REVISITING DETERMINISTIC DIFFUSION THROUGH REVERSE TRANSITION KERNELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deterministic diffusion samplers replace stochastic reverse dynamics with an iterative reconstruction map, but the conditions under which this iteration is stable remain poorly understood. We study Directly Denoising Diffusion Models (DDDMs) (Zhang et al., 2024), which refine an estimate of the clean sample by a fixed-point iteration at each noise level. Through a reverse transition kernel formulation, each reverse-time step defines a residual energy, and Tweedie's identity yields a certificate of local strong convexity for this energy—a sufficient condition for stable refinement. We audit this certificate at scale on trained models and report three findings. First, the derivative that governs convergence is taken with respect to the sampler's own running estimate, whereas Tweedie's identity concerns derivatives with respect to the noisy observation; on trained networks these are measurably different operators, and only the latter exhibits the symmetry Tweedie predicts. Second, the curvature term in the certificate can be computed exactly rather than bounded, tightening it by a median factor of 3–33. Third, across 960 configurations spanning variance-preserving models under linear and cosine schedules and variance-exploding models, the energy is strongly convex more than twice as often as even the exact certificate detects (69% versus 29%, with the worst-case bound certifying almost none); we trace the remaining conservatism to a single geometric assumption—that the certificate's two terms act along a common worst-case direction, where in measurement they are nearly orthogonal. A one-dimensional system with closed-form posterior statistics reproduces the predicted contraction rate to within 1.7% and shows that the certificate, not the sampler, loses resolution at low noise: where certification fails, the iteration itself still converges.

1 INTRODUCTION

Diffusion models (Song et al., 2021; Song & Ermon, 2020b; Ho et al., 2020) and, more recently, flow matching methods (Lipman et al., 2023; Albergo et al., 2025; Liu et al., 2025) represent two closely related paradigms that currently achieve state-of-the-art performance in generative modeling across images, video, audio, and molecular structures.

A common computational challenge in continuous-time generative models is that sample generation requires numerical integration of learned reverse-time SDEs or ODEs (Song et al., 2021; Holderrieth & Erives, 2026). Deterministic generative methods, including DDIM (Song et al., 2022), Probability Flow ODEs (Na et al., 2025), DPM-Solver (Lu et al., 2022), DPM-Solver++ (Lu et al., 2025), Flow Matching (Lipman et al., 2023), and Rectified Flow (Liu et al., 2022), generate samples by deterministically evolving the sampling state along a learned score or velocity field. Existing diffusion approaches either shift computation to training-time, requiring costly retraining for new objectives (Ho et al., 2020; Nichol & Dhariwal, 2021; Karras et al., 2022), or rely on inference-time modifications that introduce additional approximation error and computational overhead (Dhariwal & Nichol, 2021; Chung et al., 2024; Song et al., 2021). This has motivated the development of methods that reduce the number of sampling steps while maintaining generation quality. Recent work on accelerated sampling (Jutras-Dubé et al., 2025; zhangkai wu et al., 2025), have proposed flow maps (Boffi et al., 2025a; Li et al., 2025), which learn to approximate the composition of multiple diffusion or flow steps via a single neural network evaluation. This amortization of sequential inference reduces computational cost while preserving the structure of the underlying generative process.

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.

Existing DDDM formulations define an iterative denoising map, but leave the local geometry governing its reconstruction stability largely unexplored. We use the RTK formulation to associate each reverse-time subproblem with an implicit reconstruction energy, providing a tractable object for analyzing this geometry. The contribution is not the representation of a deterministic map through a residual or energy, but the analysis enabled by the specific reconstruction structure of DDDM. In particular, the resulting residual energy has a Hessian that can be explicitly decomposed into a Gauss–Newton term and a residual-dependent curvature term involving derivatives of the denoising correction. Using Tweedie's identity, we further relate these derivatives to posterior moments, allowing us to connect the local curvature of the DDDM reconstruction objective to the statistics of the underlying posterior. This connection yields a strong-convexity certificate in terms of the denoiser Jacobian, reconstruction residual, and second-order curvature, which we subsequently evaluate against the actual Hessian geometry across diffusion settings.

2 Background

2.1 Diffusion Models

DDPMs. Diffusion models (DDPMs) (Ho et al., 2020) are generative models that learn data distributions via a forward process (adding noise) and a reverse process (removing noise). Let $\{\beta_i\}_{i=1}^T$ denote a sequence of positive noise scales such that $0 < \beta_1, \beta_2, \ldots, \beta_T < 1$. The forward process gradually adds Gaussian noise to the data x_0 over time steps t in a Markov chain: $p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t;\, \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\, (1-\bar{\alpha}_t)\mathbf{I})$, where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the cumulative noise schedule and $x_0 \sim p^* \propto \exp(-f^*)$ is a sample from the true data distribution and $f^*(x)$ is the corresponding energy function. We consider a discrete-time Markov chain $x_0, x_1, \ldots, x_T$. The perturbed data distribution is denoted as $p_\alpha(\hat{\mathbf{x}}) := \int p^*(\mathbf{x})\, p_\alpha(\hat{\mathbf{x}} \mid \mathbf{x})\, d\mathbf{x}$. A variational Markov chain in the reverse direction is parameterized as $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\Big(\mathbf{x}_{t-1};\, \frac{1}{\sqrt{1-\beta_t}}\big(\mathbf{x}_t + \beta_t\, s_\theta(\mathbf{x}_t, t)\big),\, \beta_t \mathbf{I}\Big)$. Samples are produced via an ancestral sampling procedure, following $\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_i}}\,(\mathbf{x}_t + \beta_t\, s_{\theta^*}(\mathbf{x}_t, t)) + \sqrt{\beta_t}\,\mathbf{z}_t$. Here, $s_\theta(x_t, t)$ is the score function parameterized by θ, which predicts the noise component in x_t. The optimal parameters θ^* are learned by minimizing the expected denoising error between the predicted and true noise over all timesteps.

Implicit Models. The induced joint distribution over latent variables is given by $p(x_{1:T} \mid x_0) := p(x_T \mid x_0) \prod_{t=2}^{T} p(x_{t-1} \mid x_t, x_0)$. The terminal marginal distribution is specified as $p(x_T \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_T}\, x_0, (1-\bar{\alpha}_T)\mathbf{I})$, where $\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$. For all $t > 1$, the backward kernels are defined as $p(x_{t-1} \mid x_t, x_0) = \mathcal{N}\Big(\sqrt{\bar{\alpha}_{t-1}}\, x_0 + \sqrt{1-\bar{\alpha}_{t-1}-\beta_t}\, \frac{x_t - \sqrt{\bar{\alpha}_t}\, x_0}{\sqrt{1-\bar{\alpha}_t}},\ \beta_t\, \mathbf{I}\Big)$. The mean functions are constructed such that the induced marginals satisfy $p(x_t \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\mathbf{I})$ for all t, ensuring consistency across time steps. Applying Bayes' rule yields the induced forward transition $p(x_t \mid x_{t-1}, x_0) = \frac{p(x_{t-1} \mid x_t, x_0)\, p(x_t \mid x_0)}{p(x_{t-1} \mid x_0)}$. Song et al. (2022) defined an oracle reverse-time inference process conditioned on the clean data sample x_0, which serves as a theoretical reference distribution rather than a directly implementable sampling algorithm.

Implicit Forward Process. *The forward transitions in implicit diffusion models do not correspond to the Markovian noising process used in DDPMs. Instead, they arise from an oracle-based inference construction and are therefore generally non-Markovian, with explicit dependence on the clean sample x_0. A tractable generative procedure is obtained by replacing this oracle dependence with a learned prediction $\hat{x}_0 = f_\theta(x_t)$, yielding a reverse-time process consistent with the DDIM parameterization. In this setting, choosing $\sigma_t^2 = \beta_t$ corresponds to the stochastic formulation, while the limit $\sigma_t \to 0$ recovers the deterministic (implicit) case.*

Directly Denoising Diffusion. From Song & Ermon (2020b) reformulation of DDPMs as Stochastic Differential Equations (SDEs) to generalize the forward process: $d\mathbf{X}_t = -\frac{1}{2}\beta(t)\mathbf{X}_t\, dt + \sqrt{\beta(t)}\, d\mathbf{B}_t$, where $\mathbf{B}_t$ represents Brownian motion. The reverse-time VP SDE is: $d\mathbf{X}_t = \big[-\frac{1}{2}\beta(t)(\mathbf{X}_t + \nabla \log p_t(\mathbf{X}_t))\big]\, dt + \sqrt{\beta(t)}\, d\mathbf{B}_t$, where $\nabla \log p_t$ is estimated by a time-dependent score network $s_\theta(\mathbf{x}, t)$. Directly Denoising Diffusion Models (DDDMs) (Zhang et al., 2024) integrate DDPMs with the Probability Flow (PF) ODE framework, enabling faster denoising without complex solvers. The solution of the PF ODE is obtained by evaluating the integral expression

$\mathbf{x}_0 = \mathbf{x}_T + \int_0^T -\frac{1}{2}\beta(t)[\mathbf{x}_t + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)]\, dt$, where $\mathbf{x}_T \sim \mathcal{N}(0, I)$. Directly Denoising Diffusion Models (DDDM) refine the estimate of the clean state $\mathbf{x}_0$ by leveraging the probability flow ODE. Specifically, the mapping $f(\mathbf{x}_0, \mathbf{x}_t, t) = \mathbf{x}_t - F(\mathbf{x}_0, \mathbf{x}_t, t)$ is defined, where F involves an integral of the drift term parameterized by the noise schedule $\beta(t)$. A neural approximation $f_\theta(\mathbf{x}_0, \mathbf{x}_t, t) = \mathbf{x}_t - F_\theta(\mathbf{x}_0, \mathbf{x}_t, t)$ is trained such that $f_\theta \approx f$.

$$\Theta := \arg\min_\theta \; \mathbb{E}_{t\sim\mathcal{U}[1,T]}\left[\mathbb{E}_{\mathbf{x}_0\sim p_{\text{data}}(\mathbf{x}_0)}\left[\mathbb{E}_{\mathbf{x}_t\sim\mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0,(1-\bar{\alpha}_t)\mathbf{I}\right)}\left[d\big(f_\theta(\mathbf{x}_0^{(n)}, \mathbf{x}_t, t), \mathbf{x}_0\big)\right]\right]\right] \quad (1)$$

where, $d(\cdot,\cdot)$ is a suitable distance metric. We define Θ as the set of optimal neural parameters for the denoising network.

