# The Geometry of Noise: Why Diffusion Models Don't Need Noise Conditioning

**Mojtaba Sahraee-Ardakan**[*] Google

**Mauricio Delbracio** Google

**Peyman Milanfar** Google

## Abstract

Generative models typically rely on explicitly tracking the noise level to guide the sampling process. However, recent autonomous models successfully generate data using a single, time-invariant vector field that operates without explicit noise conditioning. This raises a fundamental paradox: what landscape are these networks actually optimizing when the noise level is treated as a random variable, and how can a bounded model remain stable near the data manifold where gradients typically diverge? We resolve this paradox by showing that autonomous generation is not merely "blind" denoising, but a specific form of Riemannian gradient flow on the Marginal Energy landscape, defined by integrating out the unknown noise levels. Through a novel relative energy decomposition, we demonstrate that while the raw Marginal Energy contains a severe geometric singularity normal to the data manifold, the learned field implicitly incorporates a local conformal metric. This metric perfectly preconditions the singularity, turning an infinitely deep potential well into a stable attractor. Finally, we establish strict structural stability conditions for autonomous sampling. We prove that standard noise-prediction parameterizations can structurally fail due to an inherent high gain amplification of the estimation errors. In contrast, velocity-based models are inherently stable.

## 1 Introduction

Generative modeling has seen immense progress over the past decade, tracing its roots to the foundational non-equilibrium thermodynamics approach introduced by [1]. This paradigm was popularized and scaled by Denoising Diffusion Probabilistic Models (DDPM) [2] and subsequent architectural refinements [3], score-based models [4–6], and subsequently unified under the continuous-time mathematical framework of score-based stochastic differential equations (SDEs)[7]. For a broader conceptual overview of how these perspectives on diffusion have evolved, see [8, 9]. More recently, the field has evolved beyond pure diffusion to embrace velocity-based transport formulations [10–12], notably Flow Matching [13], as well as stationary targets like Equilibrium Matching (EqM)[14]. Concurrently, these velocity-driven frameworks are being extended to respect intrinsic structural constraints, ranging from explicitly curved Riemannian and information-geometric 'spacetime' manifolds [15–17] to complex physical laws like PDEs [18, 19].

A defining characteristic of these standard generative models is their reliance on explicit time-conditioning. They typically learn a conditional score function or velocity field, such as $\epsilon_\theta(\mathbf{u}, t)$, defining a dynamic field that changes with time, i.e., a field over $\mathbb{R}^D \times [0, 1]$. In these frameworks, the network relies on the time variable $t$ to dictate the current scale of corruption and orient the trajectory.

In contrast, recent work has explored autonomous approaches, such as EqM [14] or noise-blind diffusion [20–23], which learn a single noise-agnostic vector field $f_\theta(\mathbf{u})$ independent of $t$. We denote the optimal vector field learned by such a model as $f^*(\mathbf{u})$. It is important to distinguish the

[*]Correspondence to: mojtabaa@google.com

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

model's architecture from the resulting inference dynamics: while the sampling process may still utilize a time-dependent schedule to scale the trajectory, the underlying neural network is strictly time-invariant. This autonomous approach presents a fundamental puzzle: the "correct" gradient to follow from a point $\mathbf{u}$ should depend heavily on its noise level. How can a single, static vector field effectively guide a sample from pure noise (high $t$) and also guide a sample from light noise (low $t$), all while ensuring its stationary points accurately reflect the clean data $\mathcal{X}$?

In this paper, we resolve this paradox. We show that an autonomous model does not simply average out time-dependent errors to act as a "blind" denoiser. Instead, it implicitly learns a hybrid field tied to a single, unified landscape: the marginal energy ($E_{\text{marg}}$). Rather than navigating a sequence of changing terrains dictated by a clock, the model navigates one global topographical map defined as the negative log-likelihood of the marginal data distribution $p(\mathbf{u}) = \int p(\mathbf{u}|t)p(t)dt$:

$$E_{\text{marg}}(\mathbf{u}) = -\log\left(\int p(\mathbf{u}|t)p(t)dt\right). \tag{1}$$

We build our argument as follows:

1. **The Energy Paradox:** We define the marginal energy and derive its explicit gradient. We rigorously show that this gradient diverges near the data manifold, creating an infinitely deep potential well that forbids stable gradient descent.
2. **Energy-Aligned Decomposition:** We analyze the learned autonomous vector fields, proving they decompose into exactly three geometric components: a preconditioned natural gradient, a transport correction (covariance) term, and a linear drift.
3. **Riemannian Gradient Flow:** We resolve the singularity paradox by showing that noise-agnostic models implicitly implement a preconditioned Riemannian gradient flow. The learned vector field incorporates a local scaling factor (the effective gain) that perfectly preconditions and counteracts the geometric singularity of the raw energy landscape.
4. **Stability of Sampling with Autonomous Models:** We establish the mathematical conditions necessary for sampling stability. We prove that velocity-based parameterizations (e.g., Flow Matching, EqM) succeed because they absorb posterior uncertainty into a stable drift, whereas standard noise-prediction parameterizations (e.g., DDPM/DDIM [24]) act as high-gain amplifiers for estimation errors which can lead to structural instability.

## 2 Related Work

Our work unifies and grounds three recent lines of inquiry in generative modeling: noise unconditional generation, equilibrium dynamics, and energy-based training.

**Noise-Blind Denoising.** The immense success of diffusion models has fundamentally recast generative modeling as an iterative denoising process. This deep connection has renewed interest in the classical problem of blind image denoising, where the degradation level is not known a priori and must be inferred directly from the corrupted signal itself [25–31]. Despite this historical precedent in image restoration, the prevailing paradigm in modern score-based generative modeling relies on explicitly conditioning the network on the noise level $t$ [4]. However, Sun et al. [20] recently challenged this, demonstrating that "blind" models can achieve high-fidelity generation without $t$. This connects to earlier findings by Gnanasambandam and Chan [32], who showed that a single "one-size-fits-all" denoiser could approximate an ensemble of noise-specific estimators. Our work provides the rigorous theoretical justification for these observations, identifying $E_{\text{marg}}$ as the implicit objective and connecting the gradients of this energy to the autonomous fields via concentration of the noise level given the noisy signal. This geometric perspective also conceptually grounds recent distance field-inspired approaches, such as Distance Marching, which practically reformulate time-unconditional generation by explicitly training the network to target the closest data points [33].

In a concurrent work, Kadkhodaie et al. [21] prove that for data with low intrinsic dimensionality ($d \ll D$), blind diffusion models can implicitly estimate the noise level from a single observation to robustly track a valid schedule. While their statistical analysis focuses primarily on Langevin-type SDEs within this specific regime, our work situates their concentration of measure as one asymptotic case (Regime I) within a broader geometric framework of Riemannian gradient flow on the marginal energy landscape. By generalizing to arbitrary affine diffusions, we provide a distinct mechanistic

insight: whereas they highlight that blind denoisers avoid schedule mismatch errors, we prove that noise prediction models (e.g., DDPM) are structurally unstable for autonomous generation due to gradient singularities, establishing the necessity of velocity or signal-based targets.

Beyond our primary foundational goal of explaining why empirical noise-agnostic models succeed, removing explicit $t$-conditioning offers concrete practical advantages. In standard time-conditioned models, discrepancies between training and inference trajectories induce noise shift, SNR-$t$ bias, and exposure bias [34–36]; autonomous models naturally mitigate this mismatch by inferring the effective noise scale directly from the spatial geometry of the noisy signal. Furthermore, marginalizing over $t$ acts as an implicit score smoother that aids generalization, and eliminating time-conditioning embeddings simplifies multi-time paradigms such as consistency trajectory models and flow maps by reducing two-time conditioning $(t, s)$ to a single step-size or target parameter.

**Energy Landscapes & Singularities.** While the framework of energy-based learning is well-established [37–46], explicitly learning energy functions is known to be unstable [47]. Recent approaches like "Dual Score Matching" [48] attempt to stabilize this by learning a joint energy via both space and time scores. Our work analyzes the *marginal* energy, aligning with Scarvelis et al. [49], who proved that the exact closed-form score of a finite dataset degenerates into a nearest-neighbor lookup. While they address this by smoothing the score kernel, we show that autonomous flow models resolve it implicitly via a Riemannian preconditioner.

**Equilibrium Dynamics & Flow.** Wang and Du [14] introduced Equilibrium Matching (EqM) to replace time-dependent fields with a single time-invariant gradient. This parallels Action Matching [50]. Our analysis reveals that EqM is unique: it implements a natural gradient descent on the marginal energy. This connects EqM to the fundamental JKO scheme [51], unifying "transport" and "restoration" under a single autonomous field.

## 3 Preliminaries: A Unified Schedule Formulation

To provide a general theory covering Diffusion Models (DDPM, EDM) and Flow Matching (EqM), we adopt the unified affine formulation proposed by Sun et al. [20].

Let $t \in [0, 1]$ index the noise level. The noisy observation $\mathbf{u}_t$ is constructed from clean data $\mathbf{x}$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ via method-specific schedule functions $a(t)$ and $b(t)$:

$$\mathbf{u}_t = a(t)\mathbf{x} + b(t)\boldsymbol{\epsilon}. \tag{2}$$

Generative models are typically trained to predict a linear target $r(\mathbf{x}, \boldsymbol{\epsilon}, t) = c(t)\mathbf{x} + d(t)\boldsymbol{\epsilon}$ by minimizing the Mean Squared Error (MSE):

$$\mathcal{L}(f) = \mathop{\mathbb{E}}_{\mathbf{x},\boldsymbol{\epsilon},t} \left[ \|f(\mathbf{u}_t) - (c(t)\mathbf{x} + d(t)\boldsymbol{\epsilon})\|^2 \right]. \tag{3}$$

In the standard diffusion paradigm, $f$ is explicitly conditioned on the noise level $t$, i.e. it has the form[2] $f_t(\mathbf{u})$. The minimizer of the MSE loss is then the conditional expectation of the target:

$$f_t^*(\mathbf{u}) = \mathop{\mathbb{E}}_{\mathbf{x},\boldsymbol{\epsilon}|\mathbf{u},t} [c(t)\mathbf{x} + d(t)\boldsymbol{\epsilon}]. \tag{4}$$

This function defines a time-dependent vector field that guides the generation process.

In this work, however, we focus on **autonomous models** where the network $f(\mathbf{u})$ receives only the noisy observation $\mathbf{u}$, with no access to $t$. The minimizer of the MSE loss for such a "noise-agnostic" model is the *posterior expectation* of the target (Lemma 1 in Appendix A.1):

$$f^*(\mathbf{u}) = \mathop{\mathbb{E}}_{t|\mathbf{u}} \left[ \mathop{\mathbb{E}}_{\mathbf{x},\boldsymbol{\epsilon}|\mathbf{u},t} [c(t)\mathbf{x} + d(t)\boldsymbol{\epsilon}] \right] = \mathop{\mathbb{E}}_{t|\mathbf{u}} [f_t^*(\mathbf{u})]. \tag{5}$$

This has a very intuitive interpretation: *the optimal autonomous model is a time-average of the optimal conditional model with respect to the posterior* $p(t|\mathbf{u})$. By defining the optimal conditional denoiser as $D_t^*(\mathbf{u}) = \mathbb{E}[\mathbf{x}|\mathbf{u}, t]$, we can expand this target as (Lemma 2 in Appendix A.1):

$$f^*(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}} \left[ \frac{d(t)}{b(t)}\mathbf{u} + \left( c(t) - \frac{d(t)a(t)}{b(t)} \right) D_t^*(\mathbf{u}) \right]. \tag{6}$$

[2]While a more accurate notation would have been $f(\mathbf{u}, t)$, with some abuse of notation we use $f_t(\mathbf{u})$ to denote a function of $\mathbf{u}$ and $t$ to keep the notation consistent with the autonomous models.

