A\delta_{x_0} + C_E r^2\sqrt{\tau}\,H + C_M r^2\sqrt{\tau}\frac{1 + J}{\rho_{\mathcal{M}}}.$$

Thus, $D - d$ is the flat normal-direction compression baseline, while the three terms on the right quantify tangential distortion, encoder nonlinearity, and manifold curvature, respectively.

Geometric Interpretation of FIR Deviation. Corollary B.11 combines the comparison of the curved manifold with its tangent plane, the near-isometric linear encoding of the tangent distribution, and the nonlinear encoder with its tangent-affine approximation. Together with Theorem 3.3, this identifies the following four geometric penalties.

1. Encoder-Induced Distortions: Our bounds identify local near-isometry and controlled encoder curvature as sufficient controls on FIR distortion. These controls are represented by two terms:

- *Tangential Distortion (δ):* Local metric stretching or compression (Figure 3(b)).
- *High-Frequency Penalty ($\epsilon/\sqrt{\tau}$):* This quantity compares the encoder displacement with the diffusion length scale. A fixed-amplitude distortion becomes more visible as τ decreases. For a C^2 encoder on an $O(\sqrt{\tau})$ neighborhood, the Taylor residual is $O(\tau)$, producing a local FIR penalty of order $O(\sqrt{\tau})$. Persistent low-noise inflation may therefore indicate nonsmooth behavior, increasing fine-scale curvature, global folding, or interactions between multiple local branches.

2. Intrinsic Data Penalties: The geometry of the data introduces its own penalties.

- *Dimension Mismatch ($D - d$):* The exact normal-direction contribution of dimensional compression on flat supports and the leading local compression contribution on curved supports.
- *Curvature-Dependent Penalty ($O(r^2\sqrt{\tau}(1 + J)/\rho_{\mathcal{M}})$):* The deviation of the manifold from its tangent plane produces a term proportional to $r^2\sqrt{\tau}/\rho_{\mathcal{M}}$. The corresponding latent-space error is additionally amplified by the local encoder norm J. The resulting term combines intrinsic manifold curvature with the encoder's sensitivity to normal displacement.

4 Experiments

We compare FI, FIR, and FIR deviations for pixel- and latent-space diffusion models. We begin with Gaussian toy data and simple encoders to illustrate the geometric mechanisms studied in §3 and to compare with exact population formulas where available. We then move to real image data and more complex autoencoders. Implementation details are provided in Appendix D, and additional FFHQ experiments with VAE, KL-AE, and NVAE encodings are reported in Appendix F.

4.1 FI and FIR Comparison in the Gaussian Toy Setting

Experimental Setup. We consider a 2D Gaussian dataset and its pointwise nonlinear encodings. Specifically, we generate training data $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ and construct encoded samples $\mathbf{z} = E(\mathbf{x})$,

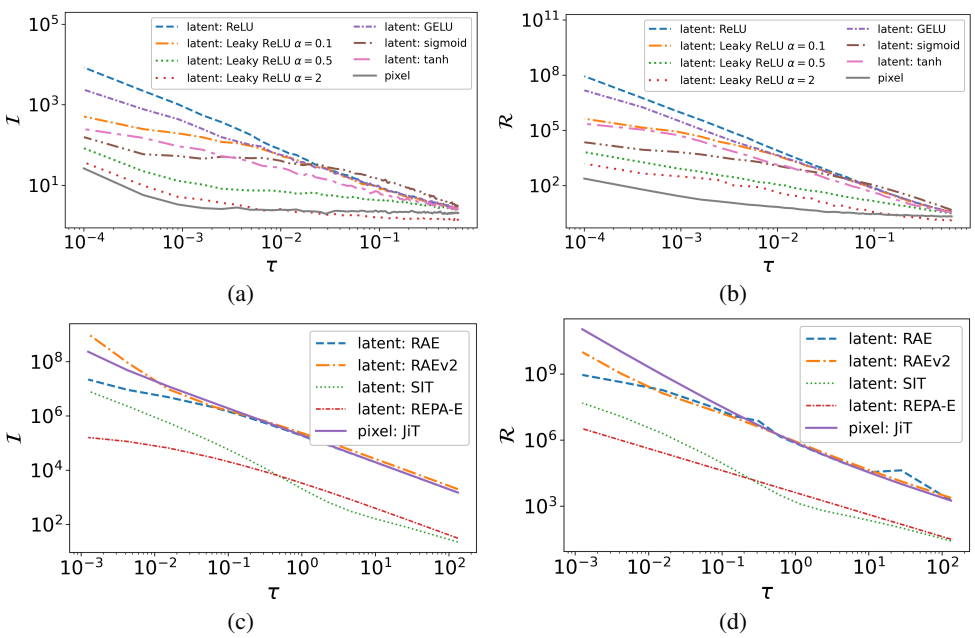


Figure 1: Values of $\mathcal{I}$ (left) and $\mathcal{R}$ (right) plotted against the noise variance τ. (a,b) Gaussian toy experiments using tiny diffusion models. *Pixel* denotes models trained on $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, while *latent* denotes models trained on $\mathbf{z} = E(\mathbf{x})$, with the pointwise activation E indicated in the legend. (c,d) Results for recent image generative models. JiT operates directly in pixel space, whereas SiT, RAE, RAEv2, and REPA-E operate in latent space. For (c,d), we show $\sqrt{\tau} \in [0.01, 80]$, excluding smaller noise levels because of numerical instability.

where the encoder E is applied pointwise to each coordinate of $\mathbf{x}$. We then train a tiny diffusion model[1] with identical architecture and training settings separately on the original data $\mathbf{x}$ and the encoded data $\mathbf{z}$. The numerical estimation details are presented in Appendix C.

Results. Figure 1(a,b) compares FI ($\mathcal{I}$) and FIR ($\mathcal{R}$) computed using diffusion models trained on the original data $\mathbf{x}$ and encoded data $\mathbf{z} = E(\mathbf{x})$. We consider encoders given by common activation functions, including ReLU, Leaky ReLU with negative slope α, GELU, sigmoid, and tanh. An expanded comparison of ReLU and Leaky ReLU over additional values of α is provided in Figure 4 in Appendix E. Appendix E.1.1 also gives the exact Gaussian-smoothed population density and a model-free FI formula for ReLU and Leaky ReLU. More generally, Lemma B.3 guarantees that every two-dimensional Gaussian-smoothed latent law satisfies $\mathcal{I}^{(2)}(\mu_{Z,\tau}) \leq 2/\tau$.

Figure 1(a) compares FI across encoders against the pixel baseline. ReLU exhibits the largest observed deviation. Leaky ReLU is globally bi-Lipschitz for every $\alpha > 0$, and its metric constants approach those of an isometry as $\alpha \to 1$. Correspondingly, its empirical FI curve approaches the pixel baseline, as further illustrated by the expanded comparison in Figure 4. However, the exact Leaky-ReLU map has an interface jump in its Jacobian determinant, so Proposition B.1 does not directly predict the Gaussian-smoothed FI curves. The exact expression (38) provides the appropriate population benchmark for this experiment. Sigmoid and tanh do not possess a positive global lower Lipschitz constant on the support of a Gaussian input, and GELU is not globally injective. We therefore interpret their FI curves empirically rather than as direct applications of Proposition B.1.

Figure 1(b) compares FIR across encoders. ReLU again exhibits the largest observed deviation. ReLU and Leaky ReLU are piecewise linear, with distributional second derivatives concentrated at the activation axes, and therefore fall outside the C^2 assumptions of Theorem B.8. For Leaky ReLU, the strength of the interface singularity is proportional to $|1 - \alpha|$, which decreases as α approaches 1, consistent with the observed FIR trend.

[1]We use the open-source implementation available at `https://github.com/tanelp/tiny-diffusion`.

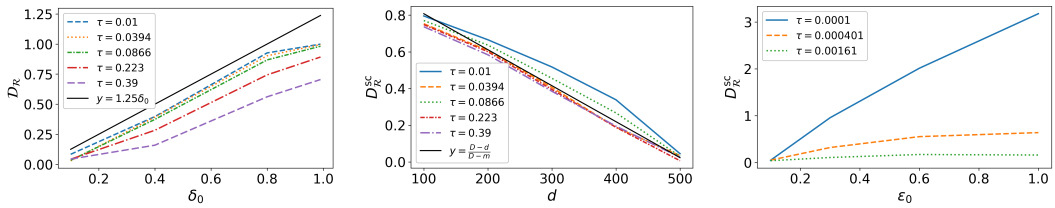


Figure 2: FIR deviation $\mathcal{D}_{\mathcal{R}}$ vs. (a) δ_0, (b) d, and (c) ε_0 in toy settings. Data $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ are embedded as $\mathbf{x} = (y_1, y_2, 0, \ldots, 0) \in \mathbb{R}^D$ and encoded as $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^d$. We compute $\mathcal{D}_{\mathcal{R}}$ from $\mathcal{R}^{(D)}(\mu_\tau)$ and $\mathcal{R}^{(d)}((\mu_Z)_\tau)$ using diffusion models trained on $\mathbf{x}$ and $\mathbf{z}$. Curves correspond to fixed noise variance τ. The solid lines $y = 1.25\delta_0$ in (a) and $y = (D-d)/(D-2)$ in (b) serve as linear references. In (a), $D = d = 2$ and $E(\mathbf{x}) = \mathbf{A}\mathbf{x}$, where $\mathbf{A} = \mathrm{diag}(\sqrt{1+\delta_0}, \sqrt{1-\delta_0})$. In (b), $D = 512$ and $\mathbf{z} = (y_1, y_2, 0, \ldots, 0) \in \mathbb{R}^d$. In (c), $D = d = 3$ and $E((x_1, x_2, 0), \varepsilon_0) = \big(\sin(\varepsilon_0 x_1)/\varepsilon_0,\ x_2,\ (1 - \cos(\varepsilon_0 x_1))/\varepsilon_0\big)$.

4.2 FIR Deviation in the Gaussian Toy Setting

Experimental Setup. Consider an intrinsic manifold of dimension $m = 2$ with data $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$. We embed $\mathbf{y}$ into $\mathbf{x} = (y_1, y_2, 0, \ldots, 0) \in \mathbb{R}^D$, and construct latent representations $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^d$, where $D, d \geq 2$. Using the same tiny diffusion model, training protocol, and FIR estimation procedure described in §4.1, we compute $\mathcal{R}^{(D)}(\mu_\tau)$ and $\mathcal{R}^{(d)}((\mu_Z)_\tau)$ using diffusion models trained on $\mathbf{x}$ and $\mathbf{z}$, respectively, and use these values to compute the FIR deviations $\mathcal{D}_{\mathcal{R}}(\tau)$ and $\mathcal{D}^{\mathrm{sc}}_{\mathcal{R}}(\tau)$.

Results. For all experiments in this section, exact Gaussian calculations for the two linear settings and additional plots of $\mathcal{D}_{\mathcal{R}}$ versus the noise variance τ are provided in Appendix E. The geometric parameters for the nonlinear setting are also discussed there.

We begin with the case $D = d = m = 2$. The encoded data are obtained via the linear transformation $\mathbf{z} = \mathbf{A}\mathbf{x}$, where $\mathbf{A} = \mathrm{diag}\left(\sqrt{1+\delta_0}, \sqrt{1-\delta_0}\right)$. In this setting, $\|\mathbf{A}^\top \mathbf{A} - \mathbf{I}_2\|_2 = \delta_0$. Figure 2(a) shows $\mathcal{D}_{\mathcal{R}}$ as a function of δ_0 for fixed noise variances τ. We report curves at relatively small noise variances τ, since for large τ the FIR values become very small due to heavy smoothing, making the relative deviation numerically unstable and less informative. The empirical curves increase monotonically and appear approximately linear over much of the tested range. The exact population calculation in (41) shows, however, that the deviation is quadratic near $\delta_0 = 0$ and satisfies $\mathcal{D}_{\mathcal{R}}(\tau) \leq 3\delta_0^2 \leq 3\delta_0$. Thus, the experiment is consistent with an $\mathcal{O}(\delta_0)$ stability bound, although its exact small-δ_0 behavior is sharper.

Next, we consider $D = 512$ and varying d. We take $\mathbf{z} = (y_1, y_2, 0, \ldots, 0) \in \mathbb{R}^d$. Here, the encoder acts as the identity on the intrinsic subspace, so $\mathbf{A} = \mathbf{I}_m$. Equation (42) gives $\mathcal{D}^{\mathrm{sc}}_{\mathcal{R}}(\tau) \leq \frac{D-d}{D-m}$. Figure 2(b) reports $\mathcal{D}^{\mathrm{sc}}_{\mathcal{R}}$ as a function of d. The observed downward trend is consistent with this exact upper bound. Finally, we consider $D = d = 3$ and $\mathbf{z} = E(\mathbf{x}, \varepsilon_0)$, where $E(\mathbf{x}, \varepsilon_0)$ is a smooth nonlinear map that locally bends the 2D data manifold into a 3D cylinder of radius $1/\varepsilon_0$, defined as $E((x_1, x_2, 0), \varepsilon_0) = \big(\sin(\varepsilon_0 x_1)/\varepsilon_0,\ x_2,\ (1 - \cos(\varepsilon_0 x_1))/\varepsilon_0\big)$ with $\varepsilon_0 > 0$. For this construction, the tangential Jacobian is an isometry, so the local near-isometry parameter satisfies $\delta_{x_0} = 0$. Moreover, $\|D^2 E\|_{\mathrm{op}} = \varepsilon_0$, while the input plane has infinite reach. Therefore, for the $r\sqrt{\tau}$-localized model of Theorem B.8, the local Taylor residual satisfies $\epsilon_{x_0,\tau} \leq \frac{r^2\tau}{2}\varepsilon_0$. Consequently, the theorem gives the local nonlinear-versus-affine bound $\left|1 - \frac{\mathcal{R}(P_{x_0,\tau})}{\mathcal{R}(Q_{x_0,\tau})}\right| \leq Cr^2\varepsilon_0\sqrt{\tau}$.

Figure 2(c) evaluates the global FIR deviation for an unbounded Gaussian input and therefore is not a direct numerical verification of this localized theorem. At each fixed noise level, however, the observed approximately linear dependence on ε_0 is qualitatively consistent with the theorem's linear dependence on the encoder-curvature magnitude. The linear trend is most evident for small τ. For larger τ, the deviation values are much smaller, which may make the FIR estimation error more prominent and thus visually weaken the linear pattern.

Table 1: Generation, geometric, and spectral evaluation results. Generation quality is evaluated using FID, spatial FID (sFID), Inception Score (IS), precision, and recall. Straightness and efficiency measure the geometry of ODE sampling trajectories, while SEC measures the fraction of latent spectral energy outside the low-frequency region, with subscripts indicating the frequency threshold. Geometric and spectral metrics are reported only for the latent diffusion models.

	Generation Metrics					Geometric and Spectral Metrics			
Model	FID↓	sFID↓	IS↑	Precision↑	Recall↑	Straightness↑	Efficiency↑	$\text{SEC}_{0.25}$ ↓	$\text{SEC}_{0.5}$ ↓
RAE	2.432	8.340	243.794	0.720	0.640	0.999	0.982	0.277	0.112
RAEv2	2.370	7.458	225.068	0.730	0.633	0.999	0.949	0.309	0.137
SiT	7.945	9.442	123.962	0.594	0.678	1.000	0.851	0.411	0.209
REPA-E	3.496	7.973	159.359	0.692	0.630	1.000	0.875	0.312	0.102
JiT	3.169	8.449	65.185	0.692	0.631	–	–	–	–

4.3 FI and FIR Comparison across ImageNet Latent Diffusion Models

Experimental Setup. We evaluate five recent image-generation models on ImageNet at 256×256 resolution: RAE (Zheng et al., 2025), RAEv2 (Singh et al., 2026), SiT (Ma et al., 2024), REPA-E (Leng et al., 2025), and JiT (Li & He, 2026). JiT performs generation directly in pixel space, whereas the other methods operate on latent representations produced by different image encoders. These models are selected because they use closely related transformer-based generative architectures of comparable scale, allowing us to examine how the choice of representation affects FI and FIR. We use the official implementations and publicly available pretrained checkpoints.

Results. Figure 1(c,d) compares the FI and FIR curves across the evaluated models. Among the latent diffusion models, RAE and RAEv2 are substantially closer to the pixel-space JiT baseline in both FI and FIR. In contrast, SiT and REPA-E show consistently lower FI/FIR values and larger deviations from the pixel-space reference. This suggests that the RAE-based latent spaces retain more of the diffusion-relevant structure of the original image space, leading to latent distributions whose complexity more closely matches the pixel-space baseline.

Table 1 reports both generation performance and geometric diagnostics of latent diffusability (Zhong et al., 2026). RAE and RAEv2 achieve the best FID scores and highest Inception Scores, while maintaining strong sFID and precision. They also exhibit the highest trajectory efficiency and relatively low SEC values, suggesting more direct ODE transport and smoother latent spectral structure. These results are broadly consistent with the FI/FIR comparison, in which RAE and RAEv2 are the latent models closest to the pixel-space JiT baseline, suggesting that FI and FIR capture properties of latent representations relevant to diffusion-based generation. The agreement between FI/FIR, generation quality, and latent-space diagnostics may partly reflect encoder choice. The DINO-style representation encoders used by RAE and RAEv2 may preserve semantic and local neighborhood structure better than the VAE-style encoders used by SiT and REPA-E, yielding more diffusible latent spaces.

5 Conclusions

We introduced a formal information-geometric framework to demystify the "diffusability" of latent spaces in generative modeling. By decomposing denoising complexity into FI and FIR, we identified geometric requirements for successful latent diffusion. Our theoretical analysis explicitly decouples the effects of dimensional compression from intrinsic geometric distortion, revealing that latent-space degradation is primarily driven by local curvature and nonlinearity injected by the encoder.

Specifically, we identified near-isometry and controlled local nonlinearity as complementary geometric requirements for preserving the FI- and FIR-based components of denoising complexity. Our toy, FFHQ, and ImageNet-based experiments show that empirical FI and FIR track several measures of generation quality and latent-space geometry in the settings tested.

FI and FIR are population quantities defined independently of a trained diffusion model. The current scalable estimators, however, generally use scores obtained from trained models, while direct empirical-mixture estimation may become unreliable in high dimension. Developing statistically reliable and scalable pre-training estimators therefore remains an important direction for future work.

AI USE STATEMENT

Generative AI tools were used only to aid and polish the writing of this manuscript. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Shun'ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, Tokyo, 2016. ISBN 978-4-431-55977-1. doi: 10.1007/978-4-431-55978-8.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Y7pYxRb.

Alan Baade, Eric Ryan Chan, Kyle Sargent, Changan Chen, Justin Johnson, Ehsan Adeli, and Li Fei-Fei. Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. *arXiv preprint arXiv:2602.11401*, 2026.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pp. 177–206. Springer, 1985.