Reverse Transition Kernel Framework. The Reverse Transition Kernel (RTK) framework decomposes reverse diffusion into a small number of transition sampling problems,

$$\hat{x}_{t_{k+1}} \sim \hat{p}_{t_{k+1}|t_k}\left(\cdot \mid \hat{x}_{t_k}\right),$$

while maintaining approximate marginal consistency with the reverse diffusion process. With an appropriate decomposition, the number of transition problems can be substantially smaller than the original diffusion discretization, potentially $\widetilde{\mathcal{O}}(1)$. When the transition targets are strongly log-concave, each subproblem has a strongly convex potential and admits a unique mode, enabling efficient and stable sampling.

3 Method

In this section, we develop an optimization-theoretic formulation of deterministic diffusion by leveraging the Reverse Transition Kernel (RTK) framework. We begin from the probability flow ODE, which characterizes deterministic reverse-time diffusion dynamics, and show that its Euler discretization naturally decomposes the reverse process into a sequence of RTK subproblems parameterized by a step size η. We then derive a proxy-energy objective associated with each reverse update, providing an optimization interpretation of deterministic denoising. Our aim is to establish a mathematical framework that supports three key capabilities:

1. **Reconstruction formulation:** Express each deterministic reverse update as an RTK reconstruction problem with an associated proxy energy.
2. **Stability characterization:** Derive local curvature and strong-convexity conditions that characterize when the iterative reconstruction is well-conditioned and stable.
3. **Empirical assessment:** Quantify the tightness of these conditions by comparing the resulting stability certificates with the actual local Hessian geometry across diffusion settings and model configurations.

3.1 Probability Flow ODE and Deterministic Reverse-Time Dynamics

Following the SDE formulation of diffusion models (Song et al., 2021), the variance-preserving (VP) forward process is given by $dX_t = -\frac{1}{2}\beta(t)X_t\, dt + \sqrt{\beta(t)}\, dB_t$, where $\beta(t)$ is the noise schedule and B_t denotes standard Brownian motion. The corresponding process admits a deterministic probability flow ODE with the same time-marginal distributions, satisfying $p_{T-t} = \overleftarrow{p}_t$ and given by $d\overleftarrow{x}_t = -\frac{1}{2}\beta(t)\left[\overleftarrow{x}_t - \nabla_x \log p_t(\overleftarrow{x}_t)\right] dt$. In practice, the unknown score $\nabla_x \log p_t(x)$ is approximated by a time-dependent neural network $s_\theta(x,t)$, yielding the neural ODE $d\bar{x}_t = -\frac{1}{2}\beta(t)\left[\bar{x}_t + s_{\theta,T-k\eta}(\bar{x}_{k\eta})\right] dt$ over $t \in [k\eta, (k+1)\eta]$. This deterministic reverse-time formulation provides the basis for the DDDM reconstruction procedure considered below.

3.2 Euler Discretization and Directly Denoising Diffusion Models

Let T denote the total diffusion horizon and let $\eta > 0$ be a step size such that $K = T/\eta$. Applying a forward Euler discretization to the probability flow ODE yields the approximation

$$\hat{x}_0 \approx x_{K\eta} - \sum_{k=0}^{K-1} \eta \cdot \frac{1}{2}\,\beta(k\eta)\left[x_{k\eta} - \nabla_x \log p_{k\eta}(x_{k\eta})\right] \quad (2)$$

Directly Denoising Diffusion Models (DDDMs) (Zhang et al., 2024) are built on the discretized probability flow formulation. Instead of simulating the stochastic reverse-time SDE as in DDPMs, DDDMs perform deterministic denoising by iteratively refining an estimate of the clean sample.

At a given reverse iteration k, starting from the noisy state $x_{k\eta}$, DDDM introduces an inner iterative refinement procedure to estimate x_0:

$$\hat{x}_0^{(m+1)} = x_{k\eta} - F_{\theta,k\eta}\left(\hat{x}_0^{(m)}, x_{k\eta}\right), \quad m = 0, \dots, M-1, \tag{3}$$

where $\hat{x}_0^{(m)}$ denotes the estimate of the clean sample at iteration m, and $F_{\theta,k\eta}$ is a neural approximation of the integrated drift term induced by the probability flow ODE.

Our goal in the subsequent sections is to understand the optimization dynamics underlying deterministic DDDM refinement and to characterize the conditions governing its stability and convergence. To this end, we develop a reconstruction-centric theoretical framework based on the local geometry of the residual-based reconstruction energy, relating its curvature to denoiser derivatives, posterior moments, and reconstruction error.

Goal 1. Formulate the deterministic DDDM reconstruction refinement as a fixed-point/reconstruction problem and characterize its induced transition through a residual-based energy.

Proposition 1 (Deterministic Reverse Kernel and Energy Formulation of DDDM)**.** *Consider the DDDM reverse refinement at step k, with current clean-sample estimate $z \equiv \hat{x}_{k\eta} \equiv \hat{x}_0^{(m)}$, defined by Eq. 3. Let the DDDM-induced refinement map be $T_{k\eta}(z) := x_{k\eta} - F_{\theta,k\eta}(z, x_{k\eta})$. Then the refinement induces the deterministic reverse transition kernel $\overleftarrow{p}_{(k+1)\eta|k\eta}(z \mid \hat{x}_{k\eta}) = \delta_{T_{k\eta}(\hat{x}_{k\eta})}(z)$.*
Defining the corresponding residual as $r_{k\eta}(z) := z - T_{k\eta}(z)$, the refinement admits the proxy energy

$$g_{k\eta}(z) = \frac{1}{2\sigma^2}\|r_{k\eta}(z)\|^2 = \frac{1}{2\sigma^2}\left\|z - x_{k\eta} + F_{\theta,k\eta}(z, x_{k\eta})\right\|^2. \tag{4}$$

In particular, any fixed point of $T_{k\eta}$ is a zero-residual minimizer of $g_{k\eta}$, while the associated proxy density $\hat{p}_{k\eta}(z) \propto \exp[-g_{k\eta}(z)]$ provides an energy-based representation whose local geometry is analyzed subsequently.

Remark 1. *The indexing $(k, k+1)$ used below is introduced for notational and interpretational convenience when relating the DDDM update to the RTK formulation. In the original DDDM formulation, the reverse process is not explicitly represented as a reverse diffusion trajectory; rather, DDDM applies a deterministic denoising map iteratively at a prescribed noise level. Thus, k indexes successive denoising refinements rather than discrete points along an explicitly constructed reverse diffusion trajectory.*

In DDDM, each reverse step is implemented as a deterministic reconstruction refinement rather than an explicit stochastic transition. While deterministic transitions can generically be represented by Dirac-valued reverse kernels; this representation provides a convenient framework for analyzing the refinement map. Our focus is therefore not on the deterministic-kernel representation itself, but on the structure and stability of the reconstruction problem induced by the DDDM denoising map.

The fixed-point view immediately yields the convergence result the refinement itself obeys, independently of any energy or curvature information. Write $f(z) := x_{k\eta} - F_{\theta,k\eta}(z, x_{k\eta})$, so that Eq. 3 is the Picard iteration $z^{(m+1)} = f(z^{(m)})$.

Theorem 1 (Convergence of the undamped refinement)**.** *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ be closed with $f(\mathcal{Z}) \subseteq \mathcal{Z}$, and suppose $\gamma := \sup_{z\in\mathcal{Z}} \|\partial F_{\theta,k\eta}/\partial z\|_2 < 1$. Then f has a unique fixed point $z^* \in \mathcal{Z}$ with $r_{k\eta}(z^*) = 0$, and $\|z^{(m)} - z^*\|_2 \le \gamma^m \|z^{(0)} - z^*\|_2$ for any $z^{(0)} \in \mathcal{Z}$. (Proof in Appendix C.)*

Corollary 1 (Clamped refinement)**.** *If Π is the orthogonal projection onto a closed convex set (e.g. the data box) with $\Pi(\mathcal{Z}) \subseteq \mathcal{Z}$, the clamped iteration $z^{(m+1)} = \Pi(f(z^{(m)}))$ converges at the same rate γ to the unique fixed point z^*_Π of $\Pi \circ f$, and $\|r_{k\eta}(z^*_\Pi)\|_2 = \|f(z^*_\Pi) - \Pi f(z^*_\Pi)\|_2$, the clamp gap.*

Two consequences frame the analysis below. The contraction constant γ is the Jacobian norm in the *estimate* slot—the argument the iteration moves—and Theorem 1 needs no step size: the undamped

update of Eq. 3 has none. A constant step size enters only if one refines by explicit gradient descent on the energy, $z \mapsto z - \tau \nabla g(z)$, a *different* algorithm from Eq. 3; under the strong convexity established below, the classical theory certifies any constant $\tau \leq 1/\Lambda$ with linear rate $1 - \tau\mu$ (Appendix C). This discharges Goal 4 in both senses, and fixes which iteration each guarantee is about.