Specific choices for the coefficients $a(t), b(t), c(t), d(t)$ yield standard architectures, as summarized in Table 1. The DDPM formulation relies on a variance-preserving schedule where $a(t) = \sqrt{\bar{\alpha}_t}$ is defined by the cumulative signal retention of its underlying discrete Markov process [2, 52].

*The central question of this paper is to understand the geometric and dynamical consequences of replacing the precise conditional field $f_t^*(\mathbf{u})$ with this autonomous posterior average $f^*(\mathbf{u})$. Does this time-invariant field still define a valid generative trajectory?*

Table 1: Unified coefficients and simplification of the general autonomous field for common generative models. For the schedule parameters, $\bar{\alpha}_t$ denotes the cumulative signal retention for DDPM [2], and $\sigma_t$ represents the continuous noise standard deviation for EDM [53].

| **Model** | $a(t)$ | $b(t)$ | $c(t)$ | $d(t)$ | **Autonomous field** $f^*(\mathbf{u}) = \mathbb{E}_{t\|\mathbf{u}}\left[f_t^*(\mathbf{u})\right]$ |
|---|---|---|---|---|---|
| **DDPM** [2] | $\sqrt{\bar{\alpha}_t}$ | $\sqrt{1-\bar{\alpha}_t}$ | 0 | 1 | $\mathbb{E}_{t\|\mathbf{u}}\left[\frac{\mathbf{u}-\sqrt{\bar{\alpha}_t}D_t^*(\mathbf{u})}{\sqrt{1-\bar{\alpha}_t}}\right] = \mathbb{E}_{t\|\mathbf{u}}[\epsilon_t^*(\mathbf{u})]$ |
| **EDM** [53] | 1 | $\sigma_t$ | 1 | 0 | $\mathbb{E}_{t\|\mathbf{u}}\left[D_t^*(\mathbf{u})\right]$ |
| **FM** [13] | $1-t$ | $t$ | $-1$ | 1 | $\mathbb{E}_{t\|\mathbf{u}}\left[\frac{\mathbf{u}-D_t^*(\mathbf{u})}{t}\right]$ |
| **EqM** [14] | $1-t$ | $t$ | $-t$ | $t$ | $\mathbb{E}_{t\|\mathbf{u}}\left[\mathbf{u}-D_t^*(\mathbf{u})\right]$ |

## 4 The Geometry of the Marginal Energy

Standard diffusion models rely on a time-dependent score function, $\nabla_{\mathbf{u}} \log p(\mathbf{u}|t)$, which explicitly guides the trajectory at every noise level. In contrast, autonomous models use a static vector field $f^*(\mathbf{u})$ that is independent of time. This fundamental difference raises a critical geometric question: *Does this static field align with the gradient of a global potential energy?* If such a potential exists, **autonomous generation** could be theoretically grounded as a form of energy minimization [37]. The most natural candidate for this potential is the *marginal energy* $E_{\text{marg}}(\mathbf{u})$, defined as the negative log-likelihood of the marginal data density $p(\mathbf{u}) = \int p(\mathbf{u}|t)p(t)dt$:

$$E_{\text{marg}}(\mathbf{u}) = -\log p(\mathbf{u}). \tag{7}$$

To determine if the learned field $f^*(\mathbf{u})$ aligns with this energy, Lemma 3 proves its gradient is the posterior expectation of the conditional scores:

$$\nabla_{\mathbf{u}} E_{\text{marg}}(\mathbf{u}) = \mathop{\mathbb{E}}_{t|\mathbf{u}}[-\nabla_{\mathbf{u}} \log p(\mathbf{u}|t)] = \mathop{\mathbb{E}}_{t|\mathbf{u}}\left[\frac{\mathbf{u} - a(t)D_t^*(\mathbf{u})}{b(t)^2}\right]. \tag{8}$$

where the second equality uses the Tweedie's formula [54, 55] (see A.3). This result establishes the link between the static geometry and the dynamic denoising process. However, it also exposes a critical flaw in the landscape itself.

### 4.1 The Energy Paradox

A key requirement for generative modeling is that the learned vector field $f^*(\mathbf{u})$ must be consistent with the clean data support at the boundary ($t \to 0$). Depending on the formulation, the resulting probability flow ODE governing the sampling dynamics (defined by the velocity $\frac{d\mathbf{u}}{dt}$) generally behaves in one of two ways once $f^*(\mathbf{u})$ is plugged into the system: an **attractor** (e.g., EqM), where the ODE velocity vanishes at the clean data ($\frac{d\mathbf{u}}{dt} = \mathbf{0}$) to create a stable fixed point, or a **transversal flow** (e.g., Flow Matching), where the velocity does not converge to zero ($\frac{d\mathbf{u}}{dt} \neq \mathbf{0}$) but ensures the trajectory intersects the data manifold at the correct time.

**The Singularity.** Regardless of this boundary behavior, the model faces a *geometric singularity*. As established in Appendix D, the posterior $p(t|\mathbf{u})$ collapses near the data manifold. The term inside the expectation in Eq. (8) becomes singular as $t \to 0$. Consequently, the marginal energy forms an infinitely deep potential well ($E_{\text{marg}} \to -\infty$, see Figure 1), causing the associated gradient field to diverge. For the case of discrete points $\{\mathbf{x}_i\}_{i=1}^N$, (e.g. the training dataset) we have

$$\lim_{\mathbf{u}\to\mathbf{x}_i} \|\nabla_{\mathbf{u}} E_{\text{marg}}(\mathbf{u})\| = \infty. \tag{9}$$

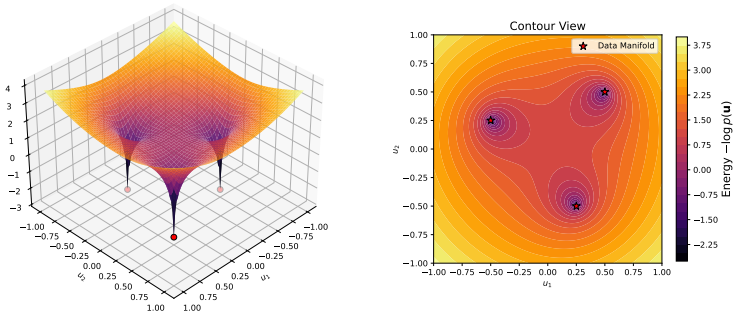



Figure 1: **The Singular Geometry of the Marginal Energy Landscape.** (Left) $E_{marg}(u) = -\log p(u)$ forms an infinitely deep potential well at the data manifold. (Right) Contour view of energy concentration around discrete data points. Though the raw gradient $\nabla_u E_{marg}(u)$ diverges near data, autonomous models maintain stability via implicit Riemannian gradient flow.

This creates a puzzle: how can a neural network learn a bounded vector field (which must be finite at the data) that aligns with a geometry defined by such a singular potential? One might argue that in practice, training is stabilized by truncating the noise level at some $t_{min} > 0$. However, this does not resolve the geometric paradox; it merely converts a mathematical singularity into an extremely stiff optimization landscape where Hessian eigenvalues scale as $1/t_{min}^2$. As such, the raw energy landscape essentially forbids stable gradient descent.

**The role of the noise prior $p(t)$.** The marginal energy landscape $E_{\text{marg}}(\mathbf{u})$ is directly sculpted by the training schedule's noise prior $p(t)$. Schedules emphasizing high noise smooth the landscape to enable global transport, whereas favoring low noise stiffens it, pushing $1/t^p$ singularities outward. In particular, as we show in detail in Appendix D, whereas a non-vanishing prior ($p(0) > 0$) yields an Inverse-Gamma posterior that collapses to a Dirac delta as $\mathbf{u}$ approaches the data manifold, an exponentially vanishing prior such as the EDM log-normal schedule [53] ($\ln b \sim \mathcal{N}(\mu, \sigma^2)$, or equivalently a logit-normal prior on $t \in (0, 1)$) decays faster than any polynomial as $b \to 0$. As proven in Appendix D.4, this super-polynomial decay absorbs the $\mathcal{O}(b^{-k})$ likelihood singularity (where $k = D - d$ is the manifold codimension, or $k = D$ for discrete points), shifting the limiting posterior to a non-degenerate log-normal distribution $\mathcal{LN}(\mu - k\sigma^2, \sigma^2)$. This regularizes $E_{\text{marg}}$ on the data manifold while maintaining a non-zero posterior variance that further reinforces the structural instability of autonomous noise prediction as we show in the later sections. Thus, the schedule serves as a fundamental geometric deformation of the target potential rather than a simple sampling heuristic.

**Memorization vs. Generalization.** Beyond optimization instability, this geometric singularity raises critical questions about generalization. While the raw marginal energy landscape $E_{\text{marg}}$ features infinitely deep potential wells at the data manifold, autonomous models avoid simply memorizing these points. Although the exact closed-form score of a finite dataset degenerates into a memorizing nearest-neighbor lookup [49], finite network capacity and training inductive biases prevent the learning of these singular fields. Because smoothing the exact score function drives generalization [49, 56–58], omitting noise conditioning acts as a powerful implicit regularizer. Removing $t$ forces the network to marginalize representations across all noise scales, effectively smoothing score estimates to enable the synthesis of novel samples.

In the next section, we address the singularity issue. We resolve this paradox by demonstrating that noise-blind models do not follow the raw energy gradient, but rather a *Riemannian gradient flow* that perfectly preconditions this geometric singularity.

## 5 Autonomous Generation as Riemannian Gradient Flow

We resolve the paradox of stable convergence despite divergent gradients by proving autonomous models implement a Riemannian gradient flow. The learned field $f^*(\mathbf{u})$ mirrors the marginal energy's natural gradient, plus a critical correction term dominating only when geometric concentration fails.

### 5.1 The Anatomy of the Autonomous Field

Recall from Lemma 2 that the optimal autonomous field is given by:

$$f^*(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}\left[\frac{d(t)}{b(t)}\mathbf{u} + \frac{c(t)b(t) - d(t)a(t)}{b(t)} D_t^*(\mathbf{u})\right]. \tag{10}$$

This reveals that the autonomous vector field is driven by two competing forces:

1. **The Repulsive Expansion** ($\frac{d(t)}{b(t)}\mathbf{u}$): A linear term scaling with the noise geometry. Assuming standard diffusion signs ($d, b > 0$), this pushes trajectories outward.
2. **The Relative Restoration**: A denoising term pulling the sample toward the data manifold.

**The Energy-Aligned Decomposition.** To understand how these forces align with the geometry of the marginal energy $E_{\text{marg}}$, we use the properties of covariance to Eq. (10) to show that the vector field decomposes into exactly three geometric components (see Appendix B):

$$f^*(\mathbf{u}) = \underbrace{\overline{\lambda}(\mathbf{u})\nabla E_{\text{marg}}(\mathbf{u})}_{\text{Natural Gradient}} + \underbrace{\mathbb{E}_{t|\mathbf{u}}\left[\left(\lambda(t) - \overline{\lambda}(\mathbf{u})\right)\left(\nabla E_t(\mathbf{u}) - \nabla E_{\text{marg}}(\mathbf{u})\right)\right]}_{\text{Transport Correction}} + \underbrace{\overline{c}(\mathbf{u})\mathbf{u}}_{\text{Linear Drift}}, \tag{11}$$

where $\overline{c}(\mathbf{u})$ and $\lambda(t)$ are scalar valued functions, and $\overline{\lambda}(\mathbf{u})$ is its posterior average defined as

$$\overline{c}(\mathbf{u}) \triangleq \mathbb{E}_{t|\mathbf{u}}[c(t)/a(t)], \quad \lambda(t) \triangleq \frac{b(t)}{a(t)}\left(d(t)a(t) - c(t)b(t)\right), \quad \overline{\lambda}(\mathbf{u}) \triangleq \mathbb{E}_{t|\mathbf{u}}[\lambda(t)]. \tag{12}$$

This decomposition (Eq. 11) resolves the paradox by isolating the model's behavior into three interpretable geometric terms: a preconditioned natural gradient flow [59, 60] (via the effective gain $\lambda$), a Transport Correction (covariance) term, and a linear drift. Importantly, this formulation exposes the mechanism of stability. While both the marginal energy gradient $\nabla E_{\text{marg}}$ and the conditional energy gradient $\nabla E_t$ become singular near the manifold, as shown in detail in Appendix E, the effective gain $\lambda(t)$ acts as a *singularity-canceling preconditioner*, vanishing at a rate that exactly counteracts this divergence. Note that $\lambda(t)$ is not strictly positive across all parameterizations (for instance, $\lambda(t) = -b(t)^2 < 0$ for signal prediction in EDM); here its primary geometric role in $f^*(\mathbf{u})$ is singularity cancellation rather than defining a positive-definite metric, and its sign is subsequently inverted by the sampler gain $\nu(t) < 0$ during ODE integration to restore descent toward the data manifold.