Nutan Chen, Alexej Klushyn, Francesco Ferroni, Justin Bayer, and Patrick Van Der Smagt. Learning flat latent manifolds with VAEs. In *International Conference on Machine Learning*, 2020.

Seunghyuk Cho, Juyong Lee, and Dongwoo Kim. Hyperbolic VAE via latent Gaussian distributions. *Advances in Neural Information Processing Systems*, 36:569–588, 2023.

Max Costa. A new entropy power inequality. *IEEE Transactions on Information Theory*, 31(6): 751–760, 1985.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181. URL https://doi.org/10.1198/jasa.2011.tm11181. PMID: 22505788.

Dongning Guo, S. Shamai, and S. Verdu. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005. doi: 10.1109/TIT.2005.844072.

Jonathan Heek, Emiel Hoogeboom, Thomas Mensink, and Tim Salimans. Unified latents (ul): How to train your latents. *arXiv preprint arXiv:2602.17270*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Rafal Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas K Garg. The spacetime of diffusion models: An information geometry perspective. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qCsbYJZRA5.

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*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *Forty-second International Conference on Machine Learning*, 2025.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Hyomin Lee, Minseon Kim, Sangwon Jang, Jongheon Jeong, and Sung Ju Hwang. Enhancing variational autoencoders with smooth robust latent encoding. *arXiv preprint arXiv:2504.17219*, 2025a.

Junho Lee, Jeongwoo Shin, Hyungwook Choi, and Joonseok Lee. Latent diffusion models with masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17422–17431, 2025b.

Wonjun Lee, Riley C. W. O'Neill, Dongmian Zou, Jeff Calder, and Gilad Lerman. Geometry-preserving encoder/decoder in latent generative models, 2025c. URL https://arxiv.org/abs/2501.09876.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18262–18272, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Alexander Lobashev, Dmitry Guskov, Maria Larchenko, and Mikhail Tamm. Hessian geometry of latent space in generative models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=H8JTsbG4KW.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Alessandro Palma, Sergei Rybakov, Leon Hetzel, Stephan Günnemann, and Fabian J Theis. Enforcing latent euclidean geometry in single-cell VAEs for manifold interpolation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DoDXFkF10S.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

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 (CVPR)*, pp. 10684–10695, June 2022.

Haitz Sáez de Ocáriz Borde, Alvaro Arroyo, Ismael Morales, Ingmar Posner, and Xiaowen Dong. Neural latent geometry search: Product manifold inference via Gromov-Hausdorff-informed Bayesian optimization. *Advances in Neural Information Processing Systems*, 36:38370–38403, 2023.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. *arXiv preprint arXiv:2510.15301*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2hEDcA7xy4.

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. URL https://openreview.net/forum?id=PxTIG12RRHS.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2009.

Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good visual tokenizers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1jBsi98fVe.

Kaiyue Yue, Menglin Jia, Junlin Hou, and Tom Goldstein. Image generation with a sphere encoder, 2026. URL https://arxiv.org/abs/2602.15030.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0u1LigJaab.

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability. *arXiv preprint arXiv:2606.03578*, 2026.

APPENDIX

The appendices are organized as follows: Appendix A reviews related literature, while Appendix B contains all missing proofs and derivations. Practical estimation of FI and FIR is detailed in Appendix C, and Appendix D provides comprehensive experimental setups. Additional mathematical derivations and numerical supplements, relevant to the toy experimental settings, are provided in Appendix E. Finally, Appendices F.1.2–G present further numerical results and spectral analyses, relevant to the real experimental settings.

A RELATED WORKS

The term "diffusability" was first introduced by Skorokhodov et al. (Skorokhodov et al., 2025) to describe the extent to which a latent space facilitates the diffusion process. The authors argued that high-frequency components in latent representations reduce diffusability, though they did not provide a rigorous quantitative measure for this effect. Subsequently, (Zheng et al., 2026) offered further empirical evidence demonstrating that standard representation encoders often induce latent distributions that are inherently difficult for diffusion models to fit.

Several follow-up works have aimed to improve diffusability through specific architectural or training modifications. For instance, (Kouzelis et al., 2025) promoted diffusability by enforcing equivariance in the latent representation, while (Yang et al., 2026) targeted the corruption robustness of the latents. Taking a different angle, (Heek et al., 2026) improved diffusability by training the encoder to produce latents that already resemble mildly noised diffusion states, allowing the diffusion prior to denoise from a smoother, better-aligned distribution. Rather than modifying the latent distribution directly, (Baade et al., 2026) proposed denoising latent features early in the pipeline and using them to condition later pixel-space denoising, effectively simplifying the overall problem.

Another distinct line of work focuses on explicitly constraining the latent geometry. In (Yue et al., 2026), latent codes are normalized to lie uniformly on a hypersphere. More broadly, alternative geometric priors for autoencoders have gained traction, including the use of hyperbolic latent spaces (Cho et al., 2023; Sáez de Ocáriz Borde et al., 2023).

While these prior works have successfully improved diffusability through empirical regularization and structural constraints, they largely lack a unified mathematical framework to quantify why certain representation spaces are easier to model than others. In contrast, our work provides a rigorous geometric analysis of diffusability. By explicitly bounding the Fisher Information Rate (FIR) of the encoded distributions, we establish formal guarantees on how linear encoders and geometric constraints fundamentally alter the stability of the diffusion process.

B SUPPLEMENTARY DERIVATIONS AND PROOFS

B.1 PROOF OF PROPOSITION 2.1

Fix $\tau > 0$. Let $p_\tau(\mathbf{x})$ be the smooth density of the noisy measure μ_τ. As defined in §2, the density is given by the convolution of the clean measure μ with the Gaussian kernel

$$\phi_\tau(\mathbf{y}) = (2\pi\tau)^{-k/2} \exp(-\|\mathbf{y}\|^2/(2\tau)) \tag{7}$$

as follows:

$$p_\tau(\mathbf{x}) = (\mu * \phi_\tau)(\mathbf{x}) = \int_{\mathbb{R}^k} \phi_\tau(\mathbf{x} - \mathbf{x}_0)\, d\mu(\mathbf{x}_0). \tag{8}$$

In view of this expression,

$$p_\tau \in C^\infty(\mathbb{R}^k), \qquad p_\tau > 0.$$

The integrations by parts below are justified using compactly supported cutoff functions and the standing regularity assumptions stated in the main text; they do not require p_τ itself to have Gaussian tails. We express the Fisher Information as:

$$\mathcal{I}^{(k)}(\mu_\tau) = \int_{\mathbb{R}^k} p_\tau(\mathbf{x}) \|\nabla_{\mathbf{x}} \log p_\tau(\mathbf{x})\|^2 \, d\mathbf{x} = \int_{\mathbb{R}^k} \frac{\|\nabla_{\mathbf{x}} p_\tau(\mathbf{x})\|^2}{p_\tau(\mathbf{x})} \, d\mathbf{x}. \tag{9}$$

Using the probabilistic heat equation $\frac{\partial p_\tau}{\partial \tau} = \frac{1}{2}\Delta_{\mathbf{x}} p_\tau$ from (1), we differentiate $\mathcal{I}^{(k)}$ with respect to τ:

$$\begin{aligned}\frac{d}{d\tau}\mathcal{I}^{(k)}(\mu_\tau) &= \int_{\mathbb{R}^k} \left(\frac{2\nabla_{\mathbf{x}} p_\tau \cdot \nabla_{\mathbf{x}}(\partial_\tau p_\tau)}{p_\tau} - \frac{\|\nabla_{\mathbf{x}} p_\tau\|^2}{p_\tau^2} \partial_\tau p_\tau \right) d\mathbf{x} \\ &= \int_{\mathbb{R}^k} \left(\frac{\nabla_{\mathbf{x}} p_\tau \cdot \nabla_{\mathbf{x}}(\Delta_{\mathbf{x}} p_\tau)}{p_\tau} - \frac{1}{2}\frac{\|\nabla_{\mathbf{x}} p_\tau\|^2}{p_\tau^2} \Delta_{\mathbf{x}} p_\tau \right) d\mathbf{x}. \end{aligned} \tag{10}$$

Integrating the first term by parts using a compactly supported cutoff and then passing to the limit:

$$\begin{aligned}\int_{\mathbb{R}^k} \frac{\nabla_{\mathbf{x}} p_\tau \cdot \nabla_{\mathbf{x}}(\Delta_{\mathbf{x}} p_\tau)}{p_\tau}\, d\mathbf{x} &= -\int_{\mathbb{R}^k} \Delta_{\mathbf{x}} p_\tau \, \operatorname{div}\left(\frac{\nabla_{\mathbf{x}} p_\tau}{p_\tau} \right) d\mathbf{x} \\ &= -\int_{\mathbb{R}^k} \Delta_{\mathbf{x}} p_\tau \left(\frac{\Delta_{\mathbf{x}} p_\tau}{p_\tau} - \frac{\|\nabla_{\mathbf{x}} p_\tau\|^2}{p_\tau^2} \right) d\mathbf{x}. \end{aligned} \tag{11}$$

Substituting this back into (10) yields:

$$\frac{d}{d\tau}\mathcal{I}^{(k)}(\mu_\tau) = \int_{\mathbb{R}^k} \left[-\frac{(\Delta_{\mathbf{x}} p_\tau)^2}{p_\tau} + \frac{1}{2}\Delta_{\mathbf{x}} p_\tau \frac{\|\nabla_{\mathbf{x}} p_\tau\|^2}{p_\tau^2} \right] d\mathbf{x}. \tag{12}$$

To simplify, let $f = \log p_\tau$, $A = \Delta_{\mathbf{x}} f$, and $B = \|\nabla_{\mathbf{x}} f\|^2$. Noting that $\frac{\Delta_{\mathbf{x}} p_\tau}{p_\tau} = \Delta_{\mathbf{x}} \log p_\tau + \|\nabla_{\mathbf{x}} \log p_\tau\|^2 = A + B$, we have:

$$\begin{aligned}\frac{d}{d\tau}\mathcal{I}^{(k)}(\mu_\tau) &= \int_{\mathbb{R}^k} p_\tau \left[-(A+B)^2 + \frac{1}{2}(A+B)B \right] d\mathbf{x} \\ &= \int_{\mathbb{R}^k} p_\tau \left[-A^2 - \frac{3}{2}AB - \frac{1}{2}B^2 \right] d\mathbf{x}. \end{aligned} \tag{13}$$

Now, let $\mathbf{H}_\tau = \nabla_{\mathbf{x}}^2 \log p_\tau$ be the Hessian. The Euclidean Bochner identity for $f = \log p_\tau$ states:

$$\frac{1}{2}\Delta_{\mathbf{x}} \|\nabla_{\mathbf{x}} f\|^2 = \|\nabla_{\mathbf{x}}^2 f\|_F^2 + \nabla_{\mathbf{x}} f \cdot \nabla_{\mathbf{x}}(\Delta_{\mathbf{x}} f). \tag{14}$$

Integrating $\frac{1}{2}\Delta_{\mathbf{x}} B$ and $-\nabla_{\mathbf{x}} f \cdot \nabla_{\mathbf{x}} A$ against the density p_τ via integration by parts, and noting that $\operatorname{div}(p_\tau \nabla_{\mathbf{x}} f) = \nabla_{\mathbf{x}} p_\tau \cdot \nabla_{\mathbf{x}} f + p_\tau \Delta_{\mathbf{x}} f = p_\tau(B + A)$, we obtain:

$$\begin{aligned}\int_{\mathbb{R}^k} p_\tau \frac{1}{2}\Delta_{\mathbf{x}} B \, d\mathbf{x} = \int_{\mathbb{R}^k} \frac{1}{2} B \Delta_{\mathbf{x}} p_\tau \, d\mathbf{x} = \int_{\mathbb{R}^k} p_\tau \left(\frac{1}{2}AB + \frac{1}{2}B^2 \right) d\mathbf{x}, \\ -\int_{\mathbb{R}^k} p_\tau \nabla_{\mathbf{x}} f \cdot \nabla_{\mathbf{x}} A \, d\mathbf{x} = \int_{\mathbb{R}^k} A \, \operatorname{div}(p_\tau \nabla_{\mathbf{x}} f)\, d\mathbf{x} = \int_{\mathbb{R}^k} p_\tau (A^2 + AB)\, d\mathbf{x}. \end{aligned}$$

Summing these yields the weighted Bochner identity:

$$\mathbb{E}_{p_\tau}[\|\mathbf{H}_\tau\|_F^2] = \mathbb{E}_{p_\tau}\left[A^2 + \frac{3}{2}AB + \frac{1}{2}B^2 \right]. \tag{15}$$

Comparing this to (13), we find:

$$\frac{d}{d\tau}\mathcal{I}^{(k)}(\mu_\tau) = -\int_{\mathbb{R}^k} p_\tau \|\mathbf{H}_\tau\|_F^2 \, d\mathbf{x} = -\mathbb{E}_{p_\tau}\left[\operatorname{Tr}(\mathbf{H}_\tau^2) \right]. \tag{16}$$

By (3), $\mathcal{R}^{(k)}(\mu_\tau) = -\frac{d}{d\tau}\mathcal{I}^{(k)}(\mu_\tau)$. Substituting the score $\mathbf{s}_\tau(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\tau$, we have $\nabla_{\mathbf{x}} \mathbf{s}_\tau = \mathbf{H}_\tau$, and $\mathcal{R}^{(k)}(\mu_\tau) = \mathbb{E}_{p_\tau}[\|\nabla_{\mathbf{x}} \mathbf{s}_\tau(\mathbf{x})\|_F^2]$, which completes the proof.

B.2 Derivation of (5) and (6)

In this section, we provide the detailed derivation for the I-MMSE identity and the Denoising Resistance decomposition. Consider a clean signal $\mathbf{x}_0 \in \mathbb{R}^k$ drawn from a distribution μ, and its corruption by additive white Gaussian noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, resulting in the noisy observation:

$$\mathbf{x} = \mathbf{x}_0 + \sqrt{\tau}\mathbf{n},$$

where τ is the noise variance. The total variance of the corruption is $\mathbb{E}[\|\mathbf{x}_0-\mathbf{x}\|^2] = \mathbb{E}[\|-\sqrt{\tau}\mathbf{n}\|^2] = \tau k$.

Let $\hat{\mathbf{x}}_0(\mathbf{x}) = \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}]$ be the MMSE estimator. We can decompose the total noise variance by adding and subtracting this estimator:

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}\|^2] &= \mathbb{E}[\|(\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x})) + (\hat{\mathbf{x}}_0(\mathbf{x}) - \mathbf{x})\|^2] \\ &= \mathbb{E}[\|\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x})\|^2] + \mathbb{E}[\|\hat{\mathbf{x}}_0(\mathbf{x}) - \mathbf{x}\|^2] + 2\mathbb{E}[(\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}))^\top(\hat{\mathbf{x}}_0(\mathbf{x}) - \mathbf{x})]. \quad (17)\end{aligned}$$

The cross-term vanishes by the tower property of conditional expectation. Since $(\hat{\mathbf{x}}_0(\mathbf{x}) - \mathbf{x})$ is deterministic given $\mathbf{x}$, we have:

$$\begin{aligned}\mathbb{E}[(\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}))^\top(\hat{\mathbf{x}}_0(\mathbf{x}) - \mathbf{x})] &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}[(\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}))^\top(\hat{\mathbf{x}}_0(\mathbf{x}) - \mathbf{x}) \mid \mathbf{x}]\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[(\mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}] - \hat{\mathbf{x}}_0(\mathbf{x}))^\top(\hat{\mathbf{x}}_0(\mathbf{x}) - \mathbf{x})\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[(\hat{\mathbf{x}}_0(\mathbf{x}) - \hat{\mathbf{x}}_0(\mathbf{x}))^\top(\hat{\mathbf{x}}_0(\mathbf{x}) - \mathbf{x})\right] = 0. \quad (18)\end{aligned}$$

Next, we invoke Tweedie's formula (Efron, 2011), which relates the residual of the MMSE estimator to the score function $\mathbf{s}_\tau(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\tau(\mathbf{x})$:

$$\hat{\mathbf{x}}_0(\mathbf{x}) - \mathbf{x} = \tau \nabla_{\mathbf{x}} \log p_\tau(\mathbf{x}) = \tau \mathbf{s}_\tau(\mathbf{x}). \quad (19)$$

Substituting (19) and the definition $\text{MMSE}^{(k)}(\mu_\tau) = \mathbb{E}[\|\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x})\|^2]$ into (17) yields:

$$\tau k = \text{MMSE}^{(k)}(\mu_\tau) + \tau^2 \mathbb{E}_{p_\tau}[\|\mathbf{s}_\tau(\mathbf{x})\|^2] = \text{MMSE}^{(k)}(\mu_\tau) + \tau^2 \mathcal{I}^{(k)}(\mu_\tau). \quad (20)$$

Rearranging for $\text{MMSE}^{(k)}(\mu_\tau)$ directly recovers the I-MMSE identity (5) as established in (Guo et al., 2005).

To derive the Denoising Resistance, we differentiate (5) with respect to the noise variance τ using the product rule:

$$\frac{d}{d\tau}\text{MMSE}^{(k)}(\mu_\tau) = \frac{d}{d\tau}\left(k\tau - \tau^2 \mathcal{I}^{(k)}(\mu_\tau)\right) = k - 2\tau \mathcal{I}^{(k)}(\mu_\tau) - \tau^2 \frac{d}{d\tau}\mathcal{I}^{(k)}(\mu_\tau). \quad (21)$$

Substituting our definition of the Fisher Information Rate (FIR) from (3), namely $\mathcal{R}^{(k)}(\mu_\tau) = -\frac{d}{d\tau}\mathcal{I}^{(k)}(\mu_\tau)$, we obtain the fundamental decomposition:

$$\frac{d}{d\tau}\text{MMSE}^{(k)}(\mu_\tau) = \underbrace{k - 2\tau \mathcal{I}^{(k)}(\mu_\tau)}_{\text{Intrinsic Noise Gain}} + \underbrace{\tau^2 \mathcal{R}^{(k)}(\mu_\tau)}_{\text{Geometric Complexity Penalty}}. \quad (22)$$

As noted in the main text, the *Intrinsic Noise Gain* represents the baseline linear increase in estimation error due to dimensionality k, attenuated by the current magnitude of information $\mathcal{I}^{(k)}$. Conversely, the *Geometric Complexity Penalty* represents the additional difficulty imposed by the local curvature of the density. By Proposition 2.1, this term is governed by $\mathbb{E}_{p_\tau}[\|\nabla \mathbf{s}_\tau\|_F^2]$, indicating that distributions with rapidly changing score functions (high curvature) exhibit a faster degradation in denoising performance as noise scales.