For a fixed noisy state x at noise level σ, we study the geometry of the energy g of Eq. 4 as a function of the reconstruction variable z: we characterize its curvature and derive sufficient conditions under which the reconstruction subproblem is locally strongly convex, i.e. conditions under which the deterministic refinement is locally well conditioned and stable.

Goal 2. Characterize the local geometry and stability of the reconstruction problem by deriving the curvature of g in terms of the Jacobian, Hessian, and reconstruction residual of the denoising correction F. We derive a sufficient condition for local strong convexity.

Proposition 2 (Local Strong Convexity of the Reverse Subproblem). *Let $g(z) = \frac{1}{2\sigma^2}\|z - x + F(z)\|_2^2$ and $r(z) := z - x + F(z)$. Its gradient and Hessian are $\nabla g = \sigma^{-2}(I + J_F)^\top r$ and $\nabla^2 g = \sigma^{-2}[(I + J_F)^\top (I + J_F) + \sum_i r_i H_{F_i}]$. Defining $L := \|J_F\|_2$, $R := \|r\|_2$, and $B_{\mathrm{sq}} := \left(\sum_i \|H_{F_i}\|_{\mathrm{op}}^2\right)^{1/2}$, we have*

$$\lambda_{\min}(\nabla^2 g) \geq \frac{\max(0,\, 1 - L)^2 - R B_{\mathrm{sq}}}{\sigma^2}. \tag{5}$$

Thus, if $\max(0, 1 - L)^2 > R B_{\mathrm{sq}}$, then g is locally μ-strongly convex with $\mu = [\max(0, 1 - L)^2 - R B_{\mathrm{sq}}]/\sigma^2 > 0$; for $L \geq 1$ the Gauss–Newton term admits no positive lower bound and the certificate is vacuous by construction.

Theorem 2 (Strong Convexity of the DDDM Reconstruction Energy). *Consider the DDDM update at reverse-time step k and the corresponding proxy reconstruction energy g defined in Eq. 4. Let $F_{\theta,k\eta} : \mathbb{R}^d \to \mathbb{R}^d$ be twice continuously differentiable on a region of interest $\mathcal{Z}$. Define $C(z) := \mathrm{Cov}(x_0 \mid z)$ and $L(z) := \left\|\sigma^{-2} C(z) - I\right\|_2$.*

Suppose there exists $\kappa \in [0, 1)$ such that, uniformly for all $z \in \mathcal{Z}$,

$$(1 - \kappa)\sigma^2 I \preceq C(z) \preceq (1 + \kappa)\sigma^2 I. \tag{6}$$

Then $L(z) \leq \kappa < 1$ and $(1 - L(z))^2 \geq (1 - \kappa)^2$.

Further, suppose that the residual and curvature are uniformly bounded on $\mathcal{Z}$: there exist finite $\bar{R}, \bar{B} \geq 0$ such that $\|r(z)\|_2 \leq \bar{R}$ and $\|H_{F,i}(z)\|_{\mathrm{op}} \leq \bar{B}/\sqrt{d}$ for $i = 1, \ldots, d$,[1] so that $B_{\mathrm{sq}}(z) \leq \bar{B}$ for all $z \in \mathcal{Z}$. If

$$(1 - \kappa)^2 > \bar{R}\,\bar{B}, \tag{7}$$

then g is μ-strongly convex on $\mathcal{Z}$ with

$$\mu = \frac{(1 - \kappa)^2 - \bar{R}\,\bar{B}}{\sigma^2} > 0. \tag{8}$$

Goal 3. Connect these regularity conditions to the underlying data distribution through Tweedie's identities, relating the Jacobian and Hessian of the optimal denoiser to posterior covariance and higher-order posterior moments.

3.3 Tweedie-Motivated Geometry of the DDDM Reconstruction

To characterize the local stability of deterministic DDDM refinement we analyze the geometry of the energy g of Eq. 4, whose curvature depends on the Jacobian and coordinate-wise Hessians of F. We characterize these quantities for the population-optimal denoiser using Tweedie's identity.

[1] By Eq. 11, for the population-optimal denoiser the curvature hypothesis holds whenever the third conditional central-moment tensor satisfies $\left(\sum_i \|\mathcal{T}_i(z)\|_{\mathrm{op}}^2\right)^{1/2} \leq \sigma^4 \bar{B}$ on $\mathcal{Z}$, so $\bar{B}$ is controlled by posterior third-moment concentration.

Under Gaussian corruption $z = x_0 + \sigma\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, let $m(z) := \mathbb{E}[x_0 \mid z]$ and $C(z) := \mathrm{Cov}(x_0 \mid z)$. Tweedie's identity gives

$$F^\star(z) := m(z) - z = \sigma^2 \nabla \log p_\sigma(z), \qquad J_{F^\star}(z) = \frac{1}{\sigma^2} C(z) - I. \tag{9}$$

Consequently,

$$L^\star(z) := \|J_{F^\star}(z)\|_2 = \left\| \frac{1}{\sigma^2} C(z) - I \right\|_2. \tag{10}$$

Differentiation gives

$$H_{F_i^\star}(z) = \frac{1}{\sigma^4} \mathcal{T}_i(z), \qquad B^\star_{\mathrm{sq}}(z) = \frac{1}{\sigma^4} \sum_{i=1}^{d} \|\mathcal{T}_i(z)\|_{\mathrm{op}}, \tag{11}$$

where $\mathcal{T}_i$ is the i-th slice of the third conditional central-moment tensor of $x_0 \mid z$. Thus, for the population-optimal denoiser, the curvature regularity is characterized by posterior covariance, third-order posterior geometry, and reconstruction error $R(z) = \|r(z)\|_2$.

Goal 4. Use the resulting curvature conditions to establish convergence of the DDDM refinement with a constant step size.

4 Experimental Setup

To empirically evaluate the conditions underlying our reconstruction analysis, we perform curvature and reconstruction diagnostics across trained diffusion models for image and molecular generation. For image generation, we consider variance-preserving (VP) models with linear and cosine schedules and a variance-exploding (VE) model, using both MSE and Pseudo-LPIPS training objectives. For molecular generation, we evaluate the VP diffusion model used for 3D molecular generation.

For CIFAR-10, we evaluate six trained checkpoints (VP-linear, VP-cosine, and VE, with two seeds each) at 20 log-spaced timesteps in $[1, 3999]$, using 8 sampled states per timestep: 160 states per checkpoint and 960 in total (the MSE certification set). The two Pseudo-LPIPS checkpoints share their image picks, noise draws, and initialization with their MSE counterparts and differ only in loss fields, so their 320 rows form a paired comparison reported separately and never merged into the certification set. The Pseudo-LPIPS runs use the pseudo-Huber modification with $c = 6.9 \times 10^{-6}$; c enters only the pseudo-Huber losses and its contribution to the loss gradient is below float32 resolution at the logged magnitudes, so it is recorded for reproducibility rather than as a tuned setting.

For each state, we compute the reconstruction residual $r(z) = z - x + F(z)$ and its energy $g(z) = ||r(z)||_2^2/(2\sigma^2)$. The Hessian is given by $\nabla^2 g = \sigma^{-2}[(I + J_F)^\top (I + J_F) + \sum_i r_i H_{F_i}]$. Writing $R = ||r||$ and $u = r/R$, the curvature term satisfies

$$\sum_i r_i H_{F_i} = R H_u, \qquad \left\| \sum_i r_i H_{F_i} \right\|_{\mathrm{op}} = R \|H_u\|_{\mathrm{op}}. \tag{12}$$

This gives an exact evaluation of the curvature contribution without random projections.

We compare three measures of the local strong-convexity certificate: $\mu_{B_s q}$, the paper-comparable bound obtained using $B_{\mathrm{sq}} \geq ||H_u||_{\mathrm{op}}$; μ_{tight}, obtained using the exact curvature quantity $R||H_u||_{\mathrm{op}}$; and the directly computed minimum Hessian eigenvalue $\lambda_{\min}(\nabla^2 g)$. We also report the Jacobian spectral norm $L = \|J_F\|_2$, the residual magnitude R, and $\Delta = \max(0, 1 - L)^2 - R B_{\mathrm{sq}}$, the numerator of the strong-convexity bound $\mu_{\mathrm{bound}} = \Delta/\sigma^2$; all certificates throughout are evaluated in this clamped form. The clamp changes no image-row verdict (18 of 960 rows have $L > 1$, none certified under either convention), but it is not cosmetic: on the molecular states of Appendix B the unclamped form would falsely certify 77 of 160 states and violate the hierarchy of Proposition 3 on 79.