The Transport Correction is a covariance bounded by the posterior variance of $p(t|\mathbf{u}) \propto p(\mathbf{u}|t)p(t)$, where a rapidly decaying prior $p(t)$ regularizes the likelihood's concentration as $t \to 0$ to tune the model's reliance on global transport. We analyze this term across two key asymptotic regimes. First, under global high-dimensional concentration (Sec. 5.2), the posterior $p(t|\mathbf{u})$ fully concentrates and the correction vanishes, yielding a pure preconditioned natural gradient flow. Conversely, under local near-manifold proximity (Sec. 5.3), raw energy gradients diverge even as posterior uncertainty shrinks. Depending on the formulation, the Transport Correction can either converge to a $\mathcal{O}(1)$ constant vector or vanish.

### 5.2 Regime I: Global Concentration in High Dimensions

A central mystery of autonomous models is how a single vector field can "know" which noise level to apply to a given input without explicit conditioning. We resolve this by observing that in high-dimensional spaces ($D \gg 1$), provided the data resides on a low-dimensional manifold ($d \ll D$), the noise level $t$ is not truly hidden; it is globally encoded in the geometry of the observation $\mathbf{u}$.

In high dimensions, the mass of a Gaussian distribution concentrates in a thin spherical shell [61, 62]. For low-dimensional data, the noisy observation $\mathbf{u}$ decomposes into a data-subspace component and an orthogonal noise component. Because the codimension is large, this orthogonal noise dominates the total norm. Consequently, the "shells" corresponding to different noise levels $b(t)$ become effectively disjoint. We prove this in Appendix C.

As a result, the input $\mathbf{u}$ becomes a deterministic proxy for the noise level $t$. This geometric structure has two profound consequences for the model's objective:

- **Posterior Concentration:** The model's uncertainty about the noise level vanishes. The posterior $p(t|\mathbf{u})$ collapses to a Dirac delta centered at an implicit estimate $\hat{t}(\mathbf{u})$.
- **Vanishing Transport Correction:** Because there is no longer a mixture of conflicting noise levels at any given point, the complex interaction between different potential fields disappears and the transport correction term vanishes.

Consequently, in this regime, the field is strictly dominated by the **Natural Gradient** flow:

$$f^*(\mathbf{u}) \approx \overline{\lambda}(\mathbf{u})\nabla E_{\text{marg}}(\mathbf{u}) + \overline{c}(\mathbf{u})\mathbf{u}. \tag{13}$$

This resolves the "blindness" paradox globally: the model implicitly sees $t$ through the separation of noise scales.

### 5.3 Regime II: Local Stability via Proximity

While high-dimensional concentration provides a global mechanism, a second, stronger mechanism ensures stability as the trajectory approaches the data, *regardless of dimension*. We analyze the decomposition in the near-manifold limit:

**1. The Near-Manifold Regime (Concentration via Proximity)** As the observation $\mathbf{u}$ approaches the data support ($\mathbf{u} \to \mathcal{X}$), the likelihood becomes dominated by the smallest noise scales. This causes the posterior $p(t|\mathbf{u})$ to concentrate sharply on $t \to 0$ simply because the observation is indistinguishable from clean data. Under standard regularity assumptions i.e. continuous schedules with $a(0) = 1$ and $b(t)$ strictly increasing from $b(0) = 0$, a continuous prior satisfying $p(0) > 0$, and codimension greater than two, we rigorously prove this in Appendix D for two cases. In the first case, we assume that the data is discrete and finite in ambient dimension $D > 2$, and show that as we approach a data point, the posterior $p(t|\mathbf{u})$ converges weakly to the Dirac measure $\delta(t)$. In the second case, we assume that the data lies on a smooth manifold of dimension $d$ with continuous bounded density $p_{\text{data}}$ in an ambient space of dimension $D$ where $k = D - d > 2$. Note that we do not require $D \gg d$ as in Section 5.2. In this case, as for discrete data, $p(t|\mathbf{u})$ converges weakly to $\delta(t)$ as we approach the data manifold. Notably, this *local concentration* occurs even in low dimensions.

However, this posterior concentration does not necessarily mean the Transport Correction term vanishes. As the trajectory approaches the data, the raw energy gradients ($\nabla E_{\text{marg}}$ and $\nabla E_t$) simultaneously diverge. This competing dynamic—shrinking uncertainty but exploding gradients—creates an indeterminate limit and a potential geometric singularity.

The vector field remains stable because the network implicitly implements a Riemannian preconditioning. The effective gain vanishes at a rate that dampens the gradient divergence. Consequently, the Transport Correction term collapses into a bounded, well-behaved vector. Depending on the model parameterization, this resolves in one of two ways: in **EDM** and **EqM** the transport correction term **vanishes**, whereas for **FM** and **DDPM** it converges to a non-zero $\mathcal{O}(1)$ constant vector (pointing normal to the manifold for DDPM, and comprising both normal and positional components for FM).

This preconditioning ensures structural stability near the manifold without explicit noise conditioning. Detailed proofs of these asymptotic behaviors are provided in Appendix D.3.

**2. The High-Noise Regime** Far from the data manifold, if the dimension $D$ is not sufficiently large to enforce global concentration (Sec. 5.2), the strict proximity cue is absent. Here, the covariance term becomes significant, "steering" the trajectories away from the raw energy gradient. This correction ensures the field satisfies the global transport requirements of the noise schedule before the local geometry takes over.

While we showed that the target field is bounded and very close to the optimal conditional field in certain regimes, the dynamics used to generate samples are sensitive to the integrator's step size and coefficients. A bounded target divided by a vanishing noise scale creates a stiff differential equation resulting in instabilities. We discuss this in the next section.

## 6 Stability Conditions for Sampling with Autonomous Models

While Section 5 established that the optimal autonomous target $f^*(\mathbf{u})$ is geometrically well-behaved and acts as an accurate proxy for the conditional field in regimes of concentration (high dimensions

Table 2: Summary of Stability Analysis for Autonomous Models. See Appendix F for details.

| Parameterization | Error Mechanism | Stability |
|---|---|---|
| Noise ($\boldsymbol{\epsilon}$) | amplified Jensen gap | **Conditionally unstable** |
| Signal ($\mathbf{x}$) | polynomial divergence vs. exponential decay | **Stable** |
| Velocity ($\mathbf{v}$) | bounded drift | **Inherently Stable** |

or near-manifold), this does not guarantee stable generation. The *dynamics* of the sampling process can amplify small errors into divergent trajectories. To quantify this, we analyze the sampling process as the integration of a time-dependent velocity field $\mathbf{v}(\mathbf{u}, t)$:

$$\frac{d\mathbf{u}}{dt} = \mathbf{v}_{\text{aut}}(\mathbf{u}, t) \triangleq \mu(t)\mathbf{u} + \nu(t) f^*(\mathbf{u}). \tag{14}$$

Here, $\mu(t)$ is the drift coefficient of the noise schedule and $\nu(t)$ is the **effective gain** of the parameterization (derived in Appendix F). Note that even though the autonomous model $f^*(\mathbf{u})$ is time-independent, the sampler velocity $\mathbf{v}_{\text{aut}}(\mathbf{u}, t)$ remains a function of time because the schedule coefficients $\mu(t)$ and $\nu(t)$ vary during integration.

We compare this autonomous velocity against an ideal "Oracle" sampler that has access to the noise level $t$. The structural stability is determined by the *Drift Perturbation Error* $\Delta\mathbf{v}$, which measures the deviation caused by substituting the conditional target with the autonomous approximation:

$$\mathbf{v}_{\text{orc}}(\mathbf{u}, t) = \mu(t)\mathbf{u} + \nu(t) f_t^*(\mathbf{u}), \qquad \mathbf{v}_{\text{aut}}(\mathbf{u}, t) = \mu(t)\mathbf{u} + \nu(t) f^*(\mathbf{u}).$$

Subtracting the two eliminates the linear term, isolating the error introduced by the target parameterization:

$$\Delta\mathbf{v}(\mathbf{u}, t) \triangleq \|\mathbf{v}_{aut}(\mathbf{u}, t) - \mathbf{v}_{orc}(\mathbf{u}, t)\| = \underbrace{|\nu(t)|}_{\text{Gain}} \cdot \underbrace{\|f^*(\mathbf{u}) - f_t^*(\mathbf{u})\|}_{\text{Estimation Error}}. \tag{15}$$

This decomposition reveals that stability is a race condition as $t \to 0$: while the posterior $p(\tau|\mathbf{u})$ concentrates and the distance to the manifold shrinks as $\mathcal{O}(b(t))$, singular target scalings ($1/b(\tau)$) or diverging sampler gains $\nu(t)$ can amplify residual uncertainty. We analyze this competition across three standard parameterizations: noise, signal, and velocity prediction.

Table 2 summarizes these regimes, with complete mathematical derivations and qualitative breakdowns provided in Appendix F. Specifically, our continuous-time analysis (Appendix F) shows that near the manifold ($t \to 0$), the noise-prediction drift error scales with the schedule derivative, $\Delta\mathbf{v}_{\text{noise}} \propto \mathcal{O}(\dot{b}(t))$, explaining why square-root schedules ($b(t) \propto \sqrt{t}$, as in VP-SDE/DDPM) suffer an $\mathcal{O}(t^{-1/2})$ boundary singularity whereas linear schedules such as flow matching ($b(t) \propto t$) remain locally bounded as $t \to 0$. Note also that while time-conditioned networks can algebraically convert between parameterizations via $t$-dependent skip connections (e.g., EDM preconditioning [53]), an autonomous network $f_\theta(\mathbf{u})$ lacks access to $t$ and cannot internally re-scale a noise prediction into a velocity or signal target without incurring posterior estimation error. It is important to note that this analysis identifies *sufficient conditions for instability*. A divergence ($\Delta\mathbf{v} \to \infty$) guarantees failure, whereas a bounded error is a necessary (but not strictly sufficient) condition for high-fidelity generation. Our results prove that noise prediction is structurally less stable for autonomous models due to error amplification.

## 7 Empirical Verification

To validate the stability conditions from Section 6, we use a toy dataset to demonstrate that high-dimensional embedding enables autonomous models to implicitly learn noise levels. We then evaluate U-Nets on CIFAR-10 and ImageNet-64 to test whether the predicted structural instability of autonomous noise prediction manifests in standard benchmarks, and if velocity or signal parameterizations resolve it. Full training details are provided in Appendix G.

### 7.1 The Impact of Dimensionality on Autonomous Generation

To empirically illustrate how high-dimensional geometry resolves the ambiguity of autonomous generation, we designed a controlled toy experiment motivated by the setup in [63]. We constructed a

2D concentric circles dataset and embedded it into a high-dimensional ambient space $\mathbb{R}^D$ using a random orthogonal projection matrix $P \in \mathbb{R}^{D\times 2}$, where $P^T P = I$. We trained a standard residual network for both Flow Matching and DDPM under conditional and autonomous ("blind") settings; to ensure the simple MLP/ResNet architecture can read out spatial shell radii without access to $t$, the network is provided with the input's Euclidean norm $\|\mathbf{u}\|$ as a purely input-derived geometric feature (Appendix G).