B.3 Proof of Proposition B.1

Proposition B.1 (Intrinsic Fisher Information under Bi-Lipschitz Encoders). *Let $\mathcal{M} \subset \mathbb{R}^D$ and $\mathcal{N} \subset \mathbb{R}^d$ be smooth m-dimensional manifolds, and let $E : \mathcal{M} \longrightarrow \mathcal{N}$ be a bi-Lipschitz homeomorphism onto $\mathcal{N}$. Assume that $c\, d_{\mathcal{M}}(x, y) \le d_{\mathcal{N}}(E(x), E(y)) \le C\, d_{\mathcal{M}}(x, y)$, for all $x, y \in \mathcal{M}$ and some constants $0 < c \le C < \infty$, where $d_{\mathcal{M}}$ and $d_{\mathcal{N}}$ denote the intrinsic Riemannian distances. By Rademacher's theorem in local charts, the tangential differential $DE(x) : T_x\mathcal{M} \longrightarrow T_{E(x)}\mathcal{N}$ exists for almost every $x \in \mathcal{M}$, and its singular values belong to $[c, C]$ almost everywhere. Define the tangential Jacobian $\mathcal{J}_E(x) := \sqrt{\det\left(DE(x)^\top DE(x)\right)}$ at points of differentiability. Let*

$$d\nu(x) = p_{\mathcal{M}}(x)\, d\,\text{vol}_{\mathcal{M}}(x),$$

where $p_{\mathcal{M}} > 0$ almost everywhere and

$$\log p_{\mathcal{M}} \in W^{1,1}_{\text{loc}}(\mathcal{M}).$$

Define the weak logarithmic score by

$$s_{\mathcal{M}} := \nabla_{\mathcal{M}} \log p_{\mathcal{M}},$$

and assume

$$s_{\mathcal{M}} \in L^2(\nu).$$

Define

$$\mathcal{I}_{\mathcal{M}} := \int_{\mathcal{M}} \|s_{\mathcal{M}}(x)\|^2 \, d\nu(x) < \infty.$$

Assume also that $\log \mathcal{J}_E \in W^{1,1}_{\mathrm{loc}}(\mathcal{M})$, *and define its weak tangential gradient by* $g_E := \nabla_{\mathcal{M}} \log \mathcal{J}_E$. *Suppose that*

$$\widetilde{\varepsilon}^2 := \int_{\mathcal{M}} \|g_E(x)\|^2 \, d\nu(x) < \infty.$$

Let $\widetilde{\nu} := E_{\#}\nu$. *Then* $\widetilde{\nu}$ *has finite intrinsic Fisher information on* $\mathcal{N}$, *and the following exact identity holds:*

$$\mathcal{I}_{\mathcal{N}} = \int_{\mathcal{M}} \left\| DE(x)^{-\top} \big(s_{\mathcal{M}}(x) - g_E(x)\big) \right\|^2 d\nu(x),$$

where $DE(x)^{-\top}$ *denotes the inverse adjoint between the tangent spaces.*

Consequently, for every $\theta \in (0,1)$,

$$\frac{1-\theta}{C^2}\mathcal{I}_{\mathcal{M}} - \frac{(\theta^{-1}-1)\widetilde{\varepsilon}^2}{C^2} \le \mathcal{I}_{\mathcal{N}} \le \frac{1+\theta}{c^2}\mathcal{I}_{\mathcal{M}} + \frac{(\theta^{-1}+1)\widetilde{\varepsilon}^2}{c^2}.$$

If E *is an isometry, then* $\mathcal{I}_{\mathcal{N}} = \mathcal{I}_{\mathcal{M}}$.

Proof. Let $\widetilde{p}_{\mathcal{N}}$ denote the density of $\widetilde{\nu} = E_{\#}\nu$ with respect to the volume measure on $\mathcal{N}$. The area formula for bi-Lipschitz maps gives, for almost every $x \in \mathcal{M}$,

$$\widetilde{p}_{\mathcal{N}}(E(x)) = \frac{p_{\mathcal{M}}(x)}{\mathcal{J}_E(x)}.$$

Hence

$$\log \widetilde{p}_{\mathcal{N}}(E(x)) = \log p_{\mathcal{M}}(x) - \log \mathcal{J}_E(x).$$

The Sobolev chain rule therefore gives

$$DE(x)^{\top} \widetilde{s}_{\mathcal{N}}(E(x)) = s_{\mathcal{M}}(x) - g_E(x)$$

for almost every x, where

$$\widetilde{s}_{\mathcal{N}} := \nabla_{\mathcal{N}} \log \widetilde{p}_{\mathcal{N}}.$$

Since $DE(x)$ is invertible between the tangent spaces almost everywhere,

$$\widetilde{s}_{\mathcal{N}}(E(x)) = DE(x)^{-\top} \big(s_{\mathcal{M}}(x) - g_E(x)\big).$$

Changing variables using $\widetilde{\nu} = E_{\#}\nu$ yields the exact identity

$$\begin{aligned} \mathcal{I}_{\mathcal{N}} &= \int_{\mathcal{N}} \|\widetilde{s}_{\mathcal{N}}(z)\|^2 \, d\widetilde{\nu}(z) \\ &= \int_{\mathcal{M}} \left\| DE(x)^{-\top} \big(s_{\mathcal{M}}(x) - g_E(x)\big) \right\|^2 d\nu(x). \end{aligned}$$

The singular values of $DE(x)$ belong to $[c, C]$ almost everywhere, so

$$\frac{1}{C}\|w\| \le \|DE(x)^{-\top} w\| \le \frac{1}{c}\|w\|.$$

For every $\theta \in (0,1)$,

$$\|a-b\|^2 \ge (1-\theta)\|a\|^2 - (\theta^{-1}-1)\|b\|^2$$

and

$$\|a-b\|^2 \le (1+\theta)\|a\|^2 + (\theta^{-1}+1)\|b\|^2.$$

Apply these inequalities with

$$a = s_{\mathcal{M}}(x), \qquad b = g_E(x),$$

and integrate to obtain the two stated bounds.

If E is an isometry, then

$$DE(x)^\top DE(x) = I_m, \qquad \mathcal{J}_E(x) = 1$$

almost everywhere. Thus $g_E = 0$ and

$$\mathcal{I}_{\mathcal{N}} = \mathcal{I}_{\mathcal{M}}.$$

□

Remark B.2 (Piecewise-smooth encoders and activation interfaces). The proposition allows weakly differentiable bi-Lipschitz encoders, but it still requires $\log \mathcal{J}_E$ to have an $L^2(\nu)$ weak gradient. For a piecewise-smooth encoder, this condition holds when its logarithmic volume Jacobian has no jump across the interfaces, together with the stated square-integrability on each smooth region.

If $\log \mathcal{J}_E$ has a nonzero jump across a codimension-one interface and the input density has a positive trace there, the deterministic pushforward density generally has a jump. Under the Sobolev definition of Fisher information compatible with the heat-semigroup identities, such a jump produces infinite Fisher information.

Leaky ReLU is globally bi-Lipschitz for every $\alpha > 0$, but for $\alpha \neq 1$ its Jacobian determinant jumps across the activation hyperplanes. It is therefore not covered directly by Proposition B.1, despite being bi-Lipschitz.

Gaussian smoothing nevertheless regularizes its latent law. More generally, as shown in Appendix B.4, if $\mu_Z := E_{\#}\nu$ is any latent probability measure, then Lemma B.3 gives

$$\mathcal{I}^{(d)}(\mu_{Z,\tau}) \leq \frac{d}{\tau}, \qquad \tau > 0.$$

Thus, the Gaussian-smoothed Leaky-ReLU latent distribution has finite FI at every positive noise variance, even though the unsmoothed deterministic pushforward may fail to have finite Sobolev FI. For the Gaussian toy setting, Appendix E.1.1 provides exact smoothed densities and a one-dimensional integral representation of the corresponding Fisher information.

B.4 Gaussian Smoothing and Fisher Information

The following result applies to an arbitrary probability measure and does not require the encoder or the unsmoothed latent law to be smooth.

Lemma B.3 (Gaussian smoothing regularizes Fisher information)**.** *Let η be an arbitrary Borel probability measure on $\mathbb{R}^d$, and let*

$$\eta_\tau := \eta * \phi_\tau^{(d)}, \qquad \tau > 0,$$

where $\phi_\tau^{(d)}$ is the centered Gaussian density with covariance τI_d. Then η_τ has a strictly positive C^∞ density p_τ, and

$$\mathcal{I}^{(d)}(\eta_\tau) \leq \frac{d}{\tau}.$$

In particular, the Gaussian-smoothed measure has finite Fisher information for every $\tau > 0$, without any moment or smoothness assumption on η.

Proof. Let $Z \sim \eta$ and $G \sim \mathcal{N}(\mathbf{0}, I_d)$ be independent, and define

$$Y = Z + \sqrt{\tau}\, G.$$

The density of Y is

$$p_\tau(y) = \int_{\mathbb{R}^d} \phi_\tau^{(d)}(y - z)\, d\eta(z).$$

Differentiating under the integral gives

$$\begin{aligned}\nabla p_\tau(y) &= -\frac{1}{\tau} \int_{\mathbb{R}^d} (y - z)\phi_\tau^{(d)}(y - z)\, d\eta(z) \\ &= -\frac{p_\tau(y)}{\sqrt{\tau}}\, \mathbb{E}[G \mid Y = y].\end{aligned}$$

Consequently, the score satisfies

$$\nabla \log p_\tau(y) = -\frac{1}{\sqrt{\tau}}\,\mathbb{E}[G \mid Y = y].$$

Therefore,

$$\begin{aligned}\mathcal{I}^{(d)}(\eta_\tau) &= \frac{1}{\tau}\mathbb{E}\left[\|\mathbb{E}[G \mid Y]\|^2\right] \\ &\le \frac{1}{\tau}\mathbb{E}\|G\|^2 = \frac{d}{\tau},\end{aligned}$$

where the inequality follows from the L^2 contraction property of conditional expectation. □

B.5 Dimensional Compression on Flat Supports

Corollary B.4 (Exact dimensional compression penalty on flat supports)**.** *Let μ_X and μ_Z be probability measures supported on affine m-dimensional subspaces of $\mathbb{R}^D$ and $\mathbb{R}^d$, respectively. Let*

$$\mathcal{I}^X_{\tan}(\tau) \quad \textit{and} \quad \mathcal{I}^Z_{\tan}(\tau)$$

denote the Fisher informations of their corresponding tangential coordinate laws after Gaussian smoothing at variance τ.

Then the ambient Fisher informations decompose exactly as

$$\mathcal{I}^{(D)}(\mu_{X,\tau}) = \mathcal{I}^X_{\tan}(\tau) + \frac{D-m}{\tau},$$

and

$$\mathcal{I}^{(d)}(\mu_{Z,\tau}) = \mathcal{I}^Z_{\tan}(\tau) + \frac{d-m}{\tau}.$$

Define the Intrinsic Noise Gain by

$$G^{(k)}(\eta_\tau) := k - 2\tau\mathcal{I}^{(k)}(\eta_\tau).$$

Then

$$G^{(d)}(\mu_{Z,\tau}) - G^{(D)}(\mu_{X,\tau}) = D - d + 2\tau\left(\mathcal{I}^X_{\tan}(\tau) - \mathcal{I}^Z_{\tan}(\tau)\right).$$

In particular, if the tangential Fisher information is preserved,

$$\mathcal{I}^X_{\tan}(\tau) = \mathcal{I}^Z_{\tan}(\tau),$$

then

$$G^{(d)}(\mu_{Z,\tau}) - G^{(D)}(\mu_{X,\tau}) = D - d.$$

Proof. Translations and orthogonal transformations preserve Fisher information, so it suffices to identify the affine supports with

$$\mathbb{R}^m \times \{\mathbf{0}\} \subset \mathbb{R}^D$$

and

$$\mathbb{R}^m \times \{\mathbf{0}\} \subset \mathbb{R}^d.$$

Let ρ_X and ρ_Z denote the corresponding tangential coordinate laws. Gaussian smoothing gives the exact factorizations

$$p_{X,\tau}(y,w) = (\rho_X * \phi^{(m)}_\tau)(y)\phi^{(D-m)}_\tau(w)$$

and

$$p_{Z,\tau}(y,v) = (\rho_Z * \phi^{(m)}_\tau)(y)\phi^{(d-m)}_\tau(v).$$

The score vectors split orthogonally into tangential and normal components. Since

$$\mathcal{I}\big(\mathcal{N}(\mathbf{0}, \tau I_\ell)\big) = \frac{\ell}{\tau},$$

we obtain

$$\mathcal{I}^{(D)}(\mu_{X,\tau}) = \mathcal{I}^X_{\tan}(\tau) + \frac{D-m}{\tau}$$

and

$$\mathcal{I}^{(d)}(\mu_{Z,\tau}) = \mathcal{I}^{Z}_{\tan}(\tau) + \frac{d-m}{\tau}.$$

Substitution into the definition of $G^{(k)}$ proves

$$G^{(d)}(\mu_{Z,\tau}) - G^{(D)}(\mu_{X,\tau}) = D - d + 2\tau\left(\mathcal{I}^{X}_{\tan}(\tau) - \mathcal{I}^{Z}_{\tan}(\tau)\right).$$

The final assertion follows when the tangential Fisher informations agree. □

Remark B.5 (Curved supports). The exact decomposition in Corollary B.4 relies on the existence of a single fixed normal subspace. For a curved manifold, the tangent and normal spaces vary with the base point, and Gaussian smoothing mixes contributions from neighboring tangent spaces. Consequently, the ambient smoothed density does not globally factor into independent tangential and normal components.

The tangent-plane model identifies

$$\frac{k-m}{\tau}$$

as the flat local baseline contributed by normal Gaussian directions. For a curved manifold, establishing a global small-noise expansion with this leading term and a controlled curvature remainder requires additional assumptions and analysis. We do not claim such a global expansion here.

Accordingly, on curved supports, the dimensional shift $D - d$ should be interpreted as the flat tangent-space compression baseline rather than as an exact global identity. The localized results of Theorem B.8 and Corollary B.11 quantify the corresponding curvature corrections for FIR.

B.6 Score Jacobian and Curvature Injection

This subsection records a formal local differentiation of the intrinsic score-transformation identity. We therefore impose stronger regularity than in Proposition B.1: assume here that E is C^3, that its tangential differential has full rank, and that the intrinsic densities are sufficiently smooth for the score fields below to be differentiable. These assumptions are used only in this explanatory subsection and not in Theorem B.8.

Let

$$\mathcal{J}_E(x) := \sqrt{\det\left(DE(x)^\top DE(x)\right)}$$

be the tangential volume Jacobian and define

$$g(x) := \nabla_{\mathcal{M}} \log \mathcal{J}_E(x).$$

The exact intrinsic score transformation from Proposition B.1 is

$$s_{\mathcal{M}}(x) = DE(x)^\top \widetilde{s}_{\mathcal{N}}(E(x)) + g(x).$$

Under these additional assumptions, differentiating the identity in flat local coordinates gives

$$\nabla s_{\mathcal{M}}(x) = \underbrace{DE(x)^\top \left(\nabla \widetilde{s}_{\mathcal{N}}\right)(E(x)) DE(x)}_{\text{inherited score curvature}} + \underbrace{\sum_{i=1}^{d} [\widetilde{s}_{\mathcal{N}}(E(x))]_i \nabla^2 E_i(x)}_{\text{encoder curvature injection}} + \underbrace{\nabla g(x)}_{\text{volume-distortion curvature}}.$$

On a general manifold, the corresponding statement is written using covariant derivatives.

Thus, encoder second derivatives contribute directly to the data-versus-latent score-Jacobian discrepancy. The volume-distortion term must also be controlled; a small pointwise metric distortion alone does not imply that ∇g is small.

B.7 Normal-Direction Decomposition for Subspace Measures

The following lemma establishes the "Dimensional Compression Penalty" by showing that normal noise directions contribute a universal baseline to the FIR. This term depends purely on the noise scale τ and the codimension $D - m$, but is constant across the manifold and independent of the data distribution. It is used to prove Theorem 3.3 in Appendix B.8.

Lemma B.6 (Exact normal-direction contribution)**.** *Let $1 \leq m \leq D$ and identify $\mathbb{R}^D \cong \mathbb{R}^m \times \mathbb{R}^{D-m}$. Let ρ be a Borel probability measure on $\mathbb{R}^m$, and define a measure μ on $\mathbb{R}^D$ by $\mu := \rho \otimes \delta_{\mathbf{0}}$. Recalling that ϕ_τ denotes the Gaussian kernel defined in* (7)*, we note that for every variance $\tau > 0$, the smoothed density μ_τ factorizes as:*

$$\mu_\tau(\mathbf{y}, \mathbf{w}) = (\rho * \phi_\tau^{(m)})(\mathbf{y})\, \phi_\tau^{(D-m)}(\mathbf{w}), \qquad (\mathbf{y}, \mathbf{w}) \in \mathbb{R}^m \times \mathbb{R}^{D-m},$$

and the Fisher Information Rate satisfies the exact splitting:

$$\mathcal{R}^{(D)}(\mu_\tau) = \mathcal{R}^{(m)}(\rho_\tau) + \frac{D-m}{\tau^2}.$$

Proof. Fix $\tau > 0$. By the definition of convolution and the product structure of the standard Gaussian kernel, we have $\mu_\tau(\mathbf{y}, \mathbf{w}) = \int_{\mathbb{R}^m} \phi_\tau^{(m)}(\mathbf{y} - \mathbf{y}')\, \phi_\tau^{(D-m)}(\mathbf{w})\, \rho(d\mathbf{y}')$. This gives the claimed factorization. Using $\log \mu_\tau(\mathbf{y}, \mathbf{w}) = \log(\rho * \phi_\tau^{(m)})(\mathbf{y}) + \log \phi_\tau^{(D-m)}(\mathbf{w})$, the Hessian $\nabla^2 \log \mu_\tau$ is block diagonal:

$$\nabla^2_{(\mathbf{y},\mathbf{w})} \log \mu_\tau = \operatorname{diag}\left(\nabla^2_{\mathbf{y}} \log(\rho * \phi_\tau^{(m)})(\mathbf{y}), \nabla^2_{\mathbf{w}} \log \phi_\tau^{(D-m)}(\mathbf{w}) \right).$$

Consequently, $\|\nabla^2 \log \mu_\tau\|_F^2 = \|\nabla^2_{\mathbf{y}} \log \rho_\tau\|_F^2 + \|\nabla^2_{\mathbf{w}} \log \phi_\tau^{(D-m)}\|_F^2$. Applying the identity from Theorem 2.1 and using Fubini's theorem to integrate over $\mathbb{R}^m \times \mathbb{R}^{D-m}$:

$$\mathcal{R}^{(D)}(\mu_\tau) = \mathcal{R}^{(m)}(\rho_\tau) + \int_{\mathbb{R}^{D-m}} \phi_\tau^{(D-m)}(\mathbf{w}) \|\nabla^2_{\mathbf{w}} \log \phi_\tau^{(D-m)}(\mathbf{w})\|_F^2 d\mathbf{w}.$$

Since $\nabla^2_{\mathbf{w}} \log \phi_\tau^{(D-m)}(\mathbf{w}) = -\frac{1}{\tau}\mathbf{I}_{D-m}$, the second term evaluates exactly to $\frac{D-m}{\tau^2}$, noting $\int \phi_\tau = 1$. □