Empirical Analysis of RTK Reconstruction. We evaluate the sufficient condition of **Proposition 2** across the image and molecular settings. Rather than using the worst-case coordinate-wise Hessian

Table 1: Empirical study of the RTK reconstruction iteration on CIFAR-10 and GEOM-LongRange. We report the reconstruction residuals R_1 and R_2 after one and two RTK iterations, respectively, together with the reconstruction change Δ, the Hessian-related quantity B_{sq}, local Jacobian spectral norm L, and stability lower bound μ_{bound} evaluated at the converged reconstruction state ($k = 2$). Lower L and positive μ_{bound} indicate a stable local regime. All models, image and molecular, are trained with MSE.

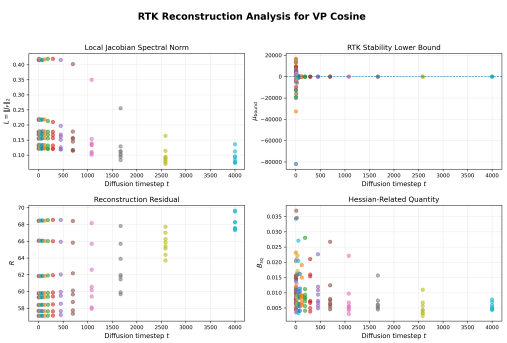
Diffusion	Schedule	Dataset	t	R_1	R_2	B_{sq}	L	μ_{bound}	Δ
VP	Linear	CIFAR-10	3999	67.13	0.001207	0.005146	0.0733	0.001289	0.764
VP	Linear	CIFAR-10	3999	67.84	0.001172	0.004776	0.0875	0.00878	0.2368
VP	Cosine	CIFAR-10	3999	67.53	0.001066	0.013796	0.004397	0.5424	0.991
VP	Cosine	CIFAR-10	3999	67.52	0.127286	0.005703	0.0732	0.002638	0.6783
VP	Cosine	GEOM-LongRange	3999	3.541	24.5652	44.35293	1.3683	-156.929	–
VE	-	CIFAR-10	3999	59.03	0.000844	2.95×10^{-7}	0.000287	1.5993×10^{-7}	–
VE	-	CIFAR-10	3999	59.13	0.00377	4.9315×10^{-7}	0.0002103	1.5993×10^{-7}	–

Table 2: Effect of the training objective on certificate inputs. Pseudo-LPIPS (PL) and MSE checkpoints are evaluated on identical (t, sample) pairs. Values report the relative change from MSE to PL.

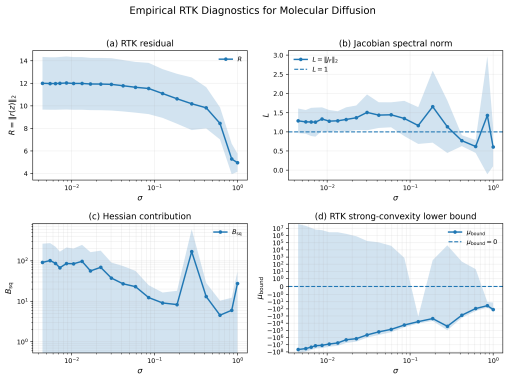
Schedule	ΔL_{est}	$\Delta(R\|H_u\|_{\mathrm{op}})$	$\mu_{B_s q} > 0$	$\mu_{\mathrm{tight}} > 0$	$\lambda_{\min} > 0$
VP-cosine	-78%	-87%	$0\% \to 0\%$	$11.25\% \to 100\%$	$45.6\% \to 100\%$
VE	-39%	-40%	$5\% \to 5\%$	$10\% \to 10\%$	$85\% \to 63\%$

bound $B_{\mathrm{sq}} \leq \bar{B}$ from Theorem 2, we directly measure the aggregate curvature contribution B_{sq} at the converged reconstruction state. At the converged reconstruction state ($k = 2$) the residual is small ($R \sim 10^{-3}$), so $(1 - L)^2 - RB_{\mathrm{sq}}$ is positive and $\mu_{\mathrm{bound}} > 0$: near a solution the certificate is nearly unconditional, since the dangerous term is weighted by R. Certification at the pre-refinement diagnostic states — the population on which stability must be predicted rather than confirmed — is far rarer and is quantified in the next section. The VE rows' μ_{bound} equals $1/\sigma^2$ to the shown precision and reflects the σ-convention; the dimensionless $\mu\sigma^2$ is the comparable quantity across formulations. Moreover, from Figure 1, we observe that the reconstruction converges rapidly, with the stability metrics remaining effectively unchanged after $k = 2$ iterations. For molecular diffusion on GEOM-LongRange (molecules with more than 100 atoms; Appendix I), the sufficient condition derived in Proposition 2 is not satisfied across the evaluated diffusion states. Specifically, $(1-L)^2 - RB_{\mathrm{sq}}$ remains non-positive, resulting in $\mu_{\mathrm{bound}} \leq 0$. This indicates that the local strong-convexity guarantee provided by Proposition 2 does not extend to the molecular setting under the evaluated model and diffusion schedule. From Table 2, relative to MSE, Pseudo-LPIPS reduces the estimate-slot Lipschitz constant by 78% and 39%, and the residual-direction curvature $R\|H_u\|_{\mathrm{op}}$ by 87% and 40%, for the VP-cosine and VE schedules, respectively. Notably, for VP-cosine, the generic B_2-based certificate remains unchanged despite this large reduction in directional curvature, whereas the direction-aware certificate reflects the change. One attribution we do not make: certification *outcomes* at epoch 49 favor whichever arm's checkpoint is less degraded past its guide-loss minimum (2 of 2 arms), and this confound is irreducible within these runs—the degradation ordering between the arms is unchanged at every measured epoch from 20 to 49, and at each run's guide-loss minimum the learned map is constant in the estimate, so every certificate holds trivially in both arms. The early constant-map phase is itself objective-independent: at epoch 20, Pseudo-LPIPS and MSE models alike have $L \approx 0.002$ and certify trivially, with refinement engaging only between epochs 20 and 49. The guide-loss rise that makes the saved epoch never the best one is caused by the curriculum: holding w fixed removes the rise entirely ($1.01\times$ vs. $1.69\times$ the minimum, one arm and seed) and yields epoch-49 models certified far more often ($\mu_{\mathrm{tight}} > 0$ on 83–91% vs. 11%), though the fixed-w models also optimize a different trade-off (their iteration loss is $13\times$ higher), so degradation is not isolated as the cause. Separating the objective's effect on outcomes would require a pair whose degradation ordering differs; the input-level differences (L, the curvature term) are not subject to this confound and remain attributable to the objective.

For the VP-linear schedule the reconstruction converges in two iterations: R falls from 67.33 at $k = 0$ to $\sim 7.5 \times 10^{-4}$ by $k = 2$, and L and μ_{bound} are unchanged thereafter (Table 3, Appendix).



(a) VP-cosine schedule on CIFAR-10.



(b) VP Cosine Schedule on GEOM-LongRange.

Figure 1: Empirical study of the RTK reconstruction. The local Jacobian norm L, Hessian-related quantity B_{sq}, and stability bound μ_{bound} are evaluated across diffusion timesteps.

This is Theorem 1's picture at the measured contraction constants (maximum estimate-slot norm 0.018–0.127 across the six checkpoints at $k \geq 1$); with the box clamp active the residual instead stalls at the clamp gap of Corollary 1 exactly (8.484 vs. 8.484 on a held-out image, 15–17% of pixels clipped), so reported residuals are pre-clamp residuals, matching the $r(z)$ of the analysis.

5 Certificate and the Gap

Proposition 3 (Certificate hierarchy)**.** *For every state with $R > 0$, the three curvature quantities satisfy*

$$\mu_{B_s q} \;\leq\; \mu_{\text{tight}} \;\leq\; \lambda_{\min}\big(\nabla^2 g(z)\big)\,. \tag{13}$$

Consequently, $\mu_{B_s q} > 0$ and $\mu_{\text{tight}} > 0$ each provide sufficient conditions for local strong convexity, with μ_{tight} being the more informative direction-aware certificate. Neither condition is necessary: the true Hessian may remain positive definite even when the corresponding certificate is non-positive.

We relegate the proof to the Appendix F. The ordering in Proposition 3 holds for all 1280 states, validating the numerical estimates. The original $B_s q$-based certificate is highly conservative: it is positive for only 1.7% of states, compared with 29% for the tighter residual-direction certificate and 69% for the directly computed minimum Hessian eigenvalue. These rates change by at most 0.11 percentage points under a numerical dead band of 10^{-3} applied to the σ^2-scaled numerator. The tight certificate is also strongly noise-dependent, being positive for 91.7% of states at $t > 2000$, but only $\sim 20\%$ at $t < 500$.

The two slots are different operators. Over the 960 MSE rows the estimate-slot and noisy-slot Jacobians exceed the contraction threshold on nearly disjoint sets (18 vs. 153 of 960 rows with $\|J\|_2 > 1$), and only the noisy slot approaches the symmetry Tweedie predicts for a covariance-derived operator (mean relative asymmetry 0.164 vs. 0.493). We report no summary correlation between the slots: the association reverses sign under grouping (positive within every checkpoint, negative within timesteps), so any single coefficient is an artifact of the grouping chosen.

To understand the gap between our certificate and the true Hessian curvature, we decompose the two terms in the bound. We find that the dominant source of conservatism is the curvature term $R\|H_u\|_{\text{op}}$, which substantially overestimates the curvature along the residual direction. In contrast, tightening the positive-curvature term $(1 - L)^2$ has a much smaller effect. This is consistent with the observed weak alignment between the relevant curvature and the residual direction. Using the measured alignment as a diagnostic substantially reduces the gap, including recovering all states in the gap population. However, this alignment-based analysis is not a valid certificate, since it does

not control the full negative spectrum of H_u. These results therefore identify directional curvature as the primary source of conservatism in the current certificate.