Figure 2 visualizes 2D-projected samples across ambient dimensions $D \in \{2, 8, 128\}$, serving as a diagnostic validation of our three theoretical geometric regimes. First, in the **low-dimensional ambiguity regime** ($D = 2$), overlapping noise shells create an ambiguous posterior $p(t|\mathbf{u})$, yielding diffuse samples for all blind models. Second, in the **parameterization stability regime** ($D = 8$), partially separated shells provide noisy implicit cues: autonomous FM leverages its bounded velocity to resolve the concentric ring structure significantly more clearly, whereas DDPM Blind suffers severe radial dispersion as its singular gain amplifies residual estimation and learning errors. Finally, in the **concentration regime** ($D = 128$), $p(t|\mathbf{u})$ concentrates sharply near a Dirac delta; because estimation error vanishes faster than DDPM's gain diverges, even structurally sensitive parameterizations converge.

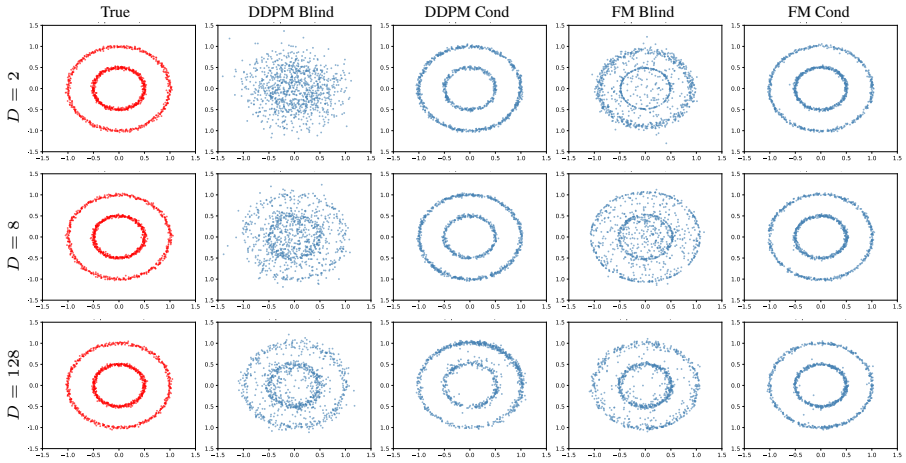



Figure 2: **2D concentric circles embedded in $\mathbb{R}^D$.** At $D = 2$, overlapping noise breaks autonomous models. At $D = 8$, bounded velocity allows autonomous Flow Matching to resolve the structures more clearly, while autonomous DDPM still exhibits severe artifacts. At $D = 128$, absolute geometric concentration eliminates error, allowing all models to converge.

### 7.2 Experiments with more realistic datasets

We verify our stability analysis against the benchmark results of Sun et al. [20] on CIFAR-10 and ImageNet-64. Using the Hackable Diffusion library [64], we trained six distinct U-Net models predicting either the flow, the signal, or the noise, with time conditioning enabled or disabled. To isolate parameterization as the sole variable, we apply a flow-matching noising process and compute the flow loss across all configurations (inspired by [63]), altering only the network's prediction target.

Figure 3 shows unguided (no CFG [65]) samples and their corresponding FID metrics for models trained on each dataset. As predicted in Section 6, the *noise prediction, time conditioning off* model fails to generate coherent samples. The resulting images are dominated by high-frequency artifacts and residual noise. In contrast, models predicting signal or flow with time conditioning off generate samples similar in quality and in metrics to their counterparts with time conditioning turned on.

## 8 Conclusion

We identified the marginal energy as the implicit objective of autonomous generative models and proved its landscape contains a fundamental gradient singularity at the data manifold. We demonstrated these models effectively implement a preconditioned Riemannian gradient flow, where the

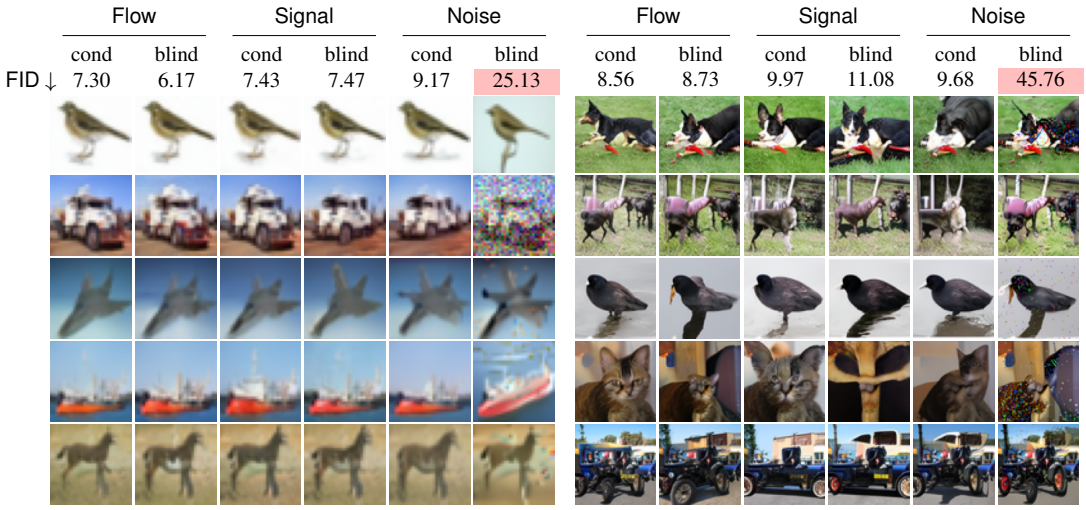



Figure 3: **CIFAR-10 (left) and ImageNet-64 (right) performance.** Columns show FID scores and generated samples for *flow*, *signal*, and *noise* prediction under Flow Matching (with vs. without time conditioning). Blind noise prediction fails consistently, causing severe artifacts.

posterior effective gain acts as a local scaling metric that counteracts the singular energy gradient. Finally, we derived the bounded vector field and sampler stability conditions, proving velocity- and signal-based parameterizations mathematically suffice to realize stable flow without explicit noise conditioning. By shifting generation from time-dependent score matching to time-invariant energy alignment, our work provides a rigorous geometric foundation for future autonomous and equilibrium-based models.

**Scope, Limitations, and Future Work.** Our theoretical analysis characterizes the exact geometry and sampling stability of ideal MMSE estimators $f^*(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}[f_t^*(\mathbf{u})]$; extending this framework to quantify finite-sample neural network optimization dynamics and approximation errors during training is an important direction for future work. Likewise, while our stability analysis focuses on deterministic ODEs to isolate intrinsic drift singularities, extending the analysis to stochastic samplers (SDEs) is a compelling avenue, as injected Langevin noise may smooth the marginal energy landscape and help regularize autonomous trajectories. Finally, from an application perspective, autonomous models are naturally suited to high-dimensional regimes where the noise posterior concentrates sharply such as image restoration, super-resolution, and high-resolution generation whereas highly multimodal conditional tasks such as open-domain text-to-image synthesis may exhibit greater posterior ambiguity in the high-noise regime when distinct conditional manifolds intersect.

## References

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2015.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 2021.

[4] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

[5] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, 2020.

[6] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[7] 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.

[8] Sander Dieleman. Perspectives on diffusion. *Sander Dieleman's Blog*, 2023. URL `https://sander.ai/2023/07/20/perspectives.html`. Published: July 20, 2023.

[9] Ruiqi Gao, Emiel Hoogeboom, Jonathan Heek, Valentin De Bortoli, Kevin Patrick Murphy, and Tim Salimans. Diffusion models and gaussian flow matching: Two sides of the same coin. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL `https://openreview.net/forum?id=C8Yyg9wy0s`.

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

[11] Mauricio Delbracio and Peyman Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=VmyFF5lL3F`. Featured Certification, Outstanding Certification.

[12] Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209): 1–80, 2025.

[13] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[14] Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.

[15] Andrey Kharitenko, Zebang Shen, Riccardo de Santi, Niao He, and Florian Doerfler. Landing with the score: Riemannian optimization through denoising. *arXiv preprint arXiv:2509.23357*, 2025.

[16] Chaoran Cheng, Yusong Wang, Yuxin Chen, Xiangxin Zhou, Nanning Zheng, and Ge Liu. Riemannian consistency model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[17] Rafał Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas Garg. The spacetime of diffusion models: An information geometry perspective. In *The Fourteenth International Conference on Learning Representations*, 2026.

[18] Giacomo Baldan, Qiang Liu, Alberto Guardone, and Nils Thuerey. Flow matching meets PDEs: A unified framework for physics-constrained generation. *arXiv preprint arXiv:2506.08604*, 2025.

[19] Yi Zhang and Difan Zou. Physics-informed distillation of diffusion models for PDE-constrained generation. *arXiv preprint arXiv:2505.22391*, 2025.

[20] Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models? In *Forty-second International Conference on Machine Learning*, 2025.

[21] Zahra Kadkhodaie, Aram-Alexandre Pooladian, Sinho Chewi, and Eero Simoncelli. Blind denoising diffusion models and the blessings of dimensionality. *arXiv preprint arXiv:2602.09639*, 2026.

[22] Jipeng Li and Yanning Shen. Is noise conditioning necessary? a unified theory of unconditional graph diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[23] Zunzhe Zhang, Runhan Huang, Yicheng Liu, Shaoting Zhu, Linzhan Mou, and Hang Zhao. Generative control as optimization: Time unconditional flow matching for adaptive and robust robotic control. *arXiv preprint arXiv:2603.17834*, 2026.

[24] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[25] Peyman Milanfar and Mauricio Delbracio. Denoising: a powerful building block for imaging, inverse problems and machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 383(2299), 2025.

[26] Guangyong Chen, Fengyuan Zhu, and Pheng Ann Heng. An efficient statistical method for image noise level estimation. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[27] Stanislav Pyatykh, Jürgen Hesser, and Lei Zheng. Image noise level estimation by principal component analysis. *IEEE transactions on image processing*, 22(2):687–699, 2012.

[28] Tamer Rabie. Robust estimation approach for blind denoising. *IEEE transactions on image processing*, 14(11):1755–1765, 2005.

[29] Marc Lebrun, Miguel Colom, and Jean-Michel Morel. The noise clinic: a blind image denoising algorithm. *Image Processing On Line*, 5:1–54, 2015.

[30] Marc Lebrun, Miguel Colom, and Jean-Michel Morel. Multiscale image blind denoising. *IEEE Transactions on Image Processing*, 24(10):3149–3161, 2015.

[31] Jae Woong Soh and Nam Ik Cho. Deep universal blind image denoising. In *2020 25th International conference on pattern recognition (ICPR)*, 2021.

[32] Abhiram Gnanasambandam and Stanley H Chan. One size fits all: Can we train one denoiser for all noise levels? *International Conference on Machine Learning*, 2020.

[33] Zimo Wang, Ishit Mehta, Haolin Lu, Chung-En Sun, Ge Yan, Tsui-Wei Weng, and Tzu-Mao Li. Distance marching for generative modeling. *arXiv preprint arXiv:2602.02928*, 2026.

[34] Jincheng Zhong, Boyuan Jiang, Xin Tao, Pengfei Wan, Kun Gai, and Mingsheng Long. Mitigating noise shift in denoising generative models with noise awareness guidance. In *International Conference on Learning Representations*, 2026.