B.8 Proof of Theorem 3.3 (Linear Stability)

By Lemma B.6, $\mathcal{R}^{(D)}(\mu_\tau) = \mathcal{R}^{(m)}(\rho_\tau) + \frac{D-m}{\tau^2}$. Similarly, μ_Z is supported on an m-plane in $\mathbb{R}^d$ with tangential coordinate law $\mathbf{A}_\# \rho$. Thus,

$$\mathcal{R}^{(d)}(\mu_{Z,\tau}) = \mathcal{R}^{(m)}((\mathbf{A}_\# \rho)_\tau) + \frac{d-m}{\tau^2}.$$

The difference is:

$$\mathcal{R}^{(D)}(\mu_\tau) - \mathcal{R}^{(d)}(\mu_{Z,\tau}) = \left(\mathcal{R}^{(m)}(\rho_\tau) - \mathcal{R}^{(m)}((\mathbf{A}_\# \rho)_\tau) \right) + \frac{D-d}{\tau^2}. \tag{23}$$

For convenience, set

$$R_p(\tau) := \mathcal{R}^{(m)}(\rho_\tau), \qquad R_q(\tau) := \mathcal{R}^{(m)}((\mathbf{A}_\# \rho)_\tau).$$

Let

$$\mathbf{A} = \mathbf{O}\mathbf{S}$$

be the polar decomposition of $\mathbf{A}$, where $\mathbf{O}$ is orthogonal and

$$\mathbf{S} = (\mathbf{A}^\top \mathbf{A})^{1/2}.$$

Since the Gaussian noise is isotropic, orthogonal transformations preserve FIR. Therefore,

$$R_q(\tau) = \mathcal{R}^{(m)}((\mathbf{S}_\# \rho)_\tau).$$

Let $\sigma_1, \ldots, \sigma_m$ be the eigenvalues of $\mathbf{S}$. The near-isometry condition implies

$$1 - \delta \leq \sigma_i^2 \leq 1 + \delta.$$

Consequently,

$$|\sigma_i - 1| = \frac{|\sigma_i^2 - 1|}{\sigma_i + 1} \le \delta,$$

and hence

$$\|\mathbf{S} - \mathbf{I}_m\|_2 \le \delta.$$

For every $\mathbf{x} \in \mathrm{supp}(\rho)$, the assumption $\tau \ge R_\rho^2$ gives

$$\|\mathbf{S}\mathbf{x}\| \le \sqrt{1+\delta}\,\|\mathbf{x}\| \le \sqrt{2}\,R_\rho \le \sqrt{2\tau},$$

and

$$\|\mathbf{x} - \mathbf{S}\mathbf{x}\| \le \delta R_\rho \le \delta\sqrt{\tau}.$$

We apply Lemma B.9, in dimension m, to the maps

$$G(\mathbf{x}) = \mathbf{x}, \qquad F(\mathbf{x}) = \mathbf{S}\mathbf{x},$$

with common prior ρ, center $\mathbf{c} = \mathbf{0}$, $L = \sqrt{2}$, and $a_0 = 1$. We obtain

$$\tau^2 |R_p(\tau) - R_q(\tau)| \le C_m^{(1)}\delta, \tag{24}$$

where $C_m^{(1)}$ depends only on m.

We also need a uniform upper bound for $R_p(\tau)$. Let

$$\bar{\rho} := (T_{1/\sqrt{\tau}})_{\#}\rho, \qquad T_{1/\sqrt{\tau}}(\mathbf{x}) := \frac{\mathbf{x}}{\sqrt{\tau}}.$$

Since $\tau \ge R_\rho^2$,

$$\mathrm{supp}(\bar{\rho}) \subseteq B_m(\mathbf{0}, 1).$$

Let $\bar{\mathbf{X}} \sim \bar{\rho}$ and

$$\bar{\mathbf{Y}} = \bar{\mathbf{X}} + \mathbf{G}, \qquad \mathbf{G} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m).$$

The second-order Tweedie identity gives

$$\nabla^2 \log(\bar{\rho} * \gamma_1)(\mathbf{y}) = -\mathbf{I}_m + \mathrm{Cov}\big(\bar{\mathbf{X}} \mid \bar{\mathbf{Y}} = \mathbf{y}\big).$$

Because $\|\bar{\mathbf{X}}\| \le 1$ almost surely,

$$\big\|\mathrm{Cov}\big(\bar{\mathbf{X}} \mid \bar{\mathbf{Y}} = \mathbf{y}\big)\big\|_F \le \mathrm{Tr}\,\mathrm{Cov}\big(\bar{\mathbf{X}} \mid \bar{\mathbf{Y}} = \mathbf{y}\big) \le 1.$$

It follows that

$$\tau^2 R_p(\tau) \le K_m, \qquad K_m := (\sqrt{m} + 1)^2. \tag{25}$$

We now consider the two cases in the theorem.

Case $m = d = D$. In this case,

$$\mathcal{D}_{\mathcal{R}}(\tau) = \frac{|R_p(\tau) - R_q(\tau)|}{R_q(\tau)}.$$

Lemma B.10, applied to $F(\mathbf{x}) = \mathbf{S}\mathbf{x}$ with $L = \sqrt{2}$, gives

$$\tau^2 R_q(\tau) \ge \frac{m}{(1+2)^2} = \frac{m}{9}.$$

Together with (24), this gives

$$\mathcal{D}_{\mathcal{R}}(\tau) \le \frac{9C_m^{(1)}}{m}\delta.$$

After enlarging the dimension-dependent constant,

$$\mathcal{D}_{\mathcal{R}}(\tau) \le C_m\delta.$$

Case $m < d \le D$**.** Set

$$\alpha := \frac{d-m}{D-m}.$$

By Lemma B.6,

$$\mathcal{R}^{(D)}(\mu_\tau) = R_p(\tau) + \frac{D-m}{\tau^2},$$

and

$$\mathcal{R}^{(d)}(\mu_{Z,\tau}) = R_q(\tau) + \frac{d-m}{\tau^2}.$$

Since $\alpha(D-m) = d-m$, the normal-direction terms cancel:

$$\begin{aligned}\mathcal{D}^{\mathrm{sc}}_{\mathcal{R}}(\tau) &= \left|1 - \alpha \frac{R_p(\tau) + (D-m)/\tau^2}{R_q(\tau) + (d-m)/\tau^2}\right| \\ &= \frac{|R_q(\tau) - \alpha R_p(\tau)|}{R_q(\tau) + (d-m)/\tau^2}. \end{aligned} \tag{26}$$

The numerator satisfies

$$|R_q - \alpha R_p| \le |R_q - R_p| + (1-\alpha)R_p,$$

where

$$1 - \alpha = \frac{D-d}{D-m}.$$

Since $d - m \ge 1$,

$$R_q(\tau) + \frac{d-m}{\tau^2} \ge \frac{1}{\tau^2}.$$

Using (24) and (25), we obtain

$$\mathcal{D}^{\mathrm{sc}}_{\mathcal{R}}(\tau) \le C_m^{(1)}\delta + K_m \frac{D-d}{D-m}.$$

After enlarging C_m and using $d - m \ge 1$, this implies

$$\mathcal{D}^{\mathrm{sc}}_{\mathcal{R}}(\tau) \le C_m \left(\frac{D-d}{D-m} + (d-m)\delta\right).$$

□

B.9 Diffusion-Scale Localization and Proof of Theorem B.8 (Localized Nonlinear Stability)

We first give a canonical construction of a localized prior at the diffusion scale. This construction makes precise the localization used in the main text, although the theorem below applies to any probability measure satisfying the same support condition.

Definition B.7 (Truncated local prior at diffusion scale)**.** Let $\mathcal{M} \subset \mathbb{R}^D$ be an m-dimensional embedded manifold, let μ be a probability measure supported on $\mathcal{M}$, and fix

$$x_0 \in \operatorname{supp}\mu, \qquad r > 0, \qquad \tau > 0.$$

Define the unnormalized localization kernel

$$K_\tau(x \mid x_0) := \exp\left(-\frac{\|x - x_0\|^2}{2\tau}\right)$$

and the truncated localized prior

$$d\lambda^{(r)}_{x_0,\tau}(x) = \frac{\mathbf{1}_{\{\|x-x_0\| \le r\sqrt{\tau}\}} K_\tau(x \mid x_0)\, d\mu(x)}{\displaystyle\int_{\{\|u-x_0\| \le r\sqrt{\tau}\}} K_\tau(u \mid x_0)\, d\mu(u)}.$$

The denominator is positive because $x_0 \in \operatorname{supp}\mu$.

When this prior is used to define a Gaussian-smoothed local law, the prior is held fixed while differentiating with respect to the smoothing variance. More precisely, for any measurable map

$$T : \mathcal{M} \to \mathbb{R}^k,$$

introduce an auxiliary smoothing variance s and define

$$P_s := \left(T_{\#}\lambda^{(r)}_{x_0,\tau}\right) * \gamma^k_s.$$

The FIR of the localized law is evaluated at $s = \tau$, with the derivative acting only on the Gaussian smoothing parameter s and not on the τ-dependence of $\lambda^{(r)}_{x_0,\tau}$.

The prior in Definition B.7 is one canonical localization at the diffusion scale. The theorem below is deliberately more general: it applies to any probability measure $\lambda^{(r)}_{x_0,\tau}$ supported on

$$\mathcal{M} \cap B(x_0, r\sqrt{\tau}),$$

whether or not it is obtained from the truncated-kernel construction. Comparing the truncated prior with an untruncated localization would require additional lower-mass or doubling assumptions on μ and is not needed below.

Theorem B.8 (Local nonlinear and curved stability of FIR)**.** *Let $\mathcal{M} \subset \mathbb{R}^D$ be a C^2, m-dimensional embedded manifold with positive reach*

$$\rho_{\mathcal{M}} := \mathrm{reach}(\mathcal{M}) > 0.$$

Fix $x_0 \in \mathcal{M}$, $r > 0$, and $\tau > 0$, and let $\lambda^{(r)}_{x_0,\tau}$ be a probability measure supported on $\mathcal{M} \cap B(x_0, r\sqrt{\tau})$. Let $E : \mathbb{R}^D \to \mathbb{R}^d$ be C^2 on a neighborhood containing all line segments joining x_0 to points in this set. Define $A_{x_0} := DE(x_0)\big|_{T_{x_0}\mathcal{M}}$, and the tangent-affine approximation $F_{x_0}(x) := E(x_0) + A_{x_0}P_{T_{x_0}\mathcal{M}}(x - x_0)$. Whenever $A^\top_{x_0}A_{x_0}$ is written, we represent A_{x_0} in an arbitrary orthonormal basis of $T_{x_0}\mathcal{M}$; the resulting operator norm is basis-independent. Define the local tangential distortion $\delta_{x_0} := \|A^\top_{x_0}A_{x_0} - I_m\|_{\mathrm{op}}$, and assume $\delta_{x_0} \le \delta_* < 1$. Let $P_{x_0,\tau} := (E_{\#}\lambda^{(r)}_{x_0,\tau}) * \gamma_\tau$, $Q_{x_0,\tau} := ((F_{x_0})_{\#}\lambda^{(r)}_{x_0,\tau}) * \gamma_\tau$. Define the local affine residual $\epsilon_{x_0,\tau} := \sup_{x\in\mathrm{supp}\,\lambda^{(r)}_{x_0,\tau}} \|E(x) - F_{x_0}(x)\|$. If $\epsilon_{x_0,\tau} \le a_0\sqrt{\tau}$, then there exists $C = C(d, r, \delta_*, a_0)$ such that $\tau^2\,|\mathcal{R}(P_{x_0,\tau}) - \mathcal{R}(Q_{x_0,\tau})| \le C\frac{\epsilon_{x_0,\tau}}{\sqrt{\tau}}$. Furthermore, $\left|1 - \frac{\mathcal{R}(P_{x_0,\tau})}{\mathcal{R}(Q_{x_0,\tau})}\right| \le \frac{C}{b_{x_0}}\frac{\epsilon_{x_0,\tau}}{\sqrt{\tau}}$, where one may take $b_{x_0} := \max\left\{d - m, \frac{d}{(1+r^2(1+\delta_*))^2}\right\}$. Moreover, suppose $\|DE(x_0)\|_{\mathrm{op}} \le J$ and $\sup_{y\in\mathcal{U}_{x_0,\tau}} \|D^2E(y)\|_{\mathrm{op}} \le H$, where $\|D^2E(y)\|_{\mathrm{op}} := \sup_{\|u\|=\|v\|=1} \|D^2E(y)[u, v]\|$, and $\mathcal{U}_{x_0,\tau}$ contains every segment $[x_0, x]$ for $x \in \mathrm{supp}\,\lambda^{(r)}_{x_0,\tau}$. Then $\epsilon_{x_0,\tau} \le \frac{r^2\tau}{2}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right)$. Consequently, whenever $\frac{r^2\sqrt{\tau}}{2}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right) \le a_0$, the preceding hypotheses are satisfied and

$$\left|1 - \frac{\mathcal{R}(P_{x_0,\tau})}{\mathcal{R}(Q_{x_0,\tau})}\right| \le \frac{Cr^2}{2b_{x_0}}\sqrt{\tau}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right).$$

To prove Theorem B.8, we first establish the following two lemmas:

Lemma B.9 (Gaussian-smoothed FIR stability under bounded perturbations)**.** *For a probability measure λ on a measurable space $\mathcal{X}$, and measurable maps $G, F : \mathcal{X} \to \mathbb{R}^d$, define*

$$P_\tau = (G_{\#}\lambda) * \gamma_\tau, \qquad Q_\tau = (F_{\#}\lambda) * \gamma_\tau.$$

Explicitly define the functional:

$$\mathcal{R}_\tau(\nu) := \int \left\|\nabla^2 \log(\nu * \gamma_\tau)(z)\right\|^2_F \, d(\nu * \gamma_\tau)(z).$$

Assume that for some $c \in \mathbb{R}^d$, $L \ge 0$, $a_0 > 0$, and $\epsilon \ge 0$,

$$\sup_{x\in\mathrm{supp}\,\lambda} \|F(x) - c\| \le L\sqrt{\tau},$$

and

$$\sup_{x\in\mathrm{supp}\,\lambda} \|G(x) - F(x)\| \le \epsilon, \qquad \frac{\epsilon}{\sqrt{\tau}} \le a_0.$$

Then there exists a constant $C(d, L, a_0)$ such that

$$\tau^2\,|\mathcal{R}_\tau(G_{\#}\lambda) - \mathcal{R}_\tau(F_{\#}\lambda)| \le C(d, L, a_0)\frac{\epsilon}{\sqrt{\tau}}.$$

Lemma B.10 (Universal FIR lower bound)**.** *Under the assumptions of Lemma B.9, we have:*

$$\tau^2 \mathcal{R}(Q_\tau) \geq \frac{d}{(1+L^2)^2}.$$

If $F(\operatorname{supp}\lambda)$ lies in an affine m-plane with $m < d$, then we also have:

$$\tau^2 \mathcal{R}(Q_\tau) \geq d - m.$$

Define $b(d, m, L) := \max\left\{d - m, \frac{d}{(1+L^2)^2}\right\}$, where the first term is interpreted as zero when $m = d$. Then:

$$\left|1 - \frac{\mathcal{R}(P_\tau)}{\mathcal{R}(Q_\tau)}\right| \leq \frac{C(d, L, a_0)}{b(d, m, L)} \frac{\epsilon}{\sqrt{\tau}}.$$

Proof of Lemma B.9. We prove the lemma in five steps.

Step 1: Remove τ by scaling.
Set

$$U = \frac{G(X) - c}{\sqrt{\tau}}, \qquad V = \frac{F(X) - c}{\sqrt{\tau}}, \qquad a = \frac{\epsilon}{\sqrt{\tau}}.$$

By the assumptions of the lemma,

$$\|V\| \leq L, \qquad \|U - V\| \leq a \leq a_0, \qquad \|U\| \leq M := L + a_0.$$

Let φ denote the standard d-dimensional Gaussian density. Define the scaled densities:

$$p(z) = \int \varphi(z - U(x))\, d\lambda(x), \qquad q(z) = \int \varphi(z - V(x))\, d\lambda(x).$$

Because $\mathcal{R}(P_\tau) = \tau^{-2}\mathcal{R}(p)$ and $\mathcal{R}(Q_\tau) = \tau^{-2}\mathcal{R}(q)$, it is sufficient to prove:

$$|\mathcal{R}(p) - \mathcal{R}(q)| \leq C(d, L, a_0)a. \tag{27}$$

Step 2: Joint Gaussian Channels and Total Variation.
Define the joint distributions on $\mathcal{X} \times \mathbb{R}^d$:

$$d\mathbf{P}(x, z) = d\lambda(x)\, \varphi(z - U(x))\, dz, \qquad d\mathbf{Q}(x, z) = d\lambda(x)\, \varphi(z - V(x))\, dz.$$

Because the conditional Gaussians have equal covariance, the KL divergence is:

$$D(\mathbf{P}\|\mathbf{Q}) = \frac{1}{2}\mathbb{E}_\lambda\|U - V\|^2 \leq \frac{a^2}{2}.$$

By the chain rule for relative entropy, $D(\mathbf{P}\|\mathbf{Q}) = D(p\|q) + \mathbb{E}_{Z\sim p} D(\pi_p^Z\|\pi_q^Z)$, which implies $D(p\|q) \leq \frac{a^2}{2}$ and $\mathbb{E}_{Z\sim p} D(\pi_p^Z\|\pi_q^Z) \leq \frac{a^2}{2}$. Pinsker's inequality subsequently gives:

$$\mathrm{TV}(p, q) \leq \frac{a}{2}, \qquad \mathbb{E}_{Z\sim p}\, \mathrm{TV}(\pi_p^Z, \pi_q^Z)^2 \leq \frac{a^2}{4}.$$

Step 3: Posterior Covariance Stability.
Let $C_p(z) = \mathrm{Cov}_{\pi_p^z}(U)$ and $C_q(z) = \mathrm{Cov}_{\pi_q^z}(V)$. We decompose the difference:

$$\|C_p(z) - C_q(z)\|_F \leq \left\|\mathrm{Cov}_{\pi_p^z}(U) - \mathrm{Cov}_{\pi_p^z}(V)\right\|_F + \left\|\mathrm{Cov}_{\pi_p^z}(V) - \mathrm{Cov}_{\pi_q^z}(V)\right\|_F.$$

The first term is bounded by $2a(M + L)$ since $\|U - V\| \leq a$. For the second term, bounded support ($\|V\| \leq L$) yields a bound of $6L^2\, \mathrm{TV}(\pi_p^z, \pi_q^z)$. Squaring and integrating under p gives the Covariance Stability Lemma:

$$\mathbb{E}_{Z\sim p}\|C_p(Z) - C_q(Z)\|_F^2 \leq C_0(L, a_0)a^2, \tag{28}$$

where $C_0(L, a_0) = 8(2L + a_0)^2 + 18L^4$.