6 Molecular Refinement

Our molecular experiments show that replacing the scalar refinement head with a 74-dimensional element-wise output removes the deterministic $\sqrt{74}$ amplification in the estimate-slot Jacobian, with $|J_{\text{self}}|_2/|\partial F/\partial \hat{x}|_2 = 1.0000$ across 160 configurations, compared with 8.6023–8.6025 for the scalar head. However, this architectural correction alone does not guarantee contraction: at intermediate checkpoints (epochs 10 and 20) the estimate-slot Jacobian is contractive only trivially—the learned map is still near-constant in the estimate—while at the final checkpoint (epoch 249) it is expansive on every evaluated state, with $\|\partial F/\partial \hat{x}\|_2$ ranging from 1.01 to 10.74 (median 4.45; 0 of 160 states contract). The results therefore indicate that removing the $\sqrt{74}$ amplification eliminates a structural source of instability, while contraction ultimately depends on the learned refinement map and, in the full sampler, on additional conditioning pathways. Further details are given in Appendix B.

7 Conclusion

We gave a convergence theory for DDDM refinement—Banach contraction for the undamped iteration in the slot that governs it, a projected variant matching the clamped sampler, and a constant-step result on the residual energy—and audited our own certificates at scale. The audit returned structure, not just rates: the certificate's conservatism reduces to one measured geometric assumption (worst-case alignment between its two terms, wrong by two to three orders of magnitude, uniformly); its inputs are trainable (an objective change moved a model from 11% to 100% certification, a change only the exact curvature term can see); its resolution—not the sampler's—runs out at low noise; and its contrapositive detected a real architectural defect from the Jacobian alone, yielding a $\sqrt{d}$ design rule for refinement heads. Rather than implying universal one-point reconstruction behavior, these results indicate that the stability of DDDM refinement depends on the data modality, the training objective, and the learned denoising geometry—in ways the theory now measures, explains, and in one case predicts. Open directions follow directly: a certificate engaging the spectral distribution of the curvature term rather than its norm, quality-matched cross-schedule and cross-objective comparisons, and stability diagnostics for a molecular refinement whose sampler operator admits contraction. For molecular generation, future work will extend the present analysis from the explicit estimate-slot contraction to the full coupled sampler, where the current estimate can also influence the conditioning inputs. This will allow us to characterize contraction and stability along all dependency pathways and to investigate how the learned Jacobian evolves during training. We also plan to connect these stability diagnostics with geometric distortion and conformational coverage under aggressive few-step deterministic sampling, providing a more direct assessment of when deterministic acceleration preserves molecular diversity and structural fidelity.

AI Use Disclosure

Large language models (Anthropic's Claude) were used substantially in this work, in two capacities. First, the experimental pipeline was executed agentically via Claude Code under author-written specifications: diagnostic sweeps, reproducibility audits, and analysis scripts were run, and their outputs reported, by the model, with all experimental designs, acceptance criteria, and stop conditions specified by the authors and all reported numbers verified by the authors against the generated artifacts (archived configurations, seeds, and per-run records). Second, portions of the manuscript text—including initial drafts of theorem statements, proofs, and expository sections—were drafted or edited with the model's assistance; all mathematical claims were subsequently checked by the authors, and several model-produced drafts were corrected or rejected in the process. The research questions, the interpretation of results, and responsibility for all content rest with the authors.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2025. URL https://arxiv.org/abs/2303.08797.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom: Energy-annotated molecular conformations for property prediction and molecular generation, 2022. URL https://arxiv.org/abs/2006.05531.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling, 2022. URL https://arxiv.org/abs/2002.00107.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation, 2025a. URL https://arxiv.org/abs/2505.18825.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models, 2025b. URL https://arxiv.org/abs/2406.07507.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling, 2023. URL https://arxiv.org/abs/2106.01357.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions, 2023a. URL https://arxiv.org/abs/2211.01916.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation, 2020. URL https://arxiv.org/abs/2009.00713.

S. Chen, G. Daras, and A. G. Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. *arXiv preprint arXiv:2303.03384*, 2023b. URL https://arxiv.org/abs/2303.03384.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast, 2023c. URL https://arxiv.org/abs/2305.11798.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints, 2024. URL https://arxiv.org/abs/2206.00941.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2015. URL https://arxiv.org/abs/1410.8516.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models, 2025. URL https://arxiv.org/abs/2410.12557.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling, 2025. URL https://arxiv.org/abs/2505.13447.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. URL https://arxiv.org/abs/2204.03458.

Peter Holderrieth and Ezra Erives. An introduction to flow matching and diffusion models, 2026. URL https://arxiv.org/abs/2506.02070.

Peter Holderrieth, Douglas Chen, Luca Eyring, Ishin Shah, Giri Anantharaman, Yutong He, Zeynep Akata, Tommi Jaakkola, Nicholas Matthew Boffi, and Max Simchowitz. Diamond maps: Efficient reward alignment via stochastic flow maps, 2026. URL https://arxiv.org/abs/2602.05993.

Pascal Jutras-Dubé, Patrick Pynadath, and Ruqi Zhang. Single-step consistent diffusion samplers, 2025. URL https://arxiv.org/abs/2502.07579.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion, 2024. URL https://arxiv.org/abs/2310.02279.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8245. doi: 10.1561/2200000056. URL http://dx.doi.org/10.1561/2200000056.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2021. URL https://arxiv.org/abs/2009.09761.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Zhiqi Li, Yuchen Sun, Greg Turk, and Bo Zhu. Functional mean flow in hilbert space, 2025. URL https://arxiv.org/abs/2511.12898.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl, 2025. URL https://arxiv.org/abs/2505.05470.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, 2022. URL https://arxiv.org/abs/2206.00927.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, June 2025. ISSN 2731-5398. doi: 10.1007/s11633-025-1562-4.

Kunwoo Na, Junghyun Lee, Se-Young Yun, and Sungbin Lim. Probability-flow ode in infinite-dimensional function spaces, 2025. URL https://arxiv.org/abs/2503.10219.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech, 2021. URL https://arxiv.org/abs/2105.06337.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.12454.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation, 2025. URL https://arxiv.org/abs/2506.14603.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020a. URL https://arxiv.org/abs/1907.05600.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models, 2020b. URL https://arxiv.org/abs/2006.09011.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023. URL https://arxiv.org/abs/2303.01469.

Dan Zhang, Jingjing Wang, and Feng Luo. Directly denoising diffusion models, 2024. URL https://arxiv.org/abs/2405.13540.

zhangkai wu, Xuhui Fan, Hongyu Wu, and Longbing Cao. SCot: Unifying consistency models and rectified flows via straight-consistent trajectories. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=GV82iAD70j.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models, 2023. URL https://arxiv.org/abs/2302.04867.

A More Related Work

Diffusion Models and PF ODE. Diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2021; Song & Ermon, 2020a), also known as score-based generative models, have achieved state-of-the-art performance across a wide range of generative tasks, including image synthesis (Dhariwal & Nichol, 2021; Ramesh et al., 2022; Saharia et al., 2022), audio generation (Kong et al., 2021; Popov et al., 2021; Chen et al., 2020), and video generation (Ho et al., 2022). However, their iterative denoising procedure requires numerous sequential function evaluations during sampling, making inference significantly slower than single-step generative models like GANs (Goodfellow et al., 2014), VAEs (Kingma & Welling, 2019) and normalizing flows (Dinh et al., 2015).

Despite their remarkable empirical success, the theoretical principles underlying their reverse-time dynamics and sampling behavior remain only partially understood. Recent years have witnessed substantial progress in the theoretical understanding of diffusion models, with convergence analyses established for both Denoising Diffusion Probabilistic Models (DDPMs) (Bortoli et al., 2023; Block

et al., 2022; Chen et al., 2023a) and Denoising Diffusion Implicit Models (DDIMs) (Chen et al., 2023b). Nevertheless, these analyses primarily study sampling dynamics rather than the optimization geometry induced by deterministic denoising mappings.

Fast Deterministic Sampling. A large body of work has sought to accelerate diffusion sampling by replacing stochastic reverse diffusion with deterministic trajectories. DDIM (Song et al., 2022) introduced a non-Markovian deterministic sampling procedure that preserves the training objective while substantially reducing the number of sampling steps. Subsequent ODE-based samplers, including probability flow ODEs (Chen et al., 2023c), DPM-Solver (Lu et al., 2022), DPM-Solver++ (Lu et al., 2025), and UniPC (Zhao et al., 2023), formulate sampling as numerical integration of an ordinary differential equation and employ higher-order solvers to improve sample quality with fewer function evaluations. More recently, consistency models (Song et al., 2023) and consistency trajectory models (Kim et al., 2024) further reduce inference cost by learning mappings that enable one-step or few-step generation. Although these methods significantly accelerate sampling, they primarily focus on improving numerical integration or directly approximating the reverse trajectory rather than reformulating the denoising process itself.