[35] Meng Yu, Lei Sun, Jianhao Zeng, Xiangxiang Chu, and Kun Zhan. Elucidating the SNR-t bias of diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[36] Hui Li, Jiayue Lyu, Fu-Yun Wang, Kaihui Cheng, Siyu Zhu, and Jingdong Wang. MixFlow training: Alleviating exposure bias with slowed interpolation mixture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[37] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[38] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

[39] Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[40] Tim Salimans and Jonathan Ho. Should EBMs model the energy or the score? In *Energy based models workshop, ICLR 2021*, 2021.

[41] Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

[42] Jiali Cui and Tian Han. Learning latent space hierarchical EBM diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[43] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. VAEBM: A symbiosis between variational autoencoders and energy-based models. In *International Conference on Learning Representations*, 2021.

[44] Jinhui Hou, Zhiyu Zhu, and Junhui Hou. Energy-oriented diffusion bridge for image restoration with foundational diffusion models. *arXiv preprint arXiv:2604.10983*, 2026.

[45] Junn Yong Loo, Leong Fang Yu, Michelle Adeline, Julia K Lau, Hwa Hui Tew, Arghya Pal, Vishnu Monn Baskaran, Chee-Ming Ting, and Raphael CW Phan. Learning energy-based generative models via potential flow: A variational principle approach to probability density homotopy matching. *Transactions on Machine Learning Research*, 2025.

[46] Adrian Robert Minut, Hazem Dewidar, and Iacopo Masi. Spilled energy in large language models. *arXiv preprint arXiv:2602.18671*, 2026.

[47] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In *International Conference on Machine Learning*, 2021.

[48] Florentin Guth, Zahra Kadkhodaie, and Eero P Simoncelli. Learning normalized image densities via dual score matching. *Advances in Neural Information Processing Systems*, 2025.

[49] Christopher Scarvelis, Haitz Sáez de Ocáriz Borde, and Justin Solomon. Closed-form diffusion models. *Transactions on Machine Learning Research*, 2025.

[50] Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *International Conference on Machine Learning*, 2023.

[51] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[52] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 2021.

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

[54] Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 388–394. Springer, 1992.

[55] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[56] Kulin Shah, Alkis Kalavasis, Adam Klivans, and Giannis Daras. Does generation require memorization? creative diffusion models using ambient diffusion. In *Forty-second International Conference on Machine Learning*, 2025.

[57] Xinyu Zhou, Jiawei Zhang, and Stephen J Wright. Smoothing the score function for generalization in diffusion models: An optimization-based explanation framework. *arXiv preprint arXiv:2601.19285*, 2026.

[58] Jacob Bamberger, Iolo Jones, Dennis Duncan, Michael M. Bronstein, Pierre Vandergheynst, and Adam Gosztolai. Carré du champ flow matching: better quality-generalisation tradeoff in generative models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[59] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2): 251–276, 1998.

[60] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

[61] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

[62] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[63] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

[64] Google LLC. Hackable diffusion. `https://github.com/google/hackable_diffusion`, 2025.

[65] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

# A General Derivations for Autonomous Models

## A.1 Derivation of the Optimal Autonomous Target

**Lemma 1** (Optimal Autonomous Target)**.** *Consider the loss functional $\mathcal{L}(f)$ defined in Eq* (3)*. The unique global minimizer $f^*(\mathbf{u})$ is given by the expectation of the target conditioned on the noise level $t$, weighted by the posterior $p(t|\mathbf{u})$:*

$$f^*(\mathbf{u}) = \mathop{\mathbb{E}}_{t|\mathbf{u}}\left[\mathop{\mathbb{E}}_{\mathbf{x},\boldsymbol{\epsilon}|\mathbf{u},t}[c(t)\mathbf{x} + d(t)\boldsymbol{\epsilon}]\right]. \tag{16}$$

*Proof.* This is an application of the Law of Iterated Expectations. □

**Lemma 2** (Denoiser Formulation)**.** *The optimal autonomous target $f^*(\mathbf{u})$ can be expressed as an affine transformation of the optimal conditional denoiser $D_t^*(\mathbf{u}) = \mathbb{E}[\mathbf{x}|\mathbf{u}, t]$:*

$$f^*(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}\left[\frac{d(t)}{b(t)}\mathbf{u} + \left(c(t) - \frac{d(t)a(t)}{b(t)}\right) D_t^*(\mathbf{u})\right]. \tag{17}$$

*Proof.* Recall the unified forward process $\mathbf{u} = a(t)\mathbf{x} + b(t)\boldsymbol{\epsilon}$. For a fixed observation $\mathbf{u}$ and noise level $t$, the noise $\boldsymbol{\epsilon}$ is deterministically related to the clean data $\mathbf{x}$ by:

$$\boldsymbol{\epsilon} = \frac{\mathbf{u} - a(t)\mathbf{x}}{b(t)}. \tag{18}$$

Substitute this into the inner expectation of Lemma 1. Note that conditioned on $\mathbf{u}$ and $t$, the terms $\mathbf{u}$, $a(t)$, and $b(t)$ are constants, leaving $\mathbf{x}$ as the only random variable:

$$\mathop{\mathbb{E}}_{\mathbf{x},\boldsymbol{\epsilon}|\mathbf{u},t}[c(t)\mathbf{x} + d(t)\boldsymbol{\epsilon}] = \mathop{\mathbb{E}}_{\mathbf{x}|\mathbf{u},t}\left[c(t)\mathbf{x} + d(t)\left(\frac{\mathbf{u} - a(t)\mathbf{x}}{b(t)}\right)\right] \tag{19}$$

$$= c(t)\,\mathbb{E}[\mathbf{x}|\mathbf{u}, t] + \frac{d(t)}{b(t)}\mathbf{u} - \frac{d(t)a(t)}{b(t)}\,\mathbb{E}[\mathbf{x}|\mathbf{u}, t]. \tag{20}$$

Identifying $D_t^*(\mathbf{u}) = \mathbb{E}[\mathbf{x}|\mathbf{u}, t]$ and grouping the coefficients for $\mathbf{u}$ and $D_t^*(\mathbf{u})$ yields the result. □

## A.2 Gradient of the Marginal Energy

**Lemma 3** (Gradient of the Marginal Energy)**.** *Let the marginal likelihood be the mixture $p(\mathbf{u}) = \int p(\mathbf{u}|t)p(t)dt$ and the marginal energy be $E_{marg}(\mathbf{u}) = -\log p(\mathbf{u})$. The gradient of the marginal energy is the posterior expectation of the conditional energy gradients:*

$$\nabla_{\mathbf{u}} E_{marg}(\mathbf{u}) = \mathop{\mathbb{E}}_{t|\mathbf{u}}\left[-\nabla_{\mathbf{u}} \log p(\mathbf{u}|t)\right]. \tag{21}$$

*Proof.* By definition, $\nabla_{\mathbf{u}} E_{\text{marg}}(\mathbf{u}) = -\frac{\nabla_{\mathbf{u}} p(\mathbf{u})}{p(\mathbf{u})}$. We differentiate the mixture integral under the sign:

$$\nabla_{\mathbf{u}} p(\mathbf{u}) = \nabla_{\mathbf{u}} \int p(\mathbf{u}|t)p(t)dt = \int \nabla_{\mathbf{u}} p(\mathbf{u}|t)p(t)dt. \tag{22}$$

We use the log-derivative trick $\nabla_{\mathbf{u}} p(\mathbf{u}|t) = p(\mathbf{u}|t)\nabla_{\mathbf{u}} \log p(\mathbf{u}|t)$ to rewrite the integrand:

$$\nabla_{\mathbf{u}} p(\mathbf{u}) = \int p(\mathbf{u}|t)\nabla_{\mathbf{u}} \log p(\mathbf{u}|t)p(t)dt. \tag{23}$$

Dividing by $p(\mathbf{u})$ allows us to identify the posterior density $p(t|\mathbf{u}) = \frac{p(\mathbf{u}|t)p(t)}{p(\mathbf{u})}$:

$$\nabla_{\mathbf{u}} E_{\text{marg}}(\mathbf{u}) = -\int \frac{p(\mathbf{u}|t)p(t)}{p(\mathbf{u})}\nabla_{\mathbf{u}} \log p(\mathbf{u}|t)dt \tag{24}$$

$$= \int p(t|\mathbf{u})\left[-\nabla_{\mathbf{u}} \log p(\mathbf{u}|t)\right] dt. \tag{25}$$

□

### A.3 Evaluating the Conditional Score via Tweedie's Formula

To evaluate the expectation derived in Lemma 3, we must compute the conditional score $\nabla_{\mathbf{u}} \log p(\mathbf{u}|t)$. For the affine forward process where the observation model is $p(\mathbf{u}|\mathbf{x}, t) = \mathcal{N}(\mathbf{u}; a(t)\mathbf{x}, b(t)^2\mathbf{I})$, we can express this score directly in terms of the optimal denoiser using Tweedie's formula [54, 55].

Tweedie's formula relates the posterior expectation of the clean data to the score of the marginal likelihood of the corrupted observation. Specifically, the optimal minimum mean squared error (MMSE) denoiser, $D_t^*(\mathbf{u}) = \mathbb{E}[\mathbf{x}|\mathbf{u}, t]$, is given by:

$$D_t^*(\mathbf{u}) = \frac{\mathbf{u} + b(t)^2 \nabla_{\mathbf{u}} \log p(\mathbf{u}|t)}{a(t)}. \tag{26}$$

By rearranging this algebraic relation, we isolate the conditional score. This provides the exact formulation used to construct the marginal energy gradients:

$$\nabla_{\mathbf{u}} \log p(\mathbf{u}|t) = \frac{a(t) D_t^*(\mathbf{u}) - \mathbf{u}}{b(t)^2}. \tag{27}$$

### A.4 Exact Analytical Forms for Autonomous Fields

To facilitate the exact stability verification in Section 6 and Appendix F, we derive the closed-form expressions for the optimal autonomous vector fields. While the neural network approximates these expectations, we can compute them exactly when the data distribution $p_{data}(\mathbf{x})$ is known.

**General Formulation.** Recall from Lemma 2 that the optimal autonomous field is an affine transformation of the posterior expectation of the conditional denoiser:

$$f^*(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}} \left[ \frac{d(t)}{b(t)} \mathbf{u} + \left( c(t) - \frac{d(t)a(t)}{b(t)} \right) D_t^*(\mathbf{u}) \right]. \tag{28}$$

The two key components required to evaluate this are:

1. **The Optimal Conditional Denoiser** $D_t^*(\mathbf{u}) = \mathbb{E}[\mathbf{x}|\mathbf{u}, t]$. By Bayes' rule, this is the center of mass of the posterior $p(\mathbf{x}|\mathbf{u}, t) \propto p(\mathbf{u}|\mathbf{x}, t) p_{data}(\mathbf{x})$.
2. **The Posterior Noise Distribution** $p(t|\mathbf{u})$. This allows us to average the conditional vector field over all possible noise levels.

**Specialization to Discrete Data.** Let the data manifold be a discrete set $\mathcal{X} = \{\mathbf{x}_k\}_{k=1}^N$ with uniform prior $p(\mathbf{x}_k) = 1/N$.