Step 4: Hessian Bound via Tweedie's Identity.
The second-order Tweedie identities give $H_p(z) := \nabla^2 \log p(z) = -I_d + C_p(z)$ and $H_q(z) =$

$-I_d + C_q(z)$. Thus, (28) directly implies $\mathbb{E}_p\|H_p - H_q\|_F^2 \le C_0 a^2$. Furthermore, $\|C_p(z)\|_F \le \operatorname{Tr} C_p(z) \le M^2$, so pointwise:

$$\|H_p(z)\|_F \le K_p := \sqrt{d} + M^2, \qquad \|H_q(z)\|_F \le K_q := \sqrt{d} + L^2.$$

Step 5: Density Change and Hessian Difference.
We decompose the FIR difference:

$$\mathcal{R}(p) - \mathcal{R}(q) = \int \left(\|H_p(z)\|_F^2 - \|H_q(z)\|_F^2\right) p(z)\, dz + \int \|H_q(z)\|_F^2 \big(p(z) - q(z)\big)\, dz.$$

Using Cauchy-Schwarz and the bounds from Step 4, the first term is bounded by $\sqrt{C_0}(K_p + K_q)a$. The change-of-density term is bounded by $2K_q^2 \operatorname{TV}(p, q) \le K_q^2 a$. Combining these proves (27):

$$|\mathcal{R}(p) - \mathcal{R}(q)| \le \left[\sqrt{C_0}(K_p + K_q) + K_q^2\right] a.$$

Scaling back proves the absolute FIR bound stated in Lemma B.9.

□

Proof of Lemma B.10. Let $Z = F(X) + \sqrt{\tau}\, G$, where $\|F(X) - c\| \le L\sqrt{\tau}$. After scaling to unit noise, define $W = \frac{Z-c}{\sqrt{\tau}} = V + G$, where $\|V\| \le L$. Therefore, the covariance satisfies:

$$\operatorname{Cov}(W) = I + \operatorname{Cov}(V) \preceq (1 + L^2) I.$$

Let

$$J(q) := \mathbb{E}_q[s_q(W) s_q(W)^\top]$$

be the Fisher information matrix of W. Because q is a bounded Gaussian location mixture, q and its derivatives have Gaussian tails, so the following integrations by parts have no boundary terms.

Writing $m_q := \mathbb{E}_q[W]$, integration by parts gives

$$\mathbb{E}_q\left[(W - m_q) s_q(W)^\top\right] = -I_d.$$

The covariance matrix of the joint random vector

$$\big(W - m_q, s_q(W)\big)$$

is positive semidefinite. Taking its Schur complement yields the matrix information inequality

$$J(q) \succeq \operatorname{Cov}(W)^{-1} \succeq \frac{1}{1 + L^2} I_d.$$

A second integration by parts gives

$$\mathbb{E}_q[\nabla^2 \log q(W)] = -J(q).$$

Hence, Jensen's inequality gives

$$\begin{aligned}
\mathcal{R}(q) &= \mathbb{E}_q\|\nabla^2 \log q(W)\|_F^2 \\
&\ge \left\|\mathbb{E}_q[\nabla^2 \log q(W)]\right\|_F^2 \\
&= \|J(q)\|_F^2 \\
&\ge \frac{d}{(1 + L^2)^2}.
\end{aligned}$$

Scaling back to the variance τ yields:

$$\tau^2 \mathcal{R}(Q_\tau) \ge \frac{d}{(1 + L^2)^2}.$$

If $F(\operatorname{supp} \lambda)$ lies in an affine m-plane with $m < d$, the orthogonal noise components contribute exactly $d - m$, yielding the secondary bound:

$$\tau^2 \mathcal{R}(Q_\tau) \ge d - m.$$

Combining these lower bounds with the absolute difference established in Lemma B.9 yields the stated relative bound. □

Proof of Theorem B.8. Apply Lemma B.9 with

$$G(x) = E(x), \qquad F(x) = F_{x_0}(x), \qquad c = E(x_0),$$

and common prior $\lambda^{(r)}_{x_0,\tau}$.

Since

$$\|A_{x_0}\|^2_{\mathrm{op}} \le 1 + \delta_*,$$

for every $x \in \operatorname{supp} \lambda^{(r)}_{x_0,\tau}$ we have

$$\|F_{x_0}(x) - E(x_0)\| = \|A_{x_0} P_{T_{x_0}\mathcal{M}}(x - x_0)\| \le r\sqrt{1+\delta_*}\sqrt{\tau}.$$

Thus the bounded-image assumption in Lemma B.9 holds with

$$L = r\sqrt{1+\delta_*}.$$

The perturbation assumption is exactly

$$\sup_{\operatorname{supp} \lambda^{(r)}_{x_0,\tau}} \|E(x) - F_{x_0}(x)\| = \epsilon_{x_0,\tau}.$$

The absolute FIR estimate therefore follows from Lemma B.9.

Moreover, the image of F_{x_0} lies in the affine space

$$E(x_0) + \operatorname{range}(A_{x_0}),$$

whose dimension is m because $\delta_* < 1$ implies that A_{x_0} is injective on $T_{x_0}\mathcal{M}$. Combining the two lower bounds in Lemma B.10 gives

$$\tau^2 \mathcal{R}(Q_{x_0,\tau}) \ge b_{x_0},$$

where

$$b_{x_0} = \max\left\{ d - m,\ \frac{d}{\left(1 + r^2(1+\delta_*)\right)^2} \right\}.$$

Dividing the absolute bound by this lower bound proves the relative estimate.

It remains to derive the explicit residual estimate. For every $x \in \operatorname{supp} \lambda^{(r)}_{x_0,\tau}$,

$$E(x) - F_{x_0}(x) = E(x) - E(x_0) - DE(x_0)(x - x_0) + DE(x_0)(I - P_{T_{x_0}\mathcal{M}})(x - x_0).$$

The second-order Taylor theorem gives

$$\|E(x) - E(x_0) - DE(x_0)(x - x_0)\| \le \frac{H}{2}\|x - x_0\|^2.$$

Positive reach gives

$$\|(I - P_{T_{x_0}\mathcal{M}})(x - x_0)\| \le \frac{\|x - x_0\|^2}{2\rho_{\mathcal{M}}}.$$

Consequently,

$$\begin{aligned}\|E(x) - F_{x_0}(x)\| &\le \frac{1}{2}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right)\|x - x_0\|^2 \\ &\le \frac{r^2\tau}{2}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right).\end{aligned}$$

Taking the supremum proves the claimed residual bound and completes the proof. □

B.10 Proof of Corollary B.11

Corollary B.11 (Combined local data-versus-latent bound)**.** *Under the setup of Theorem B.8, suppose* $\delta_{x_0} = \|A_{x_0}^\top A_{x_0} - I_m\|_{\mathrm{op}} \le \delta_* < 1$*, and assume the local derivative bounds* $\|DE(x_0)\|_{\mathrm{op}} \le J$, $\sup_{y \in \mathcal{U}_{x_0,\tau}} \|D^2E(y)\|_{\mathrm{op}} \le H$*. Fix constants* $a_X, a_Z > 0$ *and suppose* $\frac{r^2\sqrt{\tau}}{2\rho_{\mathcal{M}}} \le a_X$, $\frac{r^2\sqrt{\tau}}{2}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right) \le a_Z$*. Define the maps*

$$G_X(x) := x - x_0, \qquad F_X(x) := P_{T_{x_0}\mathcal{M}}(x - x_0),$$
$$F_Z(x) := A_{x_0} P_{T_{x_0}\mathcal{M}}(x - x_0), \qquad G_Z(x) := E(x) - E(x_0).$$

Define the corresponding local Gaussian-smoothed laws by

$$P^X_{x_0,\tau} := (G_{X\#}\lambda^{(r)}_{x_0,\tau}) * \gamma^D_\tau, \qquad Q^X_{x_0,\tau} := (F_{X\#}\lambda^{(r)}_{x_0,\tau}) * \gamma^D_\tau,$$
$$Q^Z_{x_0,\tau} := (F_{Z\#}\lambda^{(r)}_{x_0,\tau}) * \gamma^d_\tau, \qquad P^Z_{x_0,\tau} := (G_{Z\#}\lambda^{(r)}_{x_0,\tau}) * \gamma^d_\tau.$$

Then there exist constants C_A*,* C_E*,* $C_M > 0$*, depending only on* D*,* d*,* m*,* r*,* δ_**,* a_X*,* a_Z*, such that*

$$\left|\tau^2\left[\mathcal{R}(P^X_{x_0,\tau}) - \mathcal{R}(P^Z_{x_0,\tau})\right] - (D-d)\right| \le C_A\delta_{x_0} + C_E r^2\sqrt{\tau}\, H + C_M r^2\sqrt{\tau}\frac{1+J}{\rho_{\mathcal{M}}}.$$

Proof. For brevity, write

$$\lambda := \lambda^{(r)}_{x_0,\tau}$$

and let

$$P_T := P_{T_{x_0}\mathcal{M}}.$$

We decompose

$$\begin{aligned}
&\tau^2\left[\mathcal{R}(P^X_{x_0,\tau}) - \mathcal{R}(P^Z_{x_0,\tau})\right] - (D-d) \\
&= \tau^2\left[\mathcal{R}(P^X_{x_0,\tau}) - \mathcal{R}(Q^X_{x_0,\tau})\right] \\
&\quad + \left\{\tau^2\left[\mathcal{R}(Q^X_{x_0,\tau}) - \mathcal{R}(Q^Z_{x_0,\tau})\right] - (D-d)\right\} \\
&\quad + \tau^2\left[\mathcal{R}(Q^Z_{x_0,\tau}) - \mathcal{R}(P^Z_{x_0,\tau})\right].
\end{aligned} \tag{29}$$

Step 1: Manifold versus tangent plane. Apply Lemma B.9 in dimension D to

$$G_X(x) = x - x_0, \qquad F_X(x) = P_T(x - x_0).$$

On the support of λ,

$$\|F_X(x)\| \le r\sqrt{\tau}.$$

The positive-reach estimate gives

$$\begin{aligned}
\|G_X(x) - F_X(x)\| &= \|(I - P_T)(x - x_0)\| \\
&\le \frac{\|x - x_0\|^2}{2\rho_{\mathcal{M}}} \\
&\le \frac{r^2\tau}{2\rho_{\mathcal{M}}}.
\end{aligned}$$

Since

$$\frac{r^2\sqrt{\tau}}{2\rho_{\mathcal{M}}} \le a_X,$$

Lemma B.9 yields

$$\tau^2\left|\mathcal{R}(P^X_{x_0,\tau}) - \mathcal{R}(Q^X_{x_0,\tau})\right| \le C_X\frac{r^2\sqrt{\tau}}{2\rho_{\mathcal{M}}}, \tag{30}$$

where C_X depends only on D, r, a_X.

Step 2: Tangential linear distortion. Choose an isometry

$$U_0 : \mathbb{R}^m \longrightarrow T_{x_0}\mathcal{M}$$

and define tangent coordinates

$$Y(x) := U_0^\top P_T(x - x_0).$$

Then

$$\|Y(x)\| \le r\sqrt{\tau}.$$

Let

$$\overline{A} := A_{x_0} U_0.$$

Its polar decomposition has the form

$$\overline{A} = VS,$$

where $V : \mathbb{R}^m \to \mathbb{R}^d$ is an isometry and

$$S = (\overline{A}^\top \overline{A})^{1/2}.$$

Since

$$\|\overline{A}^\top \overline{A} - I_m\|_{\mathrm{op}} = \delta_{x_0} \le \delta_*,$$

the eigenvalues of S belong to

$$[\sqrt{1-\delta_*}, \sqrt{1+\delta_*}].$$

Therefore,

$$\|S - I_m\|_{\mathrm{op}} \le c_* \delta_{x_0}, \qquad c_* := \frac{1}{1+\sqrt{1-\delta_*}}.$$

Let

$$\eta := Y_\# \lambda.$$

By Lemma B.6,

$$\mathcal{R}(Q^X_{x_0,\tau}) = \mathcal{R}^{(m)}(\eta_\tau) + \frac{D-m}{\tau^2},$$

whereas orthogonal invariance and the same normal decomposition give

$$\mathcal{R}(Q^Z_{x_0,\tau}) = \mathcal{R}^{(m)}((S_\#\eta)_\tau) + \frac{d-m}{\tau^2}.$$

Hence

$$\begin{aligned}\tau^2 \left[\mathcal{R}(Q^X_{x_0,\tau}) - \mathcal{R}(Q^Z_{x_0,\tau})\right] - (D-d) \\ = \tau^2 \left[\mathcal{R}^{(m)}(\eta_\tau) - \mathcal{R}^{(m)}((S_\#\eta)_\tau)\right].\end{aligned}$$

For $y \in \operatorname{supp}\eta$,

$$\|Sy\| \le r\sqrt{1+\delta_*}\sqrt{\tau},$$

and

$$\|Sy - y\| \le c_* r \delta_{x_0} \sqrt{\tau}.$$

Applying Lemma B.9 in dimension m to the maps $y \mapsto y$ and $y \mapsto Sy$ gives

$$\left|\tau^2 \left[\mathcal{R}(Q^X_{x_0,\tau}) - \mathcal{R}(Q^Z_{x_0,\tau})\right] - (D-d)\right| \le C_A \delta_{x_0}, \tag{31}$$

where C_A depends only on m, r, δ_*.

Step 3: Affine versus nonlinear encoding. Apply Lemma B.9 in dimension d to

$$G_Z(x) = E(x) - E(x_0), \qquad F_Z(x) = A_{x_0} P_T(x - x_0).$$

We have

$$\|F_Z(x)\| \le r\sqrt{1+\delta_*}\sqrt{\tau}.$$

The residual estimate in Theorem B.8 gives

$$\|G_Z(x) - F_Z(x)\| \le \frac{r^2\tau}{2}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right).$$

Since the second smallness assumption in the corollary holds, Lemma B.9 gives

$$\tau^2 \left|\mathcal{R}(P^Z_{x_0,\tau}) - \mathcal{R}(Q^Z_{x_0,\tau})\right| \le C_Z \frac{r^2\sqrt{\tau}}{2}\left(H + \frac{J}{\rho_{\mathcal{M}}}\right), \tag{32}$$

where C_Z depends only on d, r, δ_*, a_Z.

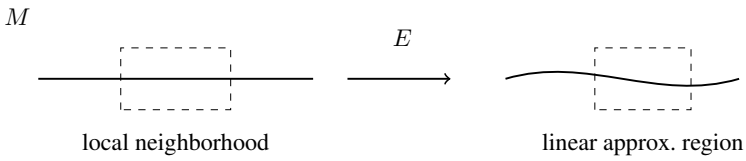
Finally, apply the triangle inequality to (29) and use (30), (31), and (32). Absorbing numerical constants into C_A, C_E, C_M yields

$$\begin{aligned}&\left|\tau^2\left[\mathcal{R}(P^X_{x_0,\tau}) - \mathcal{R}(P^Z_{x_0,\tau})\right] - (D-d)\right| \\ &\qquad \le C_A\delta_{x_0} + C_E r^2\sqrt{\tau}\,H + C_M r^2\sqrt{\tau}\frac{1+J}{\rho_{\mathcal{M}}}.\end{aligned}$$

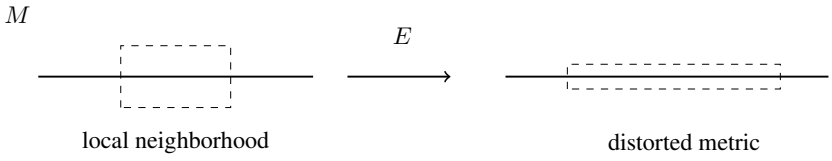
□

B.11 Visual Interpretation of Theorem B.8 and Corollary B.11

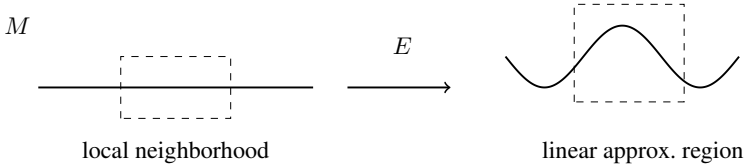
Note that Figure 3 demonstrates the encoder-induced distortion implied by Theorem B.8 and Corollary B.11 discussed in the main text.



(a) Ideal encoder: local isometry and low curvature



(b) Tangential Distortion: local stretching/compression with large δ



(c) Curvature Injection: high-frequency nonlinearity with large ε and small τ.

Figure 3: Geometric interpretation of encoder-induced distortions.

C ESTIMATION OF FI AND FIR

In this section, we describe how the Fisher Information (FI) and Fisher Information Rate (FIR) are computed in our experiments based on trained diffusion models. We specifically focus on the computational strategies that allow these metrics to scale to high-dimensional neural architectures.

Fisher Information (FI). The FI can be expressed using the score function $\mathbf{s}_\tau(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\tau(\mathbf{x})$. For a given noise variance $\tau > 0$ and dimension D, the FI is calculated as:

$$\begin{aligned} \mathcal{I}^{(D)}(\mu_\tau) &= \int_{\mathbb{R}^D} p_\tau(\mathbf{x}) \left\| \nabla_{\mathbf{x}} \log p_\tau(\mathbf{x}) \right\|^2 d\mathbf{x} \\ &= \mathbb{E}_{\mathbf{x}\sim p_\tau}\left[\|\mathbf{s}_\tau(\mathbf{x})\|^2 \right] \\ &\approx \frac{1}{N} \sum_{n=1}^{N} \left\| \mathbf{s}_\tau(\mathbf{x}^{(n)}) \right\|^2, \end{aligned} \tag{33}$$

where $\mathbf{x}^{(n)}$ are samples drawn from the smoothed distribution p_τ. In practice, these are obtained by adding Gaussian noise $\mathcal{N}(\mathbf{0}, \tau\mathbf{I})$ to clean data samples from the empirical distribution.