Flow Map-Based Acceleration. Flow maps provide a general framework for modeling the solution operator of ODEs, enabling direct prediction of integrated trajectories and few-step sampling (Kim et al., 2024; Boffi et al., 2025a;b). Recent approaches exploit this structure to reduce sampling costs: shortcut models (Frans et al., 2025) impose a semigroup property to learn flow maps, while MeanFlow (Geng et al., 2025) and Align Your Flow (Sabour et al., 2025) enforce Eulerian consistency conditions. Diamond Maps (Holderrieth et al., 2026) develop stochastic flow maps that amortize multiple generative simulation steps while enabling efficient inference-time reward alignment through consistent value-function estimation. Their focus is on efficient stochastic sampling, guidance, and search. These methods can reduce sampling to fewer than 10 NFEs, complementing other acceleration strategies such as advanced ODE solvers and distillation. However, they either sacrifice exact likelihood computation or still require full trajectory integration for likelihood evaluation.

B Diagnosing Molecular Refinement Instability

Mechanism: a scalar refinement head. The molecular failure above is a property of the refinement head rather than of molecular data. In the evaluated implementation the network emits a single scalar per node, which the sampler broadcast-subtracts across all 74 node features, so the composed refinement map has the rank-one Jacobian $J_{\text{self}} = -\mathbf{1}_{74}\,(\partial F/\partial \hat{x})^{\top}$ and therefore $\|J_{\text{self}}\|_2 = \sqrt{74}\,\|\partial F/\partial \hat{x}\|_2$; the identity matches autograd to a median relative deviation of 2.5×10^{-7} over all 160 evaluated configurations. Contraction thus requires $\|\partial F/\partial \hat{x}\|_2 < 1/\sqrt{74} \approx 0.116$, whereas the measured values lie in $[0.145,\ 7.10]$, so the composed map is expansive at every evaluated state ($\|J_{\text{self}}\|_2 \in [1.24,\ 61.1]$) and the iteration overflows within 12–16 steps. This verdict is a property of the deployed artifact rather than of the weights alone: the checkpoint's stored normalization statistics sit orders of magnitude below their well-determined population values, and recomputing them reverses both the contraction verdict and the divergence. The certificate evaluates the model as it would actually be run, normalization state included; the design rule below is structural and independent of the buffers. The mechanism generalizes to a design rule: any refinement that broadcasts a scalar correction across a d-dimensional state multiplies its contraction constant by $\sqrt{d}$, leaving essentially no contraction budget at $d = 74$.

Removing the broadcast is not sufficient. To test whether an elementwise head restores contraction, we trained a successor model with a width-74 head and an MPNN encoder feeding the GraphGPS layers, and repeated the diagnostic on the same molecules, timestep grid and seed. The amplification is removed exactly: the measured ratio $\|J_{\text{self}}\|_2/\|\partial F/\partial \hat{x}\|_2$ is 1.0000 on all 160 configurations, against 8.6023–8.6025 for the scalar head. Contraction is nonetheless not restored: measured on the raw checkpoints in float64 with canonicalized positional-encoding signs, the raw norm $\|\partial F/\partial \hat{x}\|_2$ at the final epoch is 1.01–10.74 (median 4.45), 0 of 160 states contract, and the sampler diverges from every evaluated state. The verdict is robust to the eigenvector sign convention (median shift 0.2%; no state crosses the threshold under resampled signs). The two models differ in architecture, training objective and budget as well as in head width, so this is not a controlled

ablation of the head; it establishes that the $\sqrt{d}$ amplification is removable and that removing it does not by itself yield a contractive refinement.

At the final epoch the energy nonetheless remains strongly convex on 73% of the evaluated states (117 of 160) while the clamped certificate, zero by construction past $L = 1$, certifies none—the conservatism quantified on images appears here in extreme form, and for the same reason: the binding constraint is the worst-case curvature coupling.

A second feedback path. In the molecular sampler the running estimate re-enters the network twice: as the argument of F, and through the node features, which the sampler overwrites with the broadcast estimate after every refinement step. The Jacobian L entering Proposition 2 sees only the first. The two paths are separable empirically: across five checkpoints of the elementwise-head model there is one, at epoch 20, where $\|\partial F/\partial \hat{x}\|_2$ is numerically zero and the refinement converges in two steps when the node features are held fixed, yet the sampler as implemented still diverges from every evaluated state. A vanishing contraction constant in the estimate slot is therefore not sufficient for stable refinement when the estimate also enters through the conditioning inputs. This is a scope condition on Proposition 2 rather than a counterexample to it, and one that the image models—whose sampler has no such path—do not expose. The obstruction is moreover structural for this sampler: $\partial F/\partial \hat{x}$ is rank-one to machine precision, so the update operator of the sampler as implemented retains 73 unit eigenvalues under every step size and either sign (0 of 640 damped and 0 of 84 sign-reversed runs contract); no tuning of the update can restore contraction while the conditioning overwrite is in place.

Scope: the sampler does not generate geometry. The molecular sampler as evaluated does not write atomic positions: conformers generated with the model and with the RDKit-only baseline are identical (0 Å over $1{,}766$ conformers), and the two evaluation outputs coincide. Reported RMSD/coverage numbers therefore measure the RDKit ETKDG initialization, not the model, and we make no claim about conformer quality in this work; the molecular analysis concerns refinement stability only.

C Proofs of Theorem 1, Corollary 1, and the constant-step result

Theorem 1. $J_f = -\partial F/\partial z$, so for $z, z' \in \mathcal{Z}$ the mean value inequality gives $\|f(z) - f(z')\|_2 \le \gamma\|z - z'\|_2$: f is a γ-contraction of the complete metric space $\mathcal{Z}$ into itself, and Banach's fixed-point theorem yields a unique z^* with $f(z^*) = z^*$, i.e. $r(z^*) = 0$; the geometric rate follows by iterating $\|z^{(m)} - z^*\|_2 \le \gamma\|z^{(m-1)} - z^*\|_2$. □

Corollary 1. Orthogonal projection onto a closed convex set is 1-Lipschitz, so $\Pi \circ f$ is a γ-contraction and Banach applies as above; at the fixed point $z^*_\Pi = \Pi f(z^*_\Pi)$, $r(z^*_\Pi) = z^*_\Pi - f(z^*_\Pi) = \Pi f(z^*_\Pi) - f(z^*_\Pi)$, whose norm is the clamp gap. □

Constant-step refinement. If g is μ-strongly convex and Λ-smooth on a convex invariant region containing its minimizer, then for $\tau \in (0, 1/\Lambda]$ the descent inequality $g(z - \tau\nabla g) \le g(z) - \frac{\tau}{2}\|\nabla g\|^2$ and the Polyak–Łojasiewicz inequality $\|\nabla g\|^2 \ge 2\mu(g - g^*)$ chain to $g(z^{(m+1)}) - g^* \le (1 - \tau\mu)(g(z^{(m)}) - g^*)$. □

D Proof of Proposition 2

Proof. In the subsequent analysis, we omit the subscripts $\theta, k\eta$ for notational simplicity. Let

$$g(z) = \frac{1}{2\sigma^2}\|z - (x - F(z))\|^2 = \frac{1}{2\sigma^2}\sum_{i=1}^{d} r_i(z)^2, \qquad r_i(z) = z_i - x_i + F_i(z).$$

Differentiating gives

$$\partial_j g(z) = \frac{1}{\sigma^2}\sum_{i=1}^{d} r_i(z)\,\partial_j r_i(z), \qquad \nabla g(z) = \frac{1}{\sigma^2} J_r(z)^\top r(z),$$

where $(J_r)_{ij} = \partial_j r_i(z)$. A second differentiation and substitution $\partial_j r_i(z) = \delta_{ij} + \partial_j F_i(z)$, $\partial_k \partial_j r_i(z) = \partial_k \partial_j F_i(z)$ yields the exact Hessian

$$\nabla^2 g(z) = \frac{1}{\sigma^2}\Big((I + J_F(z))^\top (I + J_F(z)) + \sum_{i=1}^{d} r_i(z)\, H_{F,i}(z)\Big). \tag{14}$$

where $J_F(z) = \nabla_z F(z)$ and $H_{F,i}(z) = \nabla^2 F_i(z)$ is the Hessian of the i-th output coordinate.

Consider an arbitrary direction $v \in \mathbb{R}^d$. Then

$$v^\top \nabla^2 g(z)\, v = \frac{1}{\sigma^2}\Big(\|(I + J_F(z))v\|^2 + \sum_{i=1}^{d} r_i(z)\, v^\top H_{F,i}(z)\, v\Big). \tag{15}$$

We bound the two terms separately.

(I) Lower bound for the quadratic term. Using the reverse triangle inequality and the operator norm of J_F,

$$\|(I + J_F)v\| = \|v + J_F v\| \geq \|v\| - \|J_F v\| \geq (1 - \|J_F\|_{\mathrm{op}})\|v\|.$$

Define $L(z) := \|J_F(z)\|_{\mathrm{op}}$; then

$$\|(I + J_F)v\|^2 \geq (1 - L(z))^2 \|v\|^2.$$

(II) Upper bound for the correction (Hessian) term. Set $a_i := v^\top H_{F,i}(z)\, v$ (scalars). By the operator-norm bound on $H_{F,i}$,

$$|a_i| \leq \|H_{F,i}(z)\|_{\mathrm{op}}\, \|v\|^2. \tag{16}$$

Thus

$$\sum_{i=1}^{d} a_i^2 \leq \|v\|^4 \sum_{i=1}^{d} \|H_{F,i}(z)\|_{\mathrm{op}}^2 = B_{\mathrm{sq}}(z)^2\, \|v\|^4. \tag{17}$$

with $B_{\mathrm{sq}}(z) := \big(\sum_{i=1}^{d} \|H_{F,i}(z)\|_{\mathrm{op}}^2\big)^{1/2}$ as in the main text. By Cauchy–Schwarz over the index i,

$$\Big|\sum_{i=1}^{d} r_i(z)\, a_i\Big| \leq \|r(z)\|_2 \sqrt{\sum_{i=1}^{d} a_i^2} \leq R(z)\, B_{\mathrm{sq}}(z)\, \|v\|^2. \tag{18}$$

where $R(z) := \|r(z)\|_2$.