1. **Conditional Denoiser:** The likelihood of observing $\mathbf{u}$ given a specific source $\mathbf{x}_k$ is Gaussian: $p(\mathbf{u}|\mathbf{x}_k, t) = \mathcal{N}(\mathbf{u}; a(t)\mathbf{x}_k, b(t)^2 I)$. The posterior probability $w_k(\mathbf{u}, t) \triangleq p(\mathbf{x}_k|\mathbf{u}, t)$ is given by the softmax of the negative log-likelihoods:

$$w_k(\mathbf{u}, t) = \frac{\exp\left( -\frac{\|\mathbf{u} - a(t)\mathbf{x}_k\|^2}{2b(t)^2} \right)}{\sum_{j=1}^N \exp\left( -\frac{\|\mathbf{u} - a(t)\mathbf{x}_j\|^2}{2b(t)^2} \right)}. \tag{29}$$

The optimal denoiser is the precision-weighted barycenter of the dataset:

$$D_t^*(\mathbf{u}) = \sum_{k=1}^N w_k(\mathbf{u}, t) \mathbf{x}_k. \tag{30}$$

2. **Posterior Noise Distribution:** To compute the outer expectation $\mathbb{E}_{t|\mathbf{u}}[\cdot]$, we require $p(t|\mathbf{u})$. The marginal likelihood $p(\mathbf{u}|t)$ is a Gaussian Mixture Model (GMM) with centers at $a(t)\mathbf{x}_k$. Assuming a uniform prior on time $p(t) \sim \mathcal{U}(0, 1)$, Bayes' rule yields:

$$p(t|\mathbf{u}) = \frac{p(\mathbf{u}|t)}{\int_0^1 p(\mathbf{u}|\tau) d\tau} \propto \frac{1}{N} \sum_{k=1}^N \mathcal{N}(\mathbf{u}; a(t)\mathbf{x}_k, b(t)^2 I). \tag{31}$$

Therefore, for the case of discrete data prior, we can evaluate the unconditional field $f^*(\mathbf{u})$ by numerically integrating Eq. (30) against this posterior $p(t|\mathbf{u})$ using high-precision quadrature on a dense grid $t \in [\epsilon, 1]$. This eliminates sampling variance and allows us to isolate the geometric stability properties of the parameterization.

3. **Closed-Form Simplification:** While numerical integration is straightforward, the discrete data assumption allows for a profound analytical simplification of the autonomous field. By substituting the explicit forms of the conditional denoiser and the posterior noise distribution into the expectation $f^*(\mathbf{u}) = \int_0^1 f_t^*(\mathbf{u}) p(t|\mathbf{u}) dt$, a structural cancellation occurs.

The softmax denominator of the denoiser weights $w_k(\mathbf{u}, t)$ exactly matches the Gaussian mixture summation in the numerator of the posterior $p(t|\mathbf{u}) \propto \sum_{i=1}^N p(\mathbf{u}|\mathbf{x}_i, t)$. Multiplying these terms inside the integral perfectly cancels the nested summations, uncoupling the conditional targets from the marginal evidence.

Factoring out the shared terms and exchanging the sum and integral yields a generalized closed-form expression for the autonomous field:

$$f^*(\mathbf{u}) = \frac{\sum_{k=1}^N \int_0^1 b(t)^{-D} \exp\left(-\frac{\|\mathbf{u}-a(t)\mathbf{x}_k\|^2}{2b(t)^2}\right) \left[\frac{d(t)}{b(t)}\mathbf{u} + \left(c(t) - \frac{d(t)a(t)}{b(t)}\right)\mathbf{x}_k\right] dt}{\sum_{j=1}^N \int_0^1 b(\tau)^{-D} \exp\left(-\frac{\|\mathbf{u}-a(\tau)\mathbf{x}_j\|^2}{2b(\tau)^2}\right) d\tau}. \tag{32}$$

This result demonstrates that for any affine schedule, the autonomous vector field bypasses intermediate softmax bottlenecks. Instead, it evaluates to a globally weighted sum of the individual conditional targets pointing toward each discrete data point.

**Remark on non-uniform priors $p(t)$.** While we assume a uniform prior $p(t) \sim \mathcal{U}(0, 1)$ in this derivation to achieve exact cancellation and analytical clarity, practical autonomous models often utilize heavily skewed distributions (e.g., the log-normal distribution in EDM [53]). Introducing a non-uniform $p(t)$ re-weights the time integrals in Eq. (32), effectively shifting the barycenter of the unconditional field. This theoretical framework provides a direct mechanism to understand how empirical schedule tuning alters the implicit data targets: shifting $p(t)$ towards higher noise levels upweights the long-range transport components of the nested integrals, smoothing the resulting vector field $f^*(\mathbf{u})$.

## B Derivation of General Energy-Aligned Decomposition

In this section, we derive the exact relationship between the learned autonomous vector field $f^*(\mathbf{u})$ and the gradient of the marginal energy $\nabla E_{\text{marg}}(\mathbf{u})$ for general affine noise schedules.

Recall the general autonomous form derived in Eq. (6):

$$f^*(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}\left[\frac{d(t)}{b(t)}\mathbf{u} + \left(c(t) - \frac{d(t)a(t)}{b(t)}\right) D_t^*(\mathbf{u})\right] \tag{33}$$

We defined the conditional energy as

$$E_t(\mathbf{u}) = -\log p(\mathbf{u}|t). \tag{34}$$

To relate this to the energy landscape, we substitute the Tweedie's formula (26), which expresses the optimal denoiser in terms of the conditional energy gradient $\nabla E_t(\mathbf{u})$:

$$D_t^*(\mathbf{u}) = \frac{\mathbf{u} - b(t)^2 \nabla E_t(\mathbf{u})}{a(t)} \tag{35}$$

Substituting this into the expression for $f^*(\mathbf{u})$, we obtain:

$$f^*(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}\left[\frac{d(t)}{b(t)}\mathbf{u} + \left(c(t) - \frac{d(t)a(t)}{b(t)}\right)\left(\frac{\mathbf{u} - b(t)^2 \nabla E_t(\mathbf{u})}{a(t)}\right)\right] \tag{36}$$

We simplify the coefficients for the linear term $\mathbf{u}$ and the gradient term $\nabla E_t(\mathbf{u})$ separately.

**1. The Linear Term:** The coefficient for $\mathbf{u}$ is:

$$\frac{d(t)}{b(t)} + \frac{1}{a(t)}\left(c(t) - \frac{d(t)a(t)}{b(t)}\right) = \frac{d(t)}{b(t)} + \frac{c(t)}{a(t)} - \frac{d(t)}{b(t)} = \frac{c(t)}{a(t)} \tag{37}$$

Thus, the linear component of the field is simply the posterior expectation of the target scale ratio:

$$f^*_{linear}(\mathbf{u}) = \mathbf{u} \cdot \mathbb{E}_{t|\mathbf{u}}\left[\frac{c(t)}{a(t)}\right] \tag{38}$$

**2. The Gradient Term:** The coefficient for $\nabla E_t(\mathbf{u})$ is:

$$-\frac{b(t)^2}{a(t)}\left(c(t) - \frac{d(t)a(t)}{b(t)}\right) = -\frac{b(t)^2 c(t)}{a(t)} + b(t)d(t) = \frac{b(t)}{a(t)}\left(d(t)a(t) - c(t)b(t)\right) \tag{39}$$

Let us define this effective gradient gain as $\lambda(t)$:

$$\lambda(t) \triangleq \frac{b(t)}{a(t)}\left(d(t)a(t) - c(t)b(t)\right) \tag{40}$$

The autonomous field can now be written as:

$$f^*(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}\left[\lambda(t)\nabla E_t(\mathbf{u})\right] + \mathbf{u} \cdot \mathbb{E}_{t|\mathbf{u}}\left[\frac{c(t)}{a(t)}\right] \tag{41}$$

Finally, we apply the covariance decomposition to the expectation term. Recall that the marginal energy gradient is the average of the conditional gradients: $\nabla E_{\text{marg}}(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}[\nabla E_t(\mathbf{u})]$. Using the identity $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + Cov(X, Y)$, we derive the **General Energy-Aligned Decomposition**:

$$f^*(\mathbf{u}) = \underbrace{\overline{\lambda}(\mathbf{u})\nabla E_{\text{marg}}(\mathbf{u})}_{\text{Natural Gradient}} + \underbrace{Cov(\lambda(t), \nabla E_t(\mathbf{u}))}_{\text{Transport Correction}} + \underbrace{\overline{c}(\mathbf{u})\mathbf{u}}_{\text{Linear Drift}} \tag{42}$$

where $\overline{\lambda}(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}[\lambda(t)]$ is the posterior effective gain, and $\overline{c}(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}[c(t)/a(t)]$ is the mean linear drift coefficient. This result proves that for any affine schedule, the learned field is a sum of the marginal energy gradient (scaled by posterior uncertainty), a covariance correction term, and a linear drift.

## C Posterior Concentration in High Dimensions

In this section, we prove that in high dimensions, the posterior distribution of the noise level concentrates sharply, effectively allowing noise-blind models to recover the time signal from the spatial geometry of the observation $\mathbf{u}$.

**Proposition 1.** *Let $\mathbb{V} \subset \mathbb{R}^D$ be a linear subspace of dimension $d < D$. Let $\mathbf{u} = \mathbf{x} + \mathbf{n}$ be an observation where $\mathbf{x} \in \mathbb{V}$ and $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 I_D)$. Assuming an improper flat prior $p(\mathbf{x}) \propto 1$ on $\mathbb{V}$, the posterior distribution $p(\sigma|\mathbf{u})$ concentrates at $\hat{\sigma} = \frac{r}{\sqrt{D-d}}$ with variance $\mathcal{O}(D^{-1})$, where $r = \min_{\mathbf{y}\in\mathbb{V}} \|\mathbf{u} - \mathbf{y}\|$.*

*Proof.* Let $\mathbf{x}^* = \text{proj}_{\mathbb{V}}(\mathbf{u})$ be the orthogonal projection of $\mathbf{u}$ onto $\mathbb{V}$. For any $\mathbf{x} \in \mathbb{V}$, the Pythagorean theorem yields the exact decomposition:

$$\|\mathbf{u} - \mathbf{x}\|^2 = \|\mathbf{u} - \mathbf{x}^*\|^2 + \|\mathbf{x}^* - \mathbf{x}\|^2 = r^2 + \|\mathbf{x}^* - \mathbf{x}\|^2$$

The marginal likelihood (or integrated likelihood) is defined as:

$$\begin{aligned} p(\mathbf{u}|\sigma) &= \int_{\mathbb{V}} p(\mathbf{u}|\mathbf{x}, \sigma) p(\mathbf{x})\, d\mathbf{x} \\ &= \frac{1}{(2\pi\sigma^2)^{D/2}} \int_{\mathbb{V}} \exp\left(-\frac{r^2 + \|\mathbf{x}^* - \mathbf{x}\|^2}{2\sigma^2}\right) d\mathbf{x} \end{aligned}$$

By defining an isometric isomorphism between $\mathbb{V}$ and $\mathbb{R}^d$, we evaluate the integral over the subspace:

$$\begin{aligned} p(\mathbf{u}|\sigma) &= \frac{\exp(-r^2/2\sigma^2)}{(2\pi\sigma^2)^{D/2}} \int_{\mathbb{R}^d} \exp\left(-\frac{\|\mathbf{z}\|^2}{2\sigma^2}\right) d\mathbf{z} \\ &= \frac{\exp(-r^2/2\sigma^2)}{(2\pi\sigma^2)^{D/2}} (2\pi\sigma^2)^{d/2} \\ &= (2\pi\sigma^2)^{-\frac{D-d}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \end{aligned}$$

The log-marginal likelihood $\ell(\sigma) = \ln p(\mathbf{u}|\sigma)$ is:

$$\ell(\sigma) = -(D-d)\ln\sigma - \frac{r^2}{2\sigma^2} + \text{const}$$

Solving $\ell'(\sigma) = 0$ gives the maximum likelihood estimator:

$$-\frac{D-d}{\sigma} + \frac{r^2}{\sigma^3} = 0 \implies \hat{\sigma}^2 = \frac{r^2}{D-d}$$

The observed Fisher Information $\mathcal{I}(\hat{\sigma}) = -\ell''(\hat{\sigma})$ evaluated at $\hat{\sigma}$ is:

$$\mathcal{I}(\hat{\sigma}) = \left[-\frac{D-d}{\sigma^2} + \frac{3r^2}{\sigma^4}\right]_{\sigma=\hat{\sigma}} = \frac{2(D-d)}{\hat{\sigma}^2}$$

As $D \to \infty$ (with $d$ fixed), the posterior variance $\text{Var}(\sigma|\mathbf{u}) \approx \mathcal{I}(\hat{\sigma})^{-1} \propto (D-d)^{-1}$ vanishes via the Laplace approximation. Furthermore, since $r^2/\sigma_0^2 \sim \chi^2_{D-d}$, the estimator $\hat{\sigma}^2$ satisfies $\mathbb{E}[\hat{\sigma}^2] = \sigma_0^2$ and $\text{Var}(\hat{\sigma}^2) = \frac{2\sigma_0^4}{D-d} \to 0$. By the Continuous Mapping Theorem, $\hat{\sigma}$ is a consistent estimator of the true noise level $\sigma_0$, and the posterior distribution $p(\sigma|\mathbf{u})$ concentrates at the true value $\sigma_0$. □

**Remark on Resolution Scaling and Log-SNR Shift.** Proposition 1 also explains why inferring the latent noise level from spatial statistics becomes easier as image resolution increases. When the ambient dimension increases from $D_{\text{low}}$ to $D_{\text{high}}$ while the intrinsic manifold dimension $d \ll D$ remains comparatively small, the variance of the implicit noise-scale estimator $\hat{\sigma}^2 = r^2/(D-d)$ shrinks as $\mathcal{O}\big((D_{\text{high}} - d)^{-1}\big)$. Simultaneously, for a fixed per-pixel noise variance $\sigma^2$, the expected squared orthogonal norm $\mathbb{E}[r^2] = (D-d)\sigma^2$ scales linearly with the codimension, which corresponds to the well-known high-resolution log-SNR shift $\Delta\text{SNR}_{\text{dB}} = 10\log_{10}(D_{\text{high}}/D_{\text{low}})$ in standard diffusion schedules.