Fisher Information Rate (FIR). The FIR is computed as the dissipation rate of the FI. Using the Hessian representation from Proposition 2.1, we have:

$$\begin{aligned} \mathcal{R}^{(D)}(\mu_\tau) &= -\frac{d}{d\tau} \mathcal{I}^{(D)}(\mu_\tau) \\ &= \mathbb{E}_{\mathbf{x}\sim p_\tau}\left[\mathrm{Tr}\Big(\big(\nabla_{\mathbf{x}} \mathbf{s}_\tau(\mathbf{x})\big)^2 \Big) \right] \\ &\approx \frac{1}{N} \sum_{n=1}^{N} \mathrm{Tr}\Big(\big(\nabla_{\mathbf{x}} \mathbf{s}_\tau(\mathbf{x}^{(n)})\big)^2 \Big). \end{aligned}$$

For high-dimensional neural networks, explicitly forming the $D \times D$ score Jacobian requires $\mathcal{O}(D^2)$ storage and many derivative evaluations. We avoid forming the full Jacobian by using the **Hutchinson trace estimator** (Hutchinson, 1989). For the exact population score,

$$\nabla_{\mathbf{x}} \mathbf{s}_\tau = \nabla_{\mathbf{x}}^2 \log p_\tau$$

is a symmetric Hessian. Hence,

$$\mathrm{Tr}\left((\nabla_{\mathbf{x}} \mathbf{s}_\tau)^2 \right) = \|\nabla_{\mathbf{x}} \mathbf{s}_\tau\|_F^2,$$

and the squared Frobenius norm can be written as an expectation over isotropic probe vectors $\mathbf{v}$:

$$\begin{aligned} \mathrm{Tr}\big((\nabla_{\mathbf{x}} \mathbf{s}_\tau(\mathbf{x}))^2 \big) &= \mathbb{E}_{\mathbf{v}}\left[\left\| \nabla_{\mathbf{x}} \mathbf{s}_\tau(\mathbf{x})\, \mathbf{v} \right\|^2 \right] \\ &\approx \frac{1}{M} \sum_{m=1}^{M} \left\| \nabla_{\mathbf{x}} \mathbf{s}_\tau(\mathbf{x})\, \mathbf{v}^{(m)} \right\|^2, \end{aligned}$$

where the probe vectors $\mathbf{v}^{(m)}$ are drawn independently from a Rademacher distribution (entries in $\{-1, 1\}$). This leads to the empirical estimator:

$$\mathcal{R}^{(D)}(\mu_\tau) \approx \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \left\| \nabla_{\mathbf{x}} \mathbf{s}_\tau(\mathbf{x}^{(n)})\, \mathbf{v}^{(m)} \right\|^2. \tag{34}$$

A learned score network need not have an exactly symmetric Jacobian. Accordingly, the empirical estimator above should be viewed as a plug-in estimator of the squared Frobenius norm of the learned score Jacobian. It agrees with the population FIR functional when the learned vector field equals the conservative population score.

Implementation via Directional Derivatives. The term $\nabla_{\mathbf{x}} \mathbf{s}_\tau(\mathbf{x})\mathbf{v}$ is a **Jacobian-vector product (JVP)**, representing the directional derivative of the score function. In modern automatic-differentiation frameworks (e.g., `torch.autograd.functional.jvp` in PyTorch), this product can be evaluated at a cost comparable to a small number of network evaluations, without explicitly computing or storing the full Jacobian matrix.

Alternatively, to avoid second-order automatic differentiation entirely or further reduce memory overhead, the term can be approximated via a first-order finite difference. For a small step size $\varepsilon > 0$:

$$\nabla_{\mathbf{x}} \mathbf{s}_\tau(\mathbf{x})\, \mathbf{v} \approx \frac{\mathbf{s}_\tau(\mathbf{x} + \varepsilon \mathbf{v}) - \mathbf{s}_\tau(\mathbf{x})}{\varepsilon}.$$

This yields the gradient-free estimator:

$$\mathcal{R}^{(D)}(\mu_\tau) \approx \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \left\| \frac{\mathbf{s}_\tau(\mathbf{x}^{(n)} + \varepsilon \mathbf{v}^{(m)}) - \mathbf{s}_\tau(\mathbf{x}^{(n)})}{\varepsilon} \right\|^2. \tag{35}$$

In our experiments, the JVP approach was found to be highly accurate and numerically stable, providing a scalable means to monitor the geometric complexity penalty during the diffusion process.

Score Function Parameterization in Experiments. In practice, the score function is obtained directly from the trained diffusion models, but the exact extraction depends on the model's training objective.

For the toy experiments described in §4.1 and §4.2, we use a noise-prediction diffusion model. The empirical score is approximated using the model's predicted noise $\boldsymbol{\epsilon}_\theta(\mathbf{x}, \tau)$:

$$\mathbf{s}_\tau(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\tau(\mathbf{x}) \approx -\frac{1}{\sqrt{\tau}} \boldsymbol{\epsilon}_\theta(\mathbf{x}, \tau).$$

For the experiments described in §4.3, JiT, SiT, RAE, RAEv2, and REPA-E employ flow-based generative models with a linear interpolation path

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where t denotes the flow time and σ_t is the corresponding noise standard-deviation coefficient. For models that predict the clean sample $\hat{\mathbf{x}}_{0,\theta}(\mathbf{x}_t, t)$, we compute the score as

$$\mathbf{s}_t(\mathbf{x}_t) = -\frac{\mathbf{x}_t - \alpha_t \hat{\mathbf{x}}_{0,\theta}(\mathbf{x}_t, t)}{\sigma_t^2}.$$

For models that instead predict the velocity field $\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t)$, we use

$$\mathbf{s}_t(\mathbf{x}_t) = \frac{\alpha_t \hat{\mathbf{v}}_\theta(\mathbf{x}_t, t) - \dot{\alpha}_t \mathbf{x}_t}{\sigma_t \left(\dot{\alpha}_t \sigma_t - \alpha_t \dot{\sigma}_t \right)},$$

where $\dot{\alpha}_t = d\alpha_t/dt$ and $\dot{\sigma}_t = d\sigma_t/dt$.

To express the flow models in the same variance parameterization used for the EDM experiments, we define

$$\widetilde{\mathbf{x}}_\tau = \frac{\mathbf{x}_t}{\alpha_t}, \qquad \tau = \left(\frac{\sigma_t}{\alpha_t} \right)^2.$$

Then $\widetilde{\mathbf{x}}_\tau = \mathbf{x}_0 + \sqrt{\tau}\boldsymbol{\epsilon}$, and the corresponding score is

$$\widetilde{\mathbf{s}}_\tau\big(\widetilde{\mathbf{x}}_\tau\big) = \alpha_t \mathbf{s}_t(\mathbf{x}_t).$$

For the experiments described in §F.1 and §F.2, we adopt the Elucidating Diffusion Models (EDM) framework (Karras et al., 2022). Following EDM standard practices, we employ a denoiser parameterization $D_\theta(\mathbf{x}, \tau)$ alongside their proposed preconditioning and noise scaling schedules. Unless otherwise specified, all core hyperparameters (e.g., network capacity, learning rate, and weighting schedules) are strictly borrowed from the default EDM configurations. Under this framework, the score is approximated directly via the denoiser:

$$\mathbf{s}_\tau(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\tau(\mathbf{x}) \approx -\frac{\mathbf{x} - D_\theta(\mathbf{x}, \tau)}{\tau}.$$

In all experiments, FI and FIR are computed using Equation (33) and Equation (34), respectively, with the required derivatives evaluated using automatic differentiation. For the large-scale models considered in §F.1, §F.2, and §4.3, we use $N = 200$ samples and $M = 20$ Hutchinson probe vectors due to their computational and memory costs. For all other experiments, we use $N = 1000$ samples and $M = 50$ Hutchinson probe vectors.

D Experimental Details

In this section, we provide the model architectures, training configurations, and sampling parameters used in our experiments. Each subsection corresponds to the experimental setting described in the main text.

D.1 FI and FIR Comparison in the Gaussian Toy Setting

For the experiment in §4.1, we first describe the architecture of the tiny diffusion model used in this setting. The denoising network in the tiny diffusion model is implemented as an MLP (multilayer perceptron). We use a hidden dimension of 128 with 3 hidden layers, and apply sinusoidal embeddings of dimension 128 to both the input data and the diffusion time.

For all experiments, we use $50{,}000$ training samples with batch size 256 and train for 200 epochs. The diffusion process uses 100 timesteps with a linear noise β-schedule. We use the AdamW optimizer with a learning rate of 5×10^{-4}.

D.2 FIR Deviation in the Gaussian Toy Setting

This subsection provides the experimental details for the study in §4.2. The model and training configuration follows the Gaussian toy experiment in §4.1.

The only modification concerns the input embedding used in the tiny diffusion model. In §4.1, sinusoidal embeddings were applied to the input coordinates. Since the experiments in this section consider data whose input dimension may exceed two, we instead use the identity embedding for all tiny diffusion models.

D.3 FI and FIR Comparison across ImageNet Latent Diffusion Models

This subsection provides the experimental details for the study described in §4.3 on ImageNet data.

We evaluate RAE (Representation Autoencoder) (Zheng et al., 2025), RAEv2 (Improved Baselines with Representation Autoencoders) (Singh et al., 2026), SiT (Scalable Interpolant Transformer) (Ma et al., 2024), REPA-E (Representation Alignment for End-to-End VAE Tuning) (Leng et al., 2025), and JiT (Just image Transformer) (Li & He, 2026) using their official GitHub implementations and publicly available pretrained checkpoints. No additional training or fine-tuning is performed. We choose these models because their latent-space or pixel-space diffusion models use closely related transformer-based architectures, and we select checkpoints with similar model scale trained on ImageNet at 256×256 resolution.

Specifically, for JiT, we choose the Huge-scale checkpoint with patch size 16 (JiT/16)[2]. For RAE, we choose the checkpoint with a DINOv2-B representation encoder without alignment guidance, together with the corresponding DiT-based XL-scale latent diffusion model for sample generation[3]. For RAEv2, we choose the checkpoint with a DINOv3-L encoder, where the multi-layer representation is obtained by summing the last 7 layers[4]. For SiT, we choose the XL-scale checkpoint with

[2]We use the official JiT implementation: `https://github.com/LTH14/JiT`.

[3]We use the official RAE implementation: `https://github.com/bytetriper/RAE/tree/main`.

[4]We use the official RAEv2 implementation: `https://github.com/nanovisionx/RAEv2/tree/main`.

patch size 2^5. For REPA-E, we choose the checkpoint with the SD-VAE tokenizer and a SiT-based XL-scale generator with patch size 2^6.

Among these methods, JiT performs generation directly in pixel space, while the other methods first encode images into a latent representation and perform generation using a latent diffusion model. For each method, we generate samples using the sampling parameters recommended by its official implementation. For evaluation, we use $50{,}000$ ImageNet validation samples as the reference set and generate $50{,}000$ samples from each model to compute the metrics.

For the experiments in §4.3, we compute the geometric diagnostics introduced by Zhong et al. (2026) using their official implementation,[7]

D.4 Computing Resources

All experiments were run on NVIDIA A100-SXM4-80GB GPUs. Each GPU has 80GB of memory.

E Gaussian Toy Experiment: Derivations and Additional Results

E.1 Derivation Relevant for the Experiment in §4.1

In this section, we provide derivations for the experiment in §4.1. We consider samples $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ and construct encoded representations $\mathbf{z} = E(\mathbf{x})$, where the encoder E is applied pointwise. For E, we consider functions given by common activation functions, including ReLU and Leaky ReLU (with negative slope α).

We distinguish the deterministic pushforward laws from the Gaussian-smoothed laws used in the diffusion experiments. Proposition B.1 concerns intrinsic FI under a deterministic pushforward when the logarithmic volume Jacobian has the required weak regularity. By contrast, Lemma B.3 applies to every Gaussian-smoothed latent law. Below, we derive the exact smoothed density and a model-free FI formula for ReLU and Leaky ReLU under the Gaussian toy distribution.

E.1.1 Exact Gaussian-smoothed FI for ReLU and Leaky ReLU

Let

$$\psi_\alpha(x) := \begin{cases} x, & x > 0, \\ \alpha x, & x \le 0, \end{cases} \qquad \alpha \ge 0.$$

Thus, $\alpha = 0$ corresponds to ReLU, while $\alpha > 0$ corresponds to Leaky ReLU.

For $v > 0$, write

$$\varphi_v(y) := \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{y^2}{2v}\right),$$

let $\varphi := \varphi_1$, and let Φ denote the standard Gaussian CDF.

Proposition E.1 (Exact smoothed activation density)**.** *Let $X, G \sim \mathcal{N}(0,1)$ be independent and define*

$$Y_{\alpha,\tau} := \psi_\alpha(X) + \sqrt{\tau}\, G, \qquad \tau > 0.$$

Then $Y_{\alpha,\tau}$ has density

$$\begin{aligned} p_{\alpha,\tau}(y) = \varphi_{1+\tau}(y)\Phi\left(\frac{y}{\sqrt{\tau(1+\tau)}}\right) \\ + \varphi_{\alpha^2+\tau}(y)\Phi\left(-\frac{\alpha y}{\sqrt{\tau(\alpha^2+\tau)}}\right). \end{aligned} \tag{36}$$

[5] We use the official SiT implementation: `https://github.com/willisma/sit`.

[6] We use the official REPA-E implementation: `https://github.com/End2End-Diffusion/REPA-E`.

[7] `https://github.com/KlingAIResearch/diffusing-right-space`

For $\alpha = 0$, the second term is interpreted as

$$\frac{1}{2}\varphi_\tau(y).$$

Define

$$a_\tau := \frac{1}{\sqrt{\tau(1+\tau)}}, \qquad b_{\alpha,\tau} := \frac{\alpha}{\sqrt{\tau(\alpha^2+\tau)}}.$$

The derivative of the density is

$$p'_{\alpha,\tau}(y) = \varphi_{1+\tau}(y)\left[-\frac{y}{1+\tau}\Phi(a_\tau y) + a_\tau\varphi(a_\tau y)\right] + \varphi_{\alpha^2+\tau}(y)\left[-\frac{y}{\alpha^2+\tau}\Phi(-b_{\alpha,\tau}y) - b_{\alpha,\tau}\varphi(b_{\alpha,\tau}y)\right]. \tag{37}$$

For the coordinatewise encoder

$$E_\alpha(x_1, x_2) = \big(\psi_\alpha(x_1), \psi_\alpha(x_2)\big)$$

applied to

$$X \sim \mathcal{N}(\mathbf{0}, I_2),$$

the two smoothed coordinates remain independent. Therefore, the exact population FI is

$$\mathcal{I}_\alpha^{(2)}(\tau) = 2\int_{\mathbb{R}} \frac{|p'_{\alpha,\tau}(y)|^2}{p_{\alpha,\tau}(y)}\,dy \le \frac{2}{\tau}. \tag{38}$$

When $\alpha = 1$, the encoder is the identity,

$$p_{1,\tau}(y) = \varphi_{1+\tau}(y),$$

and hence

$$\mathcal{I}_1^{(2)}(\tau) = \frac{2}{1+\tau}.$$

Proof. Split the defining integral according to the events $X > 0$ and $X \le 0$. For the positive branch,

$$\begin{aligned}\int_0^\infty &\varphi(x)\varphi_\tau(y-x)\,dx \\ &= \varphi_{1+\tau}(y)\mathbb{P}\left(X > 0 \,\middle|\, X + \sqrt{\tau}G = y\right) \\ &= \varphi_{1+\tau}(y)\Phi\left(\frac{y}{\sqrt{\tau(1+\tau)}}\right).\end{aligned}$$

Similarly, for the negative branch,

$$\begin{aligned}\int_{-\infty}^0 &\varphi(x)\varphi_\tau(y-\alpha x)\,dx \\ &= \varphi_{\alpha^2+\tau}(y)\Phi\left(-\frac{\alpha y}{\sqrt{\tau(\alpha^2+\tau)}}\right).\end{aligned}$$

This proves (36). Differentiating the displayed expression proves (37).

The product formula for the two-dimensional density follows from the independence of the coordinates. Its score separates coordinatewise, giving (38). The upper bound is also an immediate consequence of Lemma B.3. □

E.2 Derivation Relevant for the Experiment in §4.2

In this section, we provide derivations for the experiment in §4.2. We consider data $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ on a 2-dimensional intrinsic manifold, embedded in $\mathbb{R}^D$ as $\mathbf{x} = (y_1, y_2, 0, \ldots, 0)$, with latent representations $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^d$.

Exact FIR formulas for the linear Gaussian experiments. The compact-support assumption of Theorem 3.3 does not apply directly to the Gaussian law used in these toy experiments. Fortunately, in this setting the data-space and latent-space FIRs can be computed exactly.

More generally, let

$$\rho = \mathcal{N}(\mathbf{0}, \mathbf{I}_m),$$

and let $\lambda_1, \dots, \lambda_m$ denote the eigenvalues of $\mathbf{A}^\top \mathbf{A}$. Since the nonzero eigenvalues of $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top \mathbf{A}$ agree, Gaussian smoothing gives

$$\mathcal{R}^{(D)}(\mu_\tau) = \frac{m}{(1+\tau)^2} + \frac{D-m}{\tau^2}, \tag{39}$$

while

$$\mathcal{R}^{(d)}(\mu_{Z,\tau}) = \sum_{i=1}^{m} \frac{1}{(\lambda_i + \tau)^2} + \frac{d-m}{\tau^2}. \tag{40}$$

For the equal-dimensional experiment $D = d = m = 2$ with

$$\mathbf{A} = \operatorname{diag}\left(\sqrt{1+\delta_0}, \sqrt{1-\delta_0}\right),$$

the eigenvalues are

$$\lambda_1 = 1 + \delta_0, \qquad \lambda_2 = 1 - \delta_0.$$

Writing $a_\tau := 1 + \tau$, we obtain

$$\begin{aligned} \mathcal{D}_{\mathcal{R}}(\tau) &= 1 - \frac{2/a_\tau^2}{(a_\tau + \delta_0)^{-2} + (a_\tau - \delta_0)^{-2}} \\ &= \frac{\delta_0^2\left(3a_\tau^2 - \delta_0^2\right)}{a_\tau^2\left(a_\tau^2 + \delta_0^2\right)} \le \frac{3\delta_0^2}{(1+\tau)^2} \le 3\delta_0^2. \end{aligned} \tag{41}$$

Thus, the population FIR deviation is quadratic in δ_0 near $\delta_0 = 0$, and in particular it also satisfies an $\mathcal{O}(\delta_0)$ upper bound for $\delta_0 \in [0, 1]$.