(III) Combine the two bounds. Inserting (I) and (II) into the expression for $v^\top \nabla^2 g(z) v$ gives

$$v^\top \nabla^2 g(z)\, v \geq \frac{1}{\sigma^2}\Big((1 - L(z))^2 - R(z)\, B_{\mathrm{sq}}(z)\Big)\|v\|^2.$$

Therefore if $(1 - L(z))^2 > R(z) B_{\mathrm{sq}}(z)$ we obtain a positive lower bound; defining

$$m := \frac{1}{\sigma^2}\Big((1 - L(z))^2 - R(z) B_{\mathrm{sq}}(z)\Big) > 0. \tag{19}$$

yields $v^\top \nabla^2 g(z) v \geq \mu \|v\|^2$ for all v, i.e. $\nabla^2 g(z) \succeq \mu I_d$ and g is μ-strongly convex at z. Consequently $p(z) \propto e^{-g(z)}$ is μ-strongly log-concave in a neighborhood of z.

Affine special case. If F is affine, $F(z) = \Lambda z + b$ (so $J_F = \Lambda$ and $H_{F,i} = 0$), then

$$\nabla^2 g(z) = \frac{1}{\sigma^2}(I + \Lambda)^\top (I + \Lambda), \qquad v^\top \nabla^2 g(z) v = \frac{1}{\sigma^2}\|(I + \Lambda)v\|^2 \geq 0.$$

Hence g is convex. Moreover:

(a) If $(I + \Lambda)$ is full-rank then $\|(I + \Lambda)v\| > 0$ for all $v \neq 0$, so g is strictly (and hence strongly) convex.

(b) If $(I + \Lambda)$ is not of full rank, then there exists a nonzero vector v such that $(I + \Lambda)v = 0$. Consequently, the Hessian $\nabla^2 g(z)$ admits a zero eigenvalue corresponding to this direction, implying that g is not strictly convex along v.

□

E PROOF OF THEOREM 2

Proof. From the computation in the main text we have the exact Hessian

$$\nabla^2 g(z) = \frac{1}{\sigma^2}(I + J_F(z))^\top (I + J_F(z)) + R(z), \qquad R(z) := \frac{1}{\sigma^2}\sum_{i=1}^{d} r_i(z)\, H_{F,i}(z),$$

where $J_F(z) = \nabla_z F_{\theta,k\eta}(z)$ and $H_{F,i}(z) = \nabla^2 F_i(z)$. The first matrix is a Gram matrix and therefore positive semi-definite:

$$\frac{1}{\sigma^2}(I + J_F)^\top (I + J_F) \succeq 0.$$

By assumption $\|J_F(z)\|_2 \le \kappa < 1$. For any unit vector v the reverse triangle inequality yields

$$\|(I + J_F)v\| = \|v + J_F v\| \ge \big|\|v\| - \|J_F v\|\big| \ge 1 - \|J_F\|_2 \ge 1 - \kappa,$$

hence the smallest singular value of $I + J_F$ satisfies $\sigma_{\min}(I + J_F) \ge 1 - \kappa$.

Using the definition of singular values of a matrix A (spectral inequality from SVD),

$$A^\top A \succeq \sigma_{\min}(A)^2 I.$$

we obtain the spectral inequality

$$(I + J_F)^\top (I + J_F) \succeq (1 - \kappa)^2 I,$$

Finally, scaling by $\frac{1}{\sigma^2}$

$$\frac{1}{\sigma^2}(I + J_F)^\top (I + J_F) \succeq \frac{(1 - \kappa)^2}{\sigma^2} I.$$

We now establish an operator-norm bound for the remainder term $R(z)$. We assume that, over the region of interest, the following uniform bounds hold.

$$\|H_{F,i}(z)\|_{\mathrm{op}} \le B \qquad \text{for } i = 1, \dots, d, \qquad \text{and} \qquad \|r(z)\|_2 \le R_{\max}.$$

Then, using triangle inequality and Cauchy–Schwarz,

$$\Big\| \sum_{i=1}^{d} r_i(z)\, H_{F,i}(z) \Big\|_{\mathrm{op}} \le \sum_{i=1}^{d} |r_i(z)|\, \|H_{F,i}(z)\|_{\mathrm{op}} \le \|r(z)\|_2 \sqrt{\sum_{i=1}^{d} \|H_{F,i}(z)\|_{\mathrm{op}}^2} \le R_{\max}\sqrt{d}\, B.$$

Including the factor $1/\sigma^2$ from the definition of $R(z)$ gives

$$\|R(z)\|_{\mathrm{op}} \le \frac{R_{\max}\sqrt{d}\, B}{\sigma^2}.$$

Hence, we can define

$$C_B \;=\; \frac{R_{\max}\sqrt{d}\, B}{\sigma^2},$$

such that $\|R(z)\|_{\mathrm{op}} \le C_B$ uniformly on the region. Combining the two bounds yields the matrix inequality

$$\nabla^2 g(z) \succeq \left(\frac{(1 - \kappa)^2}{\sigma^2} - C_B \right) I.$$

Therefore, if

$$\frac{(1 - \kappa)^2}{\sigma^2} > C_B,$$

then $\nabla^2 g(z)$ is uniformly positive definite on the region and g is m-strongly convex with modulus

$$\mu \;=\; \frac{(1 - \kappa)^2}{\sigma^2} - C_B \;>\; 0.$$

□

Proof. Let

$$C(z) := \mathrm{Cov}(x_0 \mid z), \qquad L(z) := \left\| \frac{1}{\sigma^2} C(z) - I \right\|_2 .$$

By assumption, there exists $\kappa \in [0, 1)$ such that, uniformly for all $z \in \mathcal{Z}$,

$$(1-\kappa)\sigma^2 I \preceq C(z) \preceq (1+\kappa)\sigma^2 I.$$

Since $C(z)$ is symmetric positive semi-definite, it admits an eigen decomposition

$$C(z) = Q \Lambda Q^\top,$$

where $Q \in \mathbb{R}^{d \times d}$ is orthogonal and

$$\Lambda = \mathrm{diag}\big(\lambda_1(z), \ldots, \lambda_d(z)\big).$$

The assumed spectral bounds imply

$$(1-\kappa)\sigma^2 \le \lambda_i(z) \le (1+\kappa)\sigma^2, \qquad i = 1, \ldots, d.$$

Hence,

$$1 - \kappa \le \frac{\lambda_i(z)}{\sigma^2} \le 1 + \kappa,$$

and therefore

$$\left| \frac{\lambda_i(z)}{\sigma^2} - 1 \right| \le \kappa, \qquad i = 1, \ldots, d.$$

Using the eigen decomposition of $C(z)$,

$$\begin{aligned} \frac{1}{\sigma^2} C(z) - I &= \frac{1}{\sigma^2} Q \Lambda Q^\top - I \\ &= Q \left(\frac{\Lambda}{\sigma^2} - I \right) Q^\top. \end{aligned}$$

Thus, the eigenvalues of $\frac{1}{\sigma^2} C(z) - I$ are

$$\frac{\lambda_i(z)}{\sigma^2} - 1, \qquad i = 1, \ldots, d.$$

Since this matrix is symmetric, its spectral norm is the maximum absolute value of its eigenvalues. Consequently,

$$\begin{aligned} L(z) &= \left\| \frac{1}{\sigma^2} C(z) - I \right\|_2 \\ &= \max_{1 \le i \le d} \left| \frac{\lambda_i(z)}{\sigma^2} - 1 \right| \\ &\le \kappa < 1. \end{aligned}$$

Therefore,

$$1 - L(z) \ge 1 - \kappa > 0,$$

which implies

$$(1 - L(z))^2 \ge (1-\kappa)^2.$$

Dividing by $\sigma^2 > 0$ and multiplying by I gives

$$\boxed{\frac{(1-L(z))^2}{\sigma^2} I \succeq \frac{(1-\kappa)^2}{\sigma^2} I.}$$

Thus, the spectral concentration assumption on $C(z)$ provides the uniform positive-definite lower bound

$$\frac{(1-\kappa)^2}{\sigma^2} I$$

for the Hessian contribution $\frac{(1-L(z))^2}{\sigma^2} I$. □

Table 3: RTK reconstruction values for the VP linear schedule (unclamped map; the undamped iteration of Eq. 3 has no step size). The fixed point is reached by $k = 2$; later iterates agree to within 6×10^{-6} relative.

k	$R\downarrow$	$L\downarrow$	B_{sq}	$\mu_{\mathrm{bound}}\uparrow$
0	67.334	0.1057	0.00373	0.550
1	0.1249	0.0607	0.00248	0.882
2	9.60×10^{-4}	0.0606	0.00251	0.883
5	7.52×10^{-4}	0.0606	0.00294	0.883
10	7.53×10^{-4}	0.0606	0.00277	0.883
20	7.59×10^{-4}	0.0606	0.00267	0.883
50	7.58×10^{-4}	0.0606	0.00333	0.883
100	7.55×10^{-4}	0.0606	0.00282	0.883