## D Concentration of $p(t|\mathbf{u})$ via Proximity

In this section, we provide a rigorous proof that for autonomous generative models, the posterior distribution $p(t|\mathbf{u})$ converges weakly to a Dirac measure $\delta(t)$ as the observation $\mathbf{u}$ approaches the data support $\mathcal{X}$.

We prove this result for two separate cases. In Case 1, we assume that the data is discrete and finite. In this case, we prove that if the ambient dimension satisfies $D > 2$, regardless of the number of data points, as we approach a data point the posterior $p(t|\mathbf{u})$ converges weakly to the Dirac measure $\delta(t)$. In the second case, we assume that the data lies on a manifold, but the dimension of the manifold $d$ and the ambient dimension $D$ satisfy $D - d > 2$. In this case, we show that as we approach the data manifold, the posterior $p(t|\mathbf{u})$ again converges weakly to the Dirac measure $\delta(t)$.

We first establish a general lemma for the concentration of the specific family of distributions that arise in our analysis.

**Lemma 4** (Concentration of the Inverse-Gamma Kernel)**.** *Let $q(v;\beta) \propto v^{-\alpha}\exp(-\frac{\beta}{v})$ be a probability density on $v > 0$, with fixed shape $\alpha > 1$ and scale parameter $\beta > 0$. As the scale $\beta \to 0$, the distribution $q(v;\beta)$ converges weakly to a Dirac mass at zero: $q(v;\beta) \xrightarrow{w} \delta(v)$.*

*Proof.* This density corresponds to an Inverse-Gamma distribution $\mathcal{IG}(\alpha-1,\beta)$, which is well-defined for any shape parameter $\alpha > 1$. To show weak convergence to a Dirac mass at zero, it is sufficient to prove that the distribution converges in probability to zero; that is, for any constant $\epsilon > 0$, the probability $\mathbb{P}(v > \epsilon)$ vanishes as $\beta \to 0$. We evaluate this probability directly by integrating the density. Using the substitution $z = \beta/v$, we obtain:

$$\mathbb{P}(v > \epsilon) = \int_{\epsilon}^{\infty} \frac{\beta^{\alpha-1}}{\Gamma(\alpha-1)} v^{-\alpha} \exp\left(-\frac{\beta}{v}\right) dv = \frac{1}{\Gamma(\alpha-1)} \int_{0}^{\beta/\epsilon} z^{\alpha-2} \exp(-z) dz. \tag{43}$$

Since $\alpha > 1$, the exponent satisfies $\alpha - 2 > -1$, meaning the integrand is locally integrable at zero. As the scale parameter $\beta \to 0$, the upper limit of the integral $\beta/\epsilon \to 0$. Thus, the integral evaluates to zero: $\lim_{\beta\to 0} \mathbb{P}(v > \epsilon) = 0$. This demonstrates that the random variable $v$ converges in probability to 0, which strictly implies weak convergence to the Dirac measure $\delta(v)$. □

### D.1 Case 1: Discrete Data Support

**Lemma 5.** *Let the data support be a discrete set $\mathcal{X} = \{\mathbf{x}_k\}_{k=1}^N$. Let $\mathbf{u} = \mathbf{x}_j + \boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\| = \epsilon$. Assume the ambient dimension $D > 2$, the prior $p(t)$ is continuous with $p(0) > 0$, and the noise schedule $b(t)$ is continuous, strictly increasing, with $b(0) = 0$ and $a(0) = 1$. Then as $\epsilon \to 0$, $p(t|\mathbf{u}) \xrightarrow{w} \delta(t)$.*

*Proof.* Let $v = b(t)^2$ be the variance. Since $b(t)$ is strictly increasing and continuous with $b(0) = 0$, the mapping $t \mapsto v$ is a homeomorphism near zero. Therefore, proving $p(v|\mathbf{u}) \xrightarrow{w} \delta(v)$ is sufficient to imply $p(t|\mathbf{u}) \xrightarrow{w} \delta(t)$.

The marginal likelihood of the observation for general affine schedules is a mixture of Gaussians:

$$p(\mathbf{u}|v) = \frac{1}{N} \sum_{k=1}^{N} \underbrace{(2\pi v)^{-D/2} \exp\left(-\frac{\|\mathbf{u} - a(t)\mathbf{x}_k\|^2}{2v}\right)}_{L_k(v)}. \tag{44}$$

By Bayes' rule, the posterior $p(v|\mathbf{u})$ is a mixture of the individual component posteriors:

$$p(v|\mathbf{u}) = \sum_{k=1}^{N} W_k(\epsilon) p_k(v|\mathbf{u}), \tag{45}$$

where $p_k(v|\mathbf{u}) = \frac{L_k(v)p(v)}{Z_k(\epsilon)}$ is the normalized component posterior, and $Z_k(\epsilon) = \int_0^\infty L_k(z)p(z)dz$ is the component evidence. The mixing weights are determined by the ratio of component evidences: $W_k(\epsilon) = \frac{Z_k(\epsilon)}{\sum_{i=1}^N Z_i(\epsilon)}$. Note that since $p(t)$ is bounded below near zero, the induced prior $p(v)$ satisfies $p(v) \geq c_0 > 0$ in a neighborhood of $v = 0$.

We analyze the asymptotic behavior of the evidence integrals as $\epsilon \to 0$. For the nearest neighbor ($k = j$), the squared distance expands to $\|\mathbf{u} - a(t)\mathbf{x}_j\|^2 = \|\boldsymbol{\delta} + (1 - a(t))\mathbf{x}_j\|^2$. Since $a(t) \to 1$ as $t \to 0$, this distance converges to $\epsilon^2$ to leading order. The evidence integral is asymptotically:

$$Z_j(\epsilon) \approx \int_0^\infty (2\pi z)^{-D/2} \exp\left(-\frac{\epsilon^2}{2z}\right) p(z) dz. \tag{46}$$

Substituting $y = \frac{\epsilon^2}{2z}$ yields $Z_j(\epsilon) \propto (\epsilon^2)^{1-D/2} \int_0^\infty y^{D/2-2} e^{-y} p(\frac{\epsilon^2}{2y}) dy$. For $D > 2$, the integral converges to a finite value bounded away from zero. Consequently, the evidence scales as $Z_j(\epsilon) = \mathcal{O}(\epsilon^{2-D})$. Since $2 - D < 0$, $Z_j(\epsilon) \to \infty$ as $\epsilon \to 0$.

In contrast, for any other point $k \neq j$, the distance $\|\mathbf{u} - a(t)\mathbf{x}_k\|^2 \to \Delta_{jk}^2 > 0$. The evidence integral $Z_k(\epsilon)$ converges to a finite constant $Z_k(0) < \infty$, as the exponential term $\exp(-\Delta_{jk}^2/2z)$ suppresses the singularity at $z = 0$.

This divergence of the nearest-neighbor evidence implies that the mixing weights collapse:

$$W_j(\epsilon) = \frac{1}{1 + \sum_{k\neq j} \frac{Z_k(\epsilon)}{Z_j(\epsilon)}} \to 1. \tag{47}$$

Thus, the posterior is asymptotically dominated by the nearest neighbor component $p_j(v|\mathbf{u}) \propto v^{-D/2} \exp(-\frac{\epsilon^2}{2v}) p(v)$. This is an Inverse-Gamma kernel with scale parameter $\beta = \epsilon^2/2$. As $\epsilon \to 0$, the scale vanishes, and by Lemma 4, $p(v|\mathbf{u}) \xrightarrow{w} \delta(v)$. □

### D.2 Case 2: Continuous Low-Dimensional Manifold

**Lemma 6.** *Let data lie on a smooth ($C^2$) $d$-dimensional submanifold $\mathcal{M} \subset \mathbb{R}^D$ with codimension $k = D - d > 2$. Let $r = dist(\mathbf{u}, \mathcal{M})$ be the orthogonal distance. Assume $p_{data}$ is continuous and bounded on $\mathcal{M}$. As $r \to 0$, $p(t|\mathbf{u}) \xrightarrow{w} \delta(t)$.*

*Proof.* We analyze the marginal likelihood integral over the manifold $\mathcal{M}$ as a function of the variance $v = b(t)^2$:

$$p(\mathbf{u}|v) = \int_{\mathcal{M}} (2\pi v)^{-D/2} \exp\left(-\frac{\|\mathbf{u} - a(t)\mathbf{x}\|^2}{2v}\right) p_{data}(\mathbf{x}) d\mathbf{x}. \tag{48}$$

Let $\mathbf{x}_{proj}$ be the orthogonal projection of the fixed observation $\mathbf{u}$ onto $\mathcal{M}$. Since $\mathbf{u}$ is fixed, $\mathbf{x}_{proj}$ is a constant vector. We define local coordinates on the manifold centered at $\mathbf{x}_{proj}$. Let $\mathbf{y} \in \mathbb{R}^d$ represent coordinates in the tangent space $T_{\mathbf{x}_{proj}}\mathcal{M}$. For points $\mathbf{x}$ near $\mathbf{x}_{proj}$, we have the expansion $\mathbf{x}(\mathbf{y}) = \mathbf{x}_{proj} + \mathbf{J}\mathbf{y} + O(\|\mathbf{y}\|^2)$. Because $a(t) \to 1$ as $v \to 0$, the squared distance decomposes to leading order as:

$$\|\mathbf{u} - a(t)\mathbf{x}\|^2 \approx \|\mathbf{u} - \mathbf{x}_{proj} + \mathbf{x}_{proj} - \mathbf{x}\|^2 = r^2 + \|\mathbf{x}_{proj} - \mathbf{x}\|^2 \approx r^2 + \|\mathbf{y}\|^2. \tag{49}$$

(The cross-term vanishes because $\mathbf{u} - \mathbf{x}_{proj}$ is orthogonal to the tangent space).