For the dimension-varying experiment, $\mathbf{A} = \mathbf{I}_m$, and hence

$$\mathcal{D}_{\mathcal{R}}^{\mathrm{sc}}(\tau) = \frac{D-d}{D-m} \frac{m/(1+\tau)^2}{m/(1+\tau)^2 + (d-m)/\tau^2} \le \frac{D-d}{D-m}. \tag{42}$$

Linear Encoder

We now specialize the exact Gaussian formulas above to the linear encoder used in the experiment and verify its near-isometry parameter. Take $D = d = m = 2$ and consider the encoder

$$\mathbf{z} = \mathbf{A}\mathbf{x}, \qquad \mathbf{A} = \operatorname{diag}\big(\sqrt{1+\delta_0},\ \sqrt{1-\delta_0}\big).$$

Then

$$\mathbf{A}^\top \mathbf{A} - \mathbf{I}_2 = \operatorname{diag}(1+\delta_0,\ 1-\delta_0) - \mathbf{I}_2 = \operatorname{diag}(\delta_0,\ -\delta_0).$$

Therefore,

$$\|\mathbf{A}^\top \mathbf{A} - \mathbf{I}_2\|_2 = \delta_0.$$

Thus, we can take

$$\delta = \delta_0.$$

Dimension-Varying Encoder

We next specialize (42) to the dimension-varying encoder. Take $D = 512$ and $d \ge m = 2$. For $\mathbf{x} = (y_1, y_2, 0, \dots, 0) \in \mathbb{R}^D$, the encoder maps

$$\mathbf{z} = E(\mathbf{x}) = (y_1, y_2, 0, \dots, 0) \in \mathbb{R}^d.$$

Since the encoder acts as the identity on the intrinsic subspace, we may take

$$\mathbf{A} = \mathbf{I}_m.$$

Therefore

$$\mathbf{A}^\top \mathbf{A} - \mathbf{I}_m = \mathbf{0}, \qquad \|\mathbf{A}^\top \mathbf{A} - \mathbf{I}_m\|_2 = 0,$$

and the deviation parameter in Theorem 3.3 satisfies

$$\delta = 0.$$

Nonlinear Encoder Consider $D = d = 3$. For $\mathbf{x} = (y_1, y_2, 0)$, consider the encoder

$$E((x_1, x_2, 0), \varepsilon_0) = \left(\frac{\sin(\varepsilon_0 x_1)}{\varepsilon_0},\ x_2,\ \frac{1 - \cos(\varepsilon_0 x_1)}{\varepsilon_0} \right), \qquad \varepsilon_0 > 0.$$

We first derive the linear deviation parameter δ and the dimension penalty term in Theorem B.8. Since $D = d$, the dimension-penalty term in Theorem B.8 satisfies

$$\frac{D - d}{D - m} = 0.$$

Next we derive the linear deviation parameter δ in Theorem B.8. Let $M \subset \mathbb{R}^3$ denote the plane

$$M = \{(x_1, x_2, 0) : x_1, x_2 \in \mathbb{R}\}.$$

Then the distribution μ supported on M is the Gaussian measure

$$d\mu(x_1, x_2) = \frac{1}{2\pi} e^{-(x_1^2 + x_2^2)/2}\, dx_1 dx_2.$$

Let $\mathbf{x}_0 = (a, b, 0)$. With respect to the orthonormal tangent basis $\{e_1, e_2\}$ of $T_{\mathbf{x}_0} M$, the restricted Jacobian is

$$A_{\mathbf{x}_0} = DE(\mathbf{x}_0)\big|_{T_{\mathbf{x}_0} M} = \begin{pmatrix} \cos(\varepsilon_0 a) & 0 \\ 0 & 1 \\ \sin(\varepsilon_0 a) & 0 \end{pmatrix} \in \mathbb{R}^{3 \times 2}.$$

Therefore,

$$A_{\mathbf{x}_0}^\top A_{\mathbf{x}_0} = \begin{pmatrix} \cos^2(\varepsilon_0 a) + \sin^2(\varepsilon_0 a) & 0 \\ 0 & 1 \end{pmatrix} = I_2.$$

Consequently,

$$\delta_{\mathbf{x}_0} = \|A_{\mathbf{x}_0}^\top A_{\mathbf{x}_0} - I_2\|_{\mathrm{op}} = 0.$$

We next compute the localized residual appearing in Theorem B.8. Since the input manifold M is a plane,

$$\mathrm{reach}(M) = \infty.$$

The only nonzero second derivative of E is in the e_1 direction:

$$D^2 E(x)[e_1, e_1] = (-\varepsilon_0 \sin(\varepsilon_0 x_1), 0, \varepsilon_0 \cos(\varepsilon_0 x_1)).$$

Therefore,

$$\|D^2 E(x)\|_{\mathrm{op}} = \varepsilon_0$$

for every $x \in M$.

For

$$x \in \operatorname{supp} \lambda_{\mathbf{x}_0, \tau}^{(r)},$$

we have

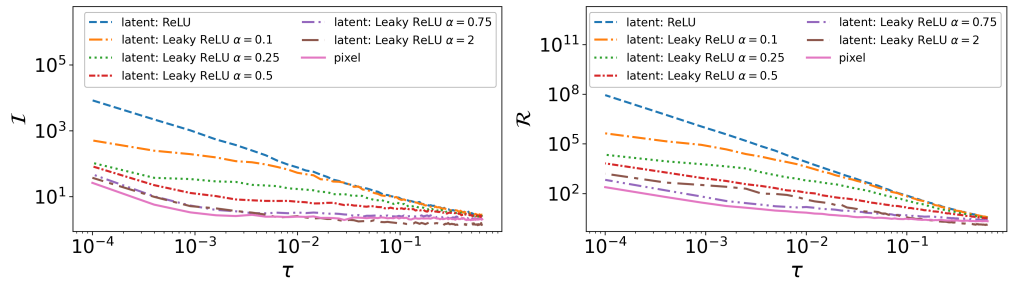
$$\|x - \mathbf{x}_0\| \le r\sqrt{\tau}.$$

The second-order Taylor theorem therefore gives

$$\begin{aligned} \|E(x) - F_{\mathbf{x}_0}(x)\| &\le \frac{\varepsilon_0}{2} \|x - \mathbf{x}_0\|^2 \\ &\le \frac{r^2 \tau}{2} \varepsilon_0. \end{aligned}$$

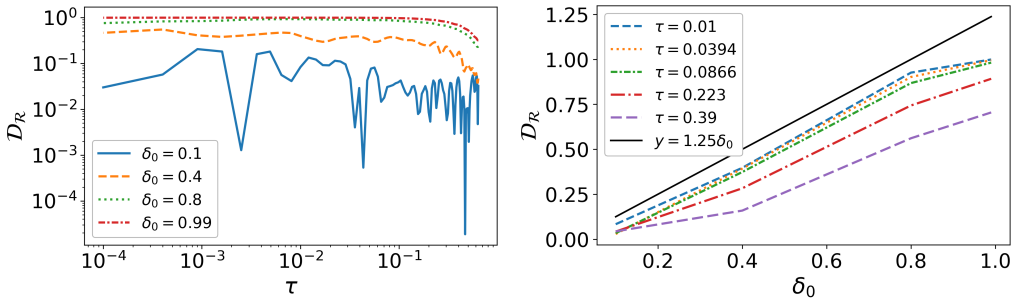
Thus,

$$\epsilon_{\mathbf{x}_0, \tau} \le \frac{r^2 \tau}{2} \varepsilon_0,$$



(a) (b)

Figure 4: Values of (c) $\mathcal{I}$ and (d) $\mathcal{R}$ plotted versus the noise variance τ for the ReLU-based experiments. *Pixel* curves correspond to models trained on $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, while *latent* curves correspond to models trained on $\mathbf{z} = E(\mathbf{x})$. For Leaky ReLU, α denotes the negative slope.



(a) (b)

Figure 5: FIR deviation $\mathcal{D}_\mathcal{R}$ vs. (a) τ and (b) δ_0 in toy settings. Data $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ are embedded as $\mathbf{x} = (y_1, y_2, 0, \ldots, 0) \in \mathbb{R}^D$ and encoded to $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^d$. We consider $D = d = 2$ and $E(\mathbf{x}) = \mathbf{A}\mathbf{x}$ with $\mathbf{A} = \mathrm{diag}(\sqrt{1+\delta_0}, \sqrt{1-\delta_0})$. We compute $\mathcal{D}_\mathcal{R}$ from $\mathcal{R}^{(D)}(\mu_\tau)$ and $\mathcal{R}^{(d)}((\mu_Z)_\tau)$ using diffusion models trained on $\mathbf{x}$ and $\mathbf{z}$. Solid line $y = 1.25\delta_0$ in (b) serves as a linear reference.

uniformly in the expansion point $\mathbf{x}_0$.

Whenever

$$\frac{r^2 \varepsilon_0 \sqrt{\tau}}{2} \leq a_0,$$

Theorem B.8 gives

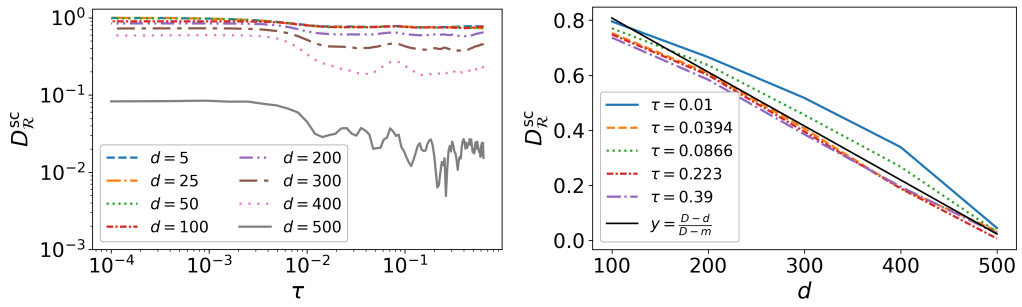
$$\left| 1 - \frac{\mathcal{R}(P_{\mathbf{x}_0,\tau})}{\mathcal{R}(Q_{\mathbf{x}_0,\tau})} \right| \leq C r^2 \varepsilon_0 \sqrt{\tau}.$$

The experiment in Figure 2(c) evaluates a global FIR deviation for an unbounded Gaussian distribution, so it illustrates the dependence on ε_0 qualitatively but is not a direct numerical verification of this localized bound.

E.3 Supplementary Plots

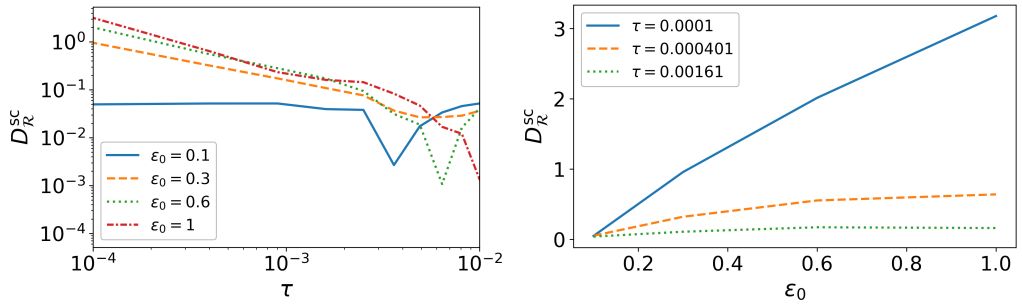
In §4.1, Figure 1(a,b) compares FI and FIR for Gaussian data encoded using several common activation functions, including ReLU, Leaky ReLU, GELU, sigmoid, and tanh. Here, we provide an extended comparison focusing on ReLU and Leaky ReLU over additional values of the negative slope α. The corresponding FI and FIR curves as functions of the noise variance τ are shown in Figure 4.

In §4.2, Figure 2 presents the FIR deviation $\mathcal{D}_\mathcal{R}$ as functions of the parameters δ_0, d, and ε_0 for fixed noise variance τ. Here we provide additional plots that instead show the dependence of $\mathcal{D}_\mathcal{R}$ on τ for the same experimental settings. These results are shown in Figures 5, 6, and 7.



(a) (b)

Figure 6: FIR deviation $\mathcal{D}_{\mathcal{R}}$ vs. (a) τ and (b) d in toy settings. Data $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ are embedded as $\mathbf{x} = (y_1, y_2, 0, \ldots, 0) \in \mathbb{R}^D$ with $D = 512$, and $\mathbf{z} = (y_1, y_2, 0, \ldots, 0) \in \mathbb{R}^d$. We compute $\mathcal{D}_{\mathcal{R}}$ from $\mathcal{R}^{(D)}(\mu_\tau)$ and $\mathcal{R}^{(d)}((\mu_Z)_\tau)$ using diffusion models trained on $\mathbf{x}$ and $\mathbf{z}$. Solid line $y = \frac{D-d}{D-2}$ in (b) serves as a linear reference.



(a) (b)

Figure 7: FIR deviation $\mathcal{D}_{\mathcal{R}}$ vs. (a) τ and (b) ε_0 in toy settings. Data $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ are embedded as $\mathbf{x} = (y_1, y_2, 0, \ldots, 0) \in \mathbb{R}^D$ and encoded to $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^d$. We consider $D = d = 3$ and $E((x_1, x_2, 0), \varepsilon_0) = \big(\sin(\varepsilon_0 x_1)/\varepsilon_0,\ x_2,\ (1 - \cos(\varepsilon_0 x_1))/\varepsilon_0\big)$. We compute $\mathcal{D}_{\mathcal{R}}$ from $\mathcal{R}^{(D)}(\mu_\tau)$ and $\mathcal{R}^{(d)}((\mu_Z)_\tau)$ using diffusion models trained on $\mathbf{x}$ and $\mathbf{z}$.

F Additional Experiments

This section presents additional FI and FIR comparisons on FFHQ using latent representations from VAE, KL-AE, and NVAE encoders.

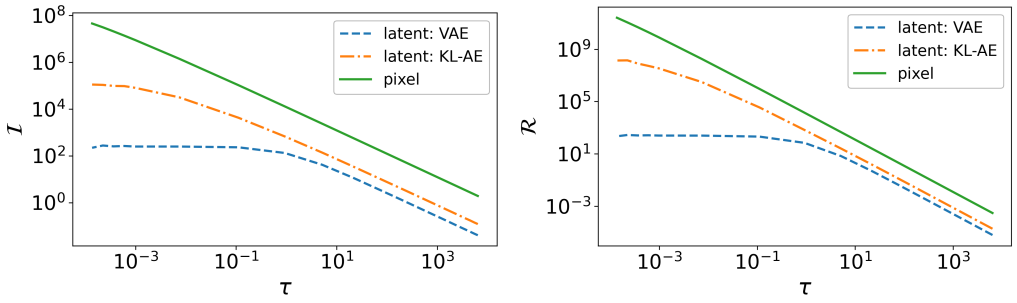
F.1 FI and FIR Comparison on FFHQ with VAE and KL-AE Encodings

F.1.1 Experimental Setup

We use the FFHQ dataset (Karras et al., 2019), which contains 70,000 high-quality face images[8], and downsample the images to 64×64. We compare FI and FIR in pixel space with those in latent representations produced by a traditional VAE and a KL-regularized autoencoder (KL-AE) (Rombach et al., 2022). For the latent-space experiments, images are encoded using either autoencoder, and the resulting latents are scaled by their empirical standard deviation following the latent-diffusion setup (Rombach et al., 2022). We compare three diffusion models: a pretrained pixel-space model and two models trained on the VAE and KL-AE latent representations, respectively.

VAE and KL-AE. The traditional VAE employs a hybrid encoder–decoder architecture that combines convolutional layers with fully connected layers for latent-space projection and feature recon-

[8]Dataset license: CC BY-NC-SA 4.0. Individual Flickr images are released under CC BY 2.0, CC BY-NC 2.0, Public Domain Mark 1.0, CC0 1.0, or U.S. Government Works licenses.



(a) (b)

Figure 8: Values of (a) $\mathcal{I}$ and (b) $\mathcal{R}$ plotted versus the noise variance τ, computed from diffusion models trained on different data representations. The *pixel* curves correspond to models trained directly on FFHQ images. The *latent* curves correspond to models trained on latent representations of an image encoder (VAE or KL-AE) pretrained on FFHQ. We show $\sqrt{\tau} \in [0.01, 80]$, excluding smaller τ due to numerical instability.

struction. We adopt the architecture configuration provided for CelebA at 64×64 resolution in the implementation[9] and set the latent dimension to 256. The VAE is trained on 64×64 FFHQ images using a batch size of 100 and a learning rate of 1×10^{-4}. We use a β-annealing strategy that gradually increases the KL-divergence weight. Specifically, we begin with warm-up stages using $\beta \in 0, 10^{-3}, 10^{-2}$, training each stage for 5 hours, and then train with $\beta = 0.1$ for 20 hours on one GPU.

The KL-AE uses a fully convolutional architecture. We use the pretrained KL autoencoder with downsampling factor $f = 4$ from the official latent-diffusion implementation[10]. This model was trained on 256×256 OpenImages data (Kuznetsova et al., 2020) as a general-purpose image autoencoder. When applied to the 64×64 FFHQ images, its downsampling factor produces latent representations of dimension $3 \times 16 \times 16$.

Diffusion Models. We use the convolutional U-Net architecture in the EDM framework (Karras et al., 2022), following the official implementation[11]. For the pixel-space experiment, we use the pretrained 64×64 FFHQ checkpoint provided in the EDM repository. For the latent-space experiments, we train separate diffusion models on the normalized representations produced by the VAE and KL-AE. The pixel-space and latent-space models use the same convolutional U-Net design and comparable training configurations, with residual block configuration $(1, 2, 2, 2)$ across resolution levels.

For latent diffusion training, we use a batch size of 256, dropout rate 0.05, and a learning-rate schedule consisting of a 10,000-kimg warm-up, where one kimg denotes one thousand training images, followed by a constant learning rate of 2×10^{-4}. Optimization is performed using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The models are trained for 200,000 kimg. The pretrained pixel-space model uses data augmentation with probability 0.15, whereas no data augmentation is applied when training the latent-space diffusion models.

For sampling, we follow the standard EDM procedure using 40 steps, $\sigma_{\min} = 0.002$, and $\sigma_{\max} = 80$. We set $S_{\text{churn}} = 0$ for the pixel-space and VAE latent models and $S_{\text{churn}} = 40$ for the KL-AE latent model.

F.1.2 Results

FI and FIR. Figure 8 compares the FI and FIR curves of the pixel-space, VAE latent-space, and KL-AE latent-space diffusion models, while Figure 9 reports the scaled FIR deviation. For the scaling factor in $\mathcal{D}^{\text{sc}}_{\mathcal{R}}$, we use the pixel dimension $D = 3 \times 64 \times 64$, the latent dimensions $d = 256$

[9] https://github.com/wonjunee/GPE_codes; license: custom research-use license.
[10] https://github.com/CompVis/latent-diffusion/tree/main; license: MIT.
[11] https://github.com/NVlabs/edm; license: CC BY-NC-SA 4.0.

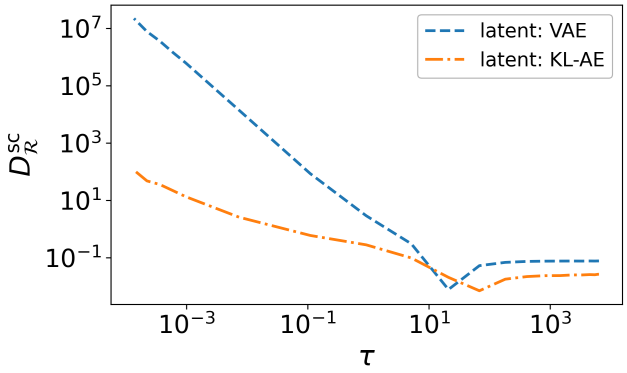
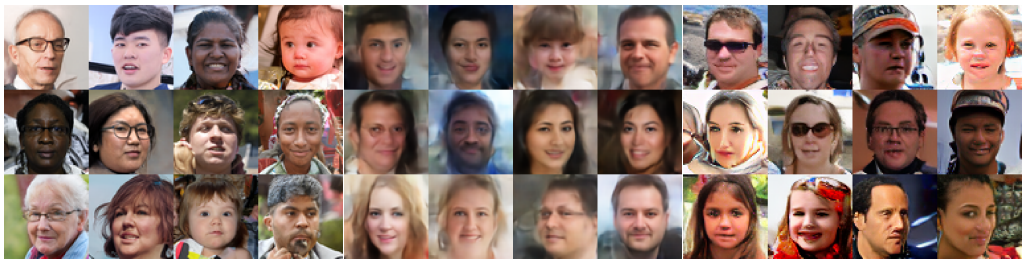


Figure 9: Dimension-normalized FIR deviation $\mathcal{D}^{sc}_{\mathcal{R}}$ vs. noise variance τ on FFHQ. Here, $D = 64 \times 64 \times 3$, $m = 20$, and the latent dimension is $d = 256$ for the VAE and $d = 3 \times 16 \times 16$ for the KL-AE. FFHQ images $\mathbf{x}$ are encoded as latents $\mathbf{z} = E(\mathbf{x})$ using the VAE or KL-AE.