F Proof of Proposition 3

Proof. Write

$$A \;=\; I + J_F(z), \qquad R \;=\; \|r(z)\|_2 \;>\; 0, \qquad u \;=\; \frac{r(z)}{R}, \qquad H_u \;=\; \sum_i u_i\, H_{F_i}(z),$$

so that $\sum_i r_i H_{F_i} = R\, H_u$ and $\nabla^2 g(z) = \sigma^{-2}\big[A^\top A + R\, H_u\big]$. Each H_{F_i} is the Hessian of a scalar coordinate function and is therefore symmetric; hence H_u is symmetric, and both certificates are understood in their clamped form,

$$\mu_{B_{\mathrm{sq}}} = \frac{\max(0, 1-L)^2 - R\, B_{\mathrm{sq}}}{\sigma^2}, \qquad \mu_{\mathrm{tight}} = \frac{\max(0, 1-L)^2 - R\, \|H_u\|_2}{\sigma^2}.$$

First inequality. By the triangle inequality and Cauchy–Schwarz over the index i, using $\|u\|_2 = 1$,

$$\|H_u\|_2 \;\le\; \sum_i |u_i|\, \|H_{F_i}\|_{\mathrm{op}} \;\le\; \Big(\sum_i u_i^2\Big)^{1/2} \Big(\sum_i \|H_{F_i}\|_{\mathrm{op}}^2\Big)^{1/2} \;=\; B_{\mathrm{sq}},$$

and the two certificates share the same first term, so $\mu_{B_{\mathrm{sq}}} \le \mu_{\mathrm{tight}}$.

Second inequality. For the Gauss–Newton term, $\lambda_{\min}(A^\top A) = \sigma_{\min}(A)^2$. By Weyl's inequality for singular values, $\sigma_{\min}(I + J_F) \ge \sigma_{\min}(I) - \|J_F\|_2 = 1 - L$, and since singular values are nonnegative this strengthens to

$$\sigma_{\min}(I + J_F) \;\ge\; \max(0,\, 1 - L), \qquad \text{hence} \qquad \lambda_{\min}(A^\top A) \;\ge\; \max(0,\, 1-L)^2.$$

(The unclamped $(1-L)^2$ is *not* a lower bound when $L > 1$: e.g. $J_F = \mathrm{diag}(-1, -3)$ has $L = 3$ and $(1-L)^2 = 4$, while $\sigma_{\min}(I + J_F) = 0$.) For the curvature term, symmetry of H_u gives $\lambda_{\min}(R\, H_u) = R\, \lambda_{\min}(H_u) \ge -R\, \|H_u\|_2$. Since $A^\top A$ and $R\, H_u$ are both symmetric, Weyl's inequality $\lambda_{\min}(X + Y) \ge \lambda_{\min}(X) + \lambda_{\min}(Y)$ yields

$$\lambda_{\min}\big(\nabla^2 g(z)\big) \;=\; \frac{1}{\sigma^2}\, \lambda_{\min}\big(A^\top A + R\, H_u\big) \;\ge\; \frac{\max(0,\, 1-L)^2 - R\, \|H_u\|_2}{\sigma^2} \;=\; \mu_{\mathrm{tight}}.$$

Combining the two inequalities gives $\mu_{B_{\mathrm{sq}}} \le \mu_{\mathrm{tight}} \le \lambda_{\min}(\nabla^2 g(z))$. □

G Certificate Conservatism and Sufficiency

We compare the proposed certificate μ_{tight} with the true minimum Hessian eigenvalue $\lambda_{\min}(\nabla^2 g)$ (Figure 2). Points above the diagonal correspond to conservative certificates. Notably, the square markers identify states for which $\mu_{\mathrm{tight}} \le 0$ while $\lambda_{\min}(\nabla^2 g) > 0$. Thus, although $\mu_{\mathrm{tight}} > 0$ certifies strong convexity, $\mu_{\mathrm{tight}} \le 0$ does not imply loss of positive definiteness. This empirically illustrates that the certificate is sufficient but not necessary. Moreover, these cases occur across

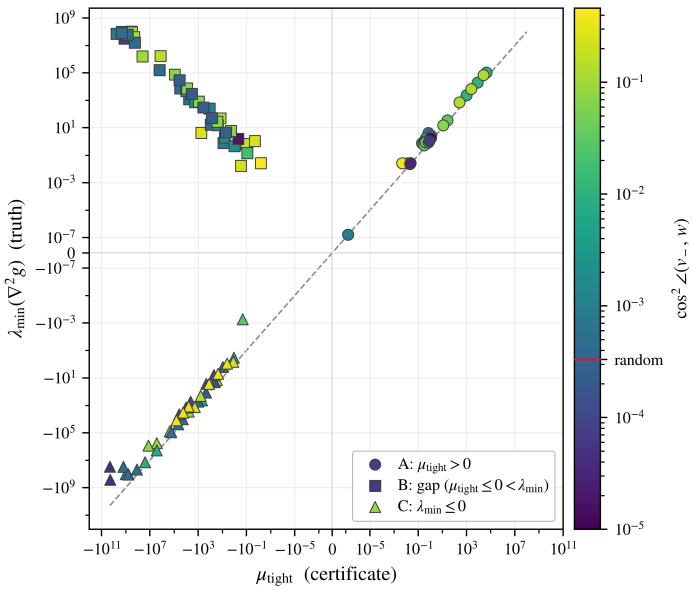


Figure 2: Comparison of the proposed certificate μ_{tight} with the minimum Hessian eigenvalue $\lambda_{\min}(\nabla^2 g)$. Points above the diagonal correspond to conservative certificates. Square markers indicate states with $\mu_{\text{tight}} \le 0$ despite $\lambda_{\min}(\nabla^2 g) > 0$.

Table 4: Directional alignment and decomposition of certificate slack.

	median $\cos^2 \angle(v_-, w)$	curvature share of slack	$\frac{s_{\min}^2}{(1-L)^2}$
A: certified	0.0030	70.4%	1.11
B: gap population	0.0017	88.6%	1.23
C: $\lambda_{\min} \le 0$	0.0024	92.1%	1.31

a range of directional-alignment values, indicating that the conservatism cannot be explained by directional misalignment alone.

Alignment dominates certificate slack. The conservatism of our certificate is primarily due to the worst-case alignment assumed for the second-order curvature term. In $d = 3072$, the measured median $\cos^2 \theta = 1.9 \times 10^{-3}$ is far below the worst-case value 1 (random-direction baseline: 3.3×10^{-4}). The curvature over-charge accounts for 70–92% of the observed slack, whereas replacing $(1-L)^2$ with the exact $s_{\min}^2(I + J_F)$ changes the bound by only 1.1–1.3×. Incorporating the measured alignment certifies all gap states as strongly convex, compared with 0% under the original bound and 22.5% when only the Jacobian bound is tightened. Thus, the certificate is a conservative sufficient condition, with worst-case alignment constituting its dominant source of slack.

H SMALL NOISE REGIME

The sufficient curvature certificate becomes non-informative in the small-noise limit, even though the reconstruction remains convergent. For the one-dimensional two-point distribution, the posterior covariance has the closed form

$$C(z) = a^2 \operatorname{sech}^2\left(\frac{az}{\sigma^2}\right).$$

Away from the symmetry point, the scaled curvature $s = C/\sigma^2$ therefore decays rapidly as $\sigma \to 0$; for example, at $\sigma/a = 0.077$ and $z/a = 1.05$, $s \approx 10^{-152}$. For the optimal denoiser, $L^* = |s - 1|$, so $L^* \to 1$ and the positive term $(1 - L^*)^2$ in our lower bound vanishes. At the same time, the Hessian correction term B^* vanishes at a comparable rate, leaving the certificate governed by the

difference between two vanishing quantities. Consequently, the bound loses a meaningful positive margin and may cease to certify strong convexity as σ decreases. This does not imply failure of the reconstruction: all 41 sampled states still converged, with a median of 7 refinement steps. Thus, in this regime, the experiment indicates a limitation of the certificate's resolution rather than a failure of the underlying reconstruction iteration.

I EXPERIMENTS

For each diffusion timestep t, we estimate the denoiser Jacobian spectral norm $\|J_F(z)\|_2$ and the curvature-dependent quantities in our Hessian bound, using them to obtain an empirical lower bound on the reconstruction objective compare the resulting admissible step-size range with those used in deterministic DDDM sampling. For image generation, we train unconditional ADM U-Nets (Dhariwal & Nichol, 2021; Zhang et al., 2024) on CIFAR-10 (Krizhevsky, 2009), using 128 base channels, channel multipliers $(1, 2, 2, 2)$, three residual blocks per stage, and dropout 0.3. For molecular generation, we use a GraphGPS-inspired architecture (Rampášek et al., 2022) combining local message passing and global attention. Models are trained on approximately 55,000 GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022) molecules with 74-dimensional node and 4-dimensional bond features using an $84\%/10\%/6\%$ train/validation/test split. We additionally evaluate generalization on GEOM-LongRange, constructed from GEOM-MoleculeNet by selecting molecules with more than 100 atoms. The molecular diagnostics are reported for two models that differ in their refinement head: the scalar-head implementation, and an elementwise-head successor trained for 250 epochs whose 74-dimensional raw node features are encoded by an MPNN and projected into the GraphGPS layers.