(a) (b) (c)

Figure 10: (a) Samples from a diffusion model trained directly on 64×64 FFHQ. (b, c) Samples from latent diffusion models trained on (b) VAE and (c) KL-AE representations. Generated latents are mapped back to image space using their respective decoders.

for VAE and $d = 3 \times 16 \times 16$ for KL-AE, and the intrinsic dimension $m = 20$. The choice of m follows intrinsic-dimension estimates of approximately 20 for CelebA (Pope et al., 2021; Liu et al., 2015), another human-face dataset.

The KL-AE curves lie consistently closer to the pixel-space baseline than the VAE curves for both FI and FIR. KL-AE also has a smaller scaled FIR deviation, indicating better agreement with the pixel-space FIR.

The smaller FIR deviation of KL-AE is qualitatively consistent with the local mechanisms isolated by Theorem B.8 and Corollary B.11. KL-AE is fully convolutional and retains a spatial latent structure of size $3 \times 16 \times 16$, whereas the traditional VAE compresses each image into a 256-dimensional vector. This architecture may better preserve local neighborhoods and may introduce smaller tangential distortion and local affine residuals.

The reported FFHQ quantity is a global FIR deviation, while the theorem concerns localized Gaussian-smoothed laws. We therefore interpret the empirical trends as qualitative evidence for the geometric mechanisms rather than as a direct numerical verification of the theorem or its constants.

Table 2: Reconstruction MSE (mean squared error) for the VAE and KL-AE trained on FFHQ.

Model	**Reconstruction MSE**
VAE	0.0094
KL-AE	0.41

Table 3: Generation quality on FFHQ. FID (Fréchet Inception Distance) is computed using two reference distributions: the original FFHQ images and the reconstructions produced by the corresponding autoencoder. The first FID column measures generation quality relative to the original image distribution, while the second measures generation quality relative to the reconstruction distribution of the corresponding latent model. For the pixel model, images are generated by sampling a diffusion model trained directly on FFHQ. For the VAE and KL-AE models, images are generated by sampling diffusion models trained in the latent space and decoding the generated latents using the corresponding decoder. All diffusion models use the U-Net architecture within the EDM framework.

Model	**FID vs. FFHQ**	**FID vs. Reconstructions**
Pixel Diffusion	2.56	–
VAE (Latent Diffusion)	62.51	–
KL-AE (Latent Diffusion)	14.97	3.58

Table 4: Empirical estimates $(\hat{c}, \hat{C})$ of the constants (c, C) in Proposition B.1 for encoders of the VAE and SD-VAE trained on FFHQ. The ratio $\hat{C}/\hat{c}$ provides an empirical estimate of the bi-Lipschitz constant.

Model	$\hat{c}$	$\hat{C}$	$\hat{C}/\hat{c}$
VAE	0.20	3.86	19.67
SD-VAE	1.60	4.83	3.02

Generation and reconstruction results. Figure 10 shows representative samples generated by the three diffusion models. Table 2 reports the reconstruction MSE for the VAE and KL-AE on FFHQ, while Table 3 reports the FID scores of the pixel-space and latent-space diffusion models.

For KL-AE, we use a pretrained checkpoint trained on 256×256 OpenImages data. When applied to the 64×64 FFHQ images, its decoder tends to sharpen the reconstructions relative to the original downsampled images, leading to a larger reconstruction MSE. To account for this autoencoder-induced distribution shift, we also report its generation FID relative to KL-AE reconstructions.

The generation results are consistent with the FI/FIR comparison. The traditional VAE latent diffusion model performs poorly. For KL-AE, the FID relative to reconstructions is much lower than the FID relative to FFHQ, but remains higher than the pixel-space FID. The generated samples similarly show that KL-AE produces higher visual quality than VAE, although it still falls short of the pixel-space baseline.

Empirical estimation of c and C. For the VAE and KL-AE encoders, we estimate the pairwise distance distortion corresponding to the constants c and C in Proposition B.1. In the proposition, these constants are defined through the tangential Jacobian bounds

$$c\|\mathbf{v}\| \le \|J_E(\mathbf{x})\mathbf{v}\| \le C\|\mathbf{v}\|, \qquad \mathbf{x} \in \mathcal{M}, \quad \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}.$$

To estimate these constants empirically, we consider the corresponding distance inequality

$$\hat{c}\|\mathbf{x}_1 - \mathbf{x}_2\| \le \|E(\mathbf{x}_1) - E(\mathbf{x}_2)\| \le \hat{C}\|\mathbf{x}_1 - \mathbf{x}_2\|.$$

To estimate $(\hat{c}, \hat{C})$, we sample 5000 images $\{\mathbf{x}_i\}$ from the FFHQ dataset and compute their encoded representations $\{E(\mathbf{x}_i)\}$. For each pair (i, j), we evaluate the distortion ratio

$$r_{ij} = \frac{\|E(\mathbf{x}_i) - E(\mathbf{x}_j)\|}{\|\mathbf{x}_i - \mathbf{x}_j\|}.$$

Using all pairs, we estimate the constants as

$$\hat{c} \approx \min_{i<j} r_{ij}, \qquad \hat{C} \approx \max_{i<j} r_{ij}.$$

The resulting estimates are reported in Table 4, together with the ratio $\hat{C}/\hat{c}$, which provides an empirical measure of the bi-Lipschitz distortion.

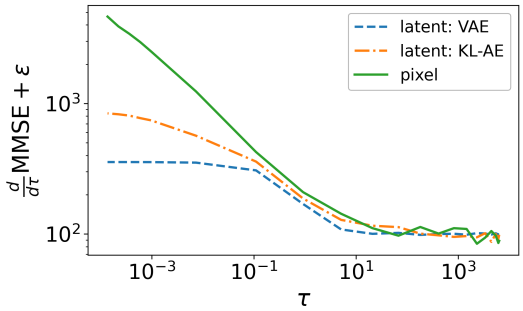


Figure 11: Sensitivity of MMSE to τ, plotted against the noise variance τ, computed from diffusion models trained on different data representations. The *pixel* curve corresponds to a model trained directly on FFHQ images, while the *latent* curves correspond to models trained on VAE and KL-AE latent representations. We show $\sqrt{\tau} \in [0.01, 80]$, excluding smaller noise levels due to numerical instability. For visualization on a logarithmic scale, we plot $\frac{d}{d\tau}\text{MMSE} + \varepsilon$ with $\varepsilon = 100$.

As shown in Table 4, KL-AE has a substantially smaller distortion ratio than the traditional VAE, with $\hat{C}/\hat{c} = 3.02$ compared with 19.67. Its ratio is therefore closer to one, which is qualitatively consistent with its improved FI alignment and the near-isometry mechanism in Proposition B.1.

These finite-sample extrema do not certify the global bi-Lipschitz constants c and C, and we do not estimate the logarithmic volume-Jacobian term $\widetilde{\varepsilon}$. Moreover, Proposition B.1 concerns intrinsic FI under deterministic pushforward, whereas the plotted quantities are ambient Gaussian-smoothed FIs. The comparison is therefore descriptive rather than a direct numerical verification of the proposition.

MMSE results. Figure 11 shows the sensitivity of MMSE to τ, as defined in Equation 6. We set $k = D = 3 \times 64 \times 64$ for the pixel-space model and $k = d$ for the latent-space models, with $d = 256$ for VAE and $d = 3 \times 16 \times 16$ for KL-AE.

The pixel-space curve has the largest sensitivity at small τ, indicating that the MMSE changes rapidly in the low-noise regime. Both latent curves lie below the pixel curve for small τ, suggesting that the latent representations are less sensitive to fine-scale noise perturbations than the original pixel space. The KL-AE curve follows the decreasing trend of the pixel curve more closely than the VAE.

F.2 FI and FIR Comparison on FFHQ Data with NVAE Encoding

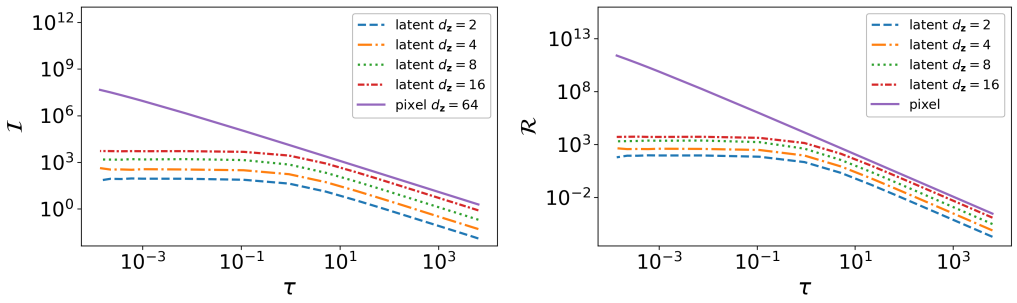
F.2.1 Experimental Setup

We conduct experiments on the FFHQ dataset (Karras et al., 2019), using images downsampled to 64×64 resolution. For the pixel-space experiment, we use the same diffusion model described in §F.1. For the latent-space experiments, we use a pretrained NVAE (Vahdat & Kautz, 2020) to encode the images at multiple spatial resolutions and train a separate diffusion model on the latent representations selected at each resolution.

NVAE. We use a publicly available pretrained NVAE model trained on FFHQ.[12] NVAE is a hierarchical variational autoencoder whose encoder produces multiple latent tensors at different spatial resolutions. To remain consistent with the pixel-space experiment, we provide 64×64 images as input to the encoder. Although the pretrained model was originally trained on 256×256 images, its fully convolutional encoder can process different input resolutions without architectural modifications.

The encoder produces latent tensors of size $20 \times d_{\mathbf{z}} \times d_{\mathbf{z}}$ at spatial resolutions $d_{\mathbf{z}} \in 32, 16, 8, 4, 2$, where 20 is the number of channels. When multiple latent tensors occur at the same resolution, we select the one closest to the input space in the encoder hierarchy. We omit $d_{\mathbf{z}} = 32$ because

[12]We use the publicly available pretrained NVAE model: `https://github.com/NVlabs/NVAE`; license: NVIDIA Source Code License.



(a) (b)

Figure 12: Values of (a) $\mathcal{I}$ and (b) $\mathcal{R}$ plotted versus the noise variance τ for diffusion models trained on different data representations. *Pixel* and *latent* curves denote models trained on FFHQ images and their pretrained NVAE latents, respectively. NVAE was pretrained on FFHQ, with spatial size $20 \times d_{\mathbf{z}} \times d_{\mathbf{z}}$. We show $\sqrt{\tau} \in [0.01, 80]$, excluding smaller τ due to numerical instability.

its dimensionality, $20 \times 32 \times 32$, exceeds that of the input image and therefore does not provide a compressed representation.

Diffusion Models. For each selected latent resolution, we train a separate diffusion model under the EDM framework (Karras et al., 2022), using a convolutional U-Net as the score network. The architecture and training configuration follow those described in §F.1. For the smallest latent resolutions, we simplify the U-Net to accommodate the limited spatial dimensions: we use residual block configurations $(1, 2, 2)$ for $d_{\mathbf{z}} = 4$ and $(1, 2)$ for $d_{\mathbf{z}} = 2$.

F.2.2 Results

Figure 12 compares FI and FIR computed using the above pixel-space and latent diffusion models. As $d_{\mathbf{z}}$ increases, the FI and FIR curves for the latent models move closer to the pixel baseline. This behavior can be explained by the hierarchical structure of NVAE. Latent variables with larger $d_{\mathbf{z}}$ lie closer to the input space and therefore preserve more of the local geometry of the data manifold, resulting in less distortion in the mapping from data space to the latent representation.

We also observe that for both FI and FIR, the pixel and latent curves remain closely aligned for large τ, while the deviation increases noticeably once τ becomes smaller. A possible explanation is that when τ is large, the distribution p_τ is strongly smoothed by Gaussian noise, so the corresponding score functions become dominated by noise and the resulting FIR values carry limited geometric information about the underlying representation.

G Spectra and Connection to Fisher Geometry

In the main text, we introduce Fisher information (FI) and Fisher information rate (FIR) as tools for analyzing the diffusability of autoencoder representations. In this section, we examine spectral analysis as an alternative perspective for studying diffusability. Corresponding to the experiments in Sections F.1 and F.2, which compare FI and FIR between pixel space and latent space, we analyze the power spectra of FFHQ images and of latent representations produced by the VAE, KL-AE, and NVAE encoders. This analysis provides a complementary view of the signal structure and allows us to contrast the information revealed by spectral behavior with that captured by FI and FIR.

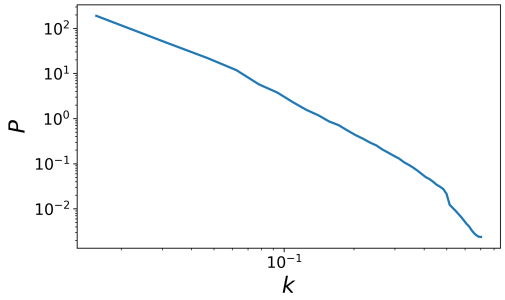
We use the same 64×64 resolution FFHQ images and the same VAE, KL-AE, and NVAE models trained on FFHQ data as in the experiments in Sections F.1 and F.2. For FFHQ images, KL-AE latent representations, and NVAE latent representations, which are two-dimensional spatial tensors, we compute the two-dimensional power spectrum by applying the Fast Fourier Transform (FFT) over the spatial dimensions. For multi-channel inputs, the resulting spectra are averaged across channels to obtain a single spectrum per sample.

For the latent representations produced by the VAE encoder, the latent variable for each sample is a one-dimensional vector. To analyze their frequency characteristics, we compute the one-dimensional power spectrum using the discrete Fourier transform (FFT) along the latent dimension. Each latent vector is normalized prior to the transform.

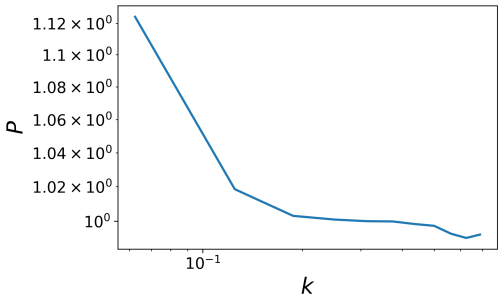
Both the frequency axis k and the power values P are normalized by the signal length (or spatial resolution for the 2D case). To obtain a stable estimate, the spectrum is averaged over $10{,}000$ samples drawn from the dataset. The zero-frequency (DC) component corresponding to the mean signal level is excluded from the plots.

Figure 13(a) shows the power spectrum of FFHQ images. Figures (b) and (c) show the spectra of NVAE latent representations at different spatial resolutions, while Figures (d) and (e) show the spectra of the VAE and KL-AE latent representations. Although the spectra in (b), (c), and (e) also decrease with frequency, their shapes differ from the pixel-space spectrum in (a), and the decay behavior varies across resolutions, indicating a mismatch with the spectral structure of the original images. Unlike the pixel spectrum in (a), the VAE latent spectrum in (d) exhibits strong oscillations and lacks a clear monotonic decay, further deviating from the spectral behavior of the original images.

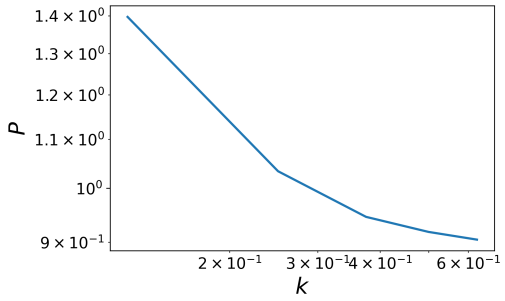
Overall, the latent spectra do not always align with the spectrum of the original images. This suggests that spectral statistics alone may not be the most reliable diagnostic for assessing diffusability. In contrast, the FI and FIR analysis in Sections F.1 and F.2 provides a more consistent characterization of the diffusability of the corresponding representations. We also note that spectral statistics are not permutation invariant: for example, randomly shuffling the pixels of an image would destroy its spectrum by removing the spatial structure on which the Fourier transform relies. In contrast, FI and FIR depend on the geometry of the probability density through the score norm $\|\nabla_{\mathbf{x}} \log p(\mathbf{x})\|^2$, and are therefore invariant under permutations of coordinates.



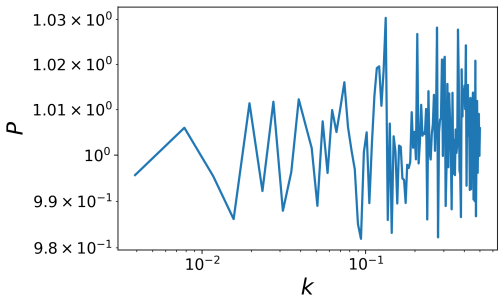
(a) FFHQ Images



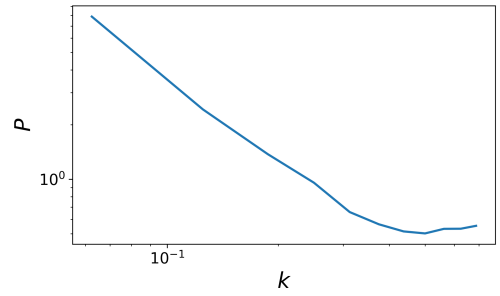
(b) NVAE latents ($d_z = 16$)



(c) NVAE latents ($d_z = 8$)



(d) VAE latents



(e) KL-AE latents

Figure 13: (a) Power spectrum of FFHQ images. (b)(c) Power spectra of NVAE latent representations with spatial resolutions $d_z = 16$ and 8. (d)(e) Power spectra of VAE and KL-AE latent representations. The NVAE, VAE, and KL-AE models are pretrained on FFHQ images. For FFHQ images, NVAE latents, and KL-AE latents, the spectrum is computed using the two-dimensional FFT over the spatial dimensions and averaged over channels. For the one-dimensional latent vectors produced by the VAE encoder, the spectrum is computed using a one-dimensional FFT along the latent dimension. The spectra are averaged over 10,000 samples drawn from the dataset. Both the frequency axis k and the power values P are normalized by the signal length (or spatial resolution for the 2D case). The zero-frequency (DC) component corresponding to the mean signal level is excluded.