Substituting this into the integral:

$$p(\mathbf{u}|v) \approx (2\pi v)^{-D/2} e^{-\frac{r^2}{2v}} \int_{\mathbb{R}^d} e^{-\frac{\|\mathbf{y}\|^2}{2v}} p_{data}(\mathbf{x}(\mathbf{y})) d\mathbf{y}. \tag{50}$$

We apply the Laplace method for asymptotic integrals as $v \to 0$ (which corresponds to the high-likelihood regime). The Gaussian kernel $e^{-\|\mathbf{y}\|^2/2v}$ concentrates mass entirely at $\mathbf{y} = 0$ (i.e., $\mathbf{x} = \mathbf{x}_{proj}$). Since $p_{data}$ is continuous, we can pull the value $p_{data}(\mathbf{x}_{proj})$ out of the integral:

$$\int_{\mathbb{R}^d} e^{-\frac{\|\mathbf{y}\|^2}{2v}} p_{data}(\mathbf{x}(\mathbf{y})) d\mathbf{y} \approx p_{data}(\mathbf{x}_{proj}) \int_{\mathbb{R}^d} e^{-\frac{\|\mathbf{y}\|^2}{2v}} d\mathbf{y}. \tag{51}$$

The remaining integral is a standard unnormalized Gaussian integral over $d$ dimensions, equal to $(2\pi v)^{d/2}$. Combining the pre-factors, let $k = D - d$ be the codimension. The likelihood takes the form of an Inverse-Gamma kernel:

$$p(\mathbf{u}|v) \propto v^{-k/2} \exp\left(-\frac{r^2}{2v}\right). \tag{52}$$

For a standard uniform time prior $p(t) = 1$ and a linear schedule $b(t) = t$, the induced variance prior is $p(v) \propto v^{-1/2}$. Multiplying the likelihood by this prior yields the posterior:

$$p(v|\mathbf{u}) \propto v^{-(k+1)/2} \exp\left(-\frac{r^2}{2v}\right) \tag{53}$$

This is exactly an Inverse-Gamma distribution $\mathcal{IG}(\alpha, \beta)$ with shape $\alpha = \frac{k-1}{2}$ and scale $\beta = \frac{r^2}{2}$. Provided the codimension $k > 1$, the shape parameter $\alpha > 0$. As the distance to the manifold $r \to 0$, the scale parameter $\beta \to 0$. By Lemma 4, the distribution converges weakly to a Dirac mass: $p(v|\mathbf{u}) \xrightarrow{w} \delta(v)$. This implies $p(t|\mathbf{u}) \xrightarrow{w} \delta(t)$. □

### D.3 Asymptotic Stability of the Transport Correction Term

In Section 5, we established that the Transport Correction term takes the form of the posterior covariance between the effective gain and the conditional energy gradient:

$$\mathrm{Cov}_{t|\mathbf{u}}(\lambda(t), \nabla E_t(\mathbf{u})) = \mathbb{E}_{t|\mathbf{u}}\left[\left(\lambda(t) - \overline{\lambda}(\mathbf{u})\right)\left(\nabla E_t(\mathbf{u}) - \nabla E_{\text{marg}}(\mathbf{u})\right)\right]. \tag{54}$$

In this subsection, we rigorously prove that despite the presence of divergent conditional energy gradients near the data manifold, this Transport Correction term remains strictly bounded. We will show that by reformulating the variance in terms of an invariant Inverse-Gamma variable, the vanishing rate of the effective gain perfectly offsets the singular $\mathcal{O}(1/b(t)^2)$ growth of the gradient. Because the distance to the manifold asymptotically scales with the noise schedule near the data (i.e., $\|\mathbf{u}-\mathbf{x}\| \sim \mathcal{O}(b(t))$), we will demonstrate that this covariance evaluates to a bounded $\mathcal{O}(1)$ vector for linear-gain parameterizations (e.g., Flow Matching, DDPM), and an $\mathcal{O}(b(t))$ vector that smoothly vanishes for quadratic-gain parameterizations (e.g., EDM, EqM).

While the posterior concentrates as $\mathbf{u}$ approaches the data manifold, the raw energy gradients diverge, creating a potential geometric singularity. We now rigorously show how the inverse-gamma asymptotics derived above resolve this indeterminate form.

Let $\mathbf{u} = \mathbf{x}_j + \boldsymbol{\delta}$, where $\mathbf{x}_j$ is the nearest data point on the manifold, $\|\boldsymbol{\delta}\| = \epsilon$, and $\mathbf{n} = \boldsymbol{\delta}/\epsilon$ is the unit normal vector pointing away from the manifold. For general affine schedules $\mathbf{u}_t = a(t)\mathbf{x} + b(t)\boldsymbol{\epsilon}$, the conditional score near the manifold is dominated by this nearest neighbor:

$$\nabla E_t(\mathbf{u}) \approx \frac{\mathbf{u} - a(t)\mathbf{x}_j}{b(t)^2} = \frac{\boldsymbol{\delta} + (1-a(t))\mathbf{x}_j}{b(t)^2}. \tag{55}$$

From the previous lemmas, we established that as $\epsilon \to 0$, the posterior for the variance $v = b(t)^2$ concentrates as an Inverse-Gamma distribution. We can decouple the local distance scale $\epsilon$ from the random variable by defining the normalized variable $z = 2v/\epsilon^2 = 2b(t)^2/\epsilon^2$. By the scaling properties of the Inverse-Gamma distribution, $z \sim \mathcal{IG}(\frac{k-1}{2}, 1)$, which is an invariant distribution entirely independent of $\epsilon$. Consequently, the noise scale is given directly by $b(t) = \frac{\epsilon}{\sqrt{2}} z^{1/2}$.

**Unified Asymptotic Framework.** Because both the effective gain $\lambda(t)$ and the signal decay $1-a(t)$ depend on the random variable $t$, neither can be factored outside the covariance before expressing its functional dependence on $z$. To evaluate all schedules rigorously under a single unified law, we parameterize the asymptotic behavior of the effective gain and the signal decay as $t \to 0$ by power laws of $b(t)$:

$$\lambda(t) \approx c_\lambda b(t)^m, \qquad 1 - a(t) \approx c_a b(t)^q, \tag{56}$$

where $m > 0$ and $q > 0$ are the respective scaling exponents (using $q$ to distinguish the signal decay exponent from the manifold codimension $k = D - d$), and $c_\lambda, c_a$ are schedule-specific constants. Substituting $b(t) = \frac{\epsilon}{\sqrt{2}} z^{1/2}$ into both terms *prior* to evaluating the covariance yields:

$$\lambda(t) \approx c_\lambda \left(\frac{\epsilon}{\sqrt{2}}\right)^m z^{m/2}, \qquad \nabla E_t(\mathbf{u}) \approx \frac{2}{\epsilon}\mathbf{n}\, z^{-1} + c_a 2^{1-q/2}\epsilon^{q-2}\mathbf{x}_j\, z^{q/2-1}. \tag{57}$$

By the bilinearity of covariance, all $z$-dependent factors remain strictly inside $\mathrm{Cov}_z(\cdot,\cdot)$, while the deterministic constants and powers of $\epsilon$ factor out cleanly:

$$\begin{aligned}\mathrm{Cov}_{t|\mathbf{u}}\big(\lambda(t), \nabla E_t(\mathbf{u})\big) &\approx \mathrm{Cov}_{t|\mathbf{u}}\left(\lambda(t), \frac{\boldsymbol{\delta}}{b(t)^2}\right) + \mathrm{Cov}_{t|\mathbf{u}}\left(\lambda(t), \frac{(1-a(t))\mathbf{x}_j}{b(t)^2}\right) \\ &\propto \underbrace{\epsilon^{m-1}\mathbf{n}\,\mathrm{Cov}_z\left(z^{m/2}, z^{-1}\right)}_{\text{Normal Component}} + \underbrace{\epsilon^{m+q-2}\mathbf{x}_j\,\mathrm{Cov}_z\left(z^{m/2}, z^{q/2-1}\right)}_{\text{Positional Component}}.\end{aligned} \tag{58}$$

Because $z$ has an invariant $\mathcal{IG}(\frac{k-1}{2}, 1)$ distribution independent of $\epsilon$, the covariance $\mathrm{Cov}_z(z^{m/2}, \cdot)$ requires the $(m/2)$-th moment of $z$ to exist, which holds whenever the shape parameter satisfies $\alpha = \frac{k-1}{2} > \frac{m}{2}$ (i.e., $k > 2$ for linear gain $m = 1$, and $k > 3$ for quadratic gain $m = 2$; on bounded $t \in [0,1]$, $\epsilon\,\mathbb{E}[z] = \mathcal{O}(\epsilon \ln(1/\epsilon)) \to 0$ even at the boundary $k = 3$ for $m = 2$). Thus, Eq. (58) immediately determines the exact near-manifold limit for every standard architecture via its exponents $(m, q)$:

- **EDM** ($a(t) = 1 \implies c_a = 0$ [or $q = \infty$]; $\lambda(t) = -b(t)^2 \implies m = 2$): Because $1 - a(t) = 0$ identically, the positional component is zero. For the normal component, the exponent is $m - 1 = 1$, so the covariance scales as $\mathcal{O}(\epsilon) = \mathcal{O}(b(t)) \to \mathbf{0}$.

- **DDPM** ($a(t) = \sqrt{1-b(t)^2} \approx 1 - \frac{1}{2}b(t)^2 \implies q = 2$; $\lambda(t) \approx b(t) \implies m = 1$): In the positional component, the second argument of the covariance evaluates to $z^{q/2-1} = z^{2/2-1} = z^0 = 1$. Because the covariance of any random variable with a constant is identically zero, $\mathrm{Cov}_z(z^{1/2}, 1) = 0$, and the positional component vanishes completely. In the normal component, $\epsilon^{m-1} = \epsilon^0 = 1$, so the Transport Correction converges to a non-zero $\mathcal{O}(1)$ constant vector pointing strictly along the manifold normal $\mathbf{n}$.
- **Flow Matching** ($a(t) = 1 - t$, $b(t) = t \implies q = 1$; $\lambda(t) = \frac{t}{1-t} \approx b(t) \implies m = 1$): Both scaling exponents evaluate to zero ($m - 1 = 0$ and $m + q - 2 = 0$), while $\mathrm{Cov}_z(z^{1/2}, z^{-1})$ and $\mathrm{Cov}_z(z^{1/2}, z^{-1/2})$ evaluate to finite non-zero constants. Thus, $\epsilon$ cancels in both terms and the Transport Correction converges to a bounded $\mathcal{O}(1)$ constant vector comprising both normal and positional components.
- **Equilibrium Matching (EqM)** ($a(t) = 1 - t$, $b(t) = t \implies q = 1$; $\lambda(t) = \frac{t^2}{1-t} \approx b(t)^2 \implies m = 2$): Both exponents evaluate to one ($m - 1 = 1$ and $m + q - 2 = 1$). Consequently, both the normal and positional components scale as $\mathcal{O}(\epsilon) = \mathcal{O}(b(t))$ and smoothly vanish as $\mathbf{u} \to \mathcal{X}$.

In all cases, the potentially fatal $\mathcal{O}(1/b(t)^2)$ singularity of the raw energy landscape is neutralized. The implicit geometry of the learned vector field acts as a Riemannian preconditioner, strictly bounding the transport dynamics to ensure stability during autonomous generation.

### D.4 Behavior Under Exponentially Vanishing Priors (Log-Normal Schedule)

In Lemmas 4 and 5, we assumed a non-vanishing prior near the boundary ($p(0) > 0$), which produces an Inverse-Gamma posterior that collapses to a Dirac delta $\delta(t)$ as the distance to the manifold $r = \mathrm{dist}(\mathbf{u}, \mathcal{M}) \to 0$. Here we analyze how an exponentially vanishing prior—specifically the EDM log-normal noise schedule [53], $\ln b \sim \mathcal{N}(\mu, \sigma^2)$ (or equivalently a logit-normal prior on $t \in (0, 1)$, $\ln \frac{t}{1-t} \sim \mathcal{N}(\mu, \sigma^2)$, which has identical $\exp(-(\ln t - \mu)^2/(2\sigma^2))$ asymptotics as $t \to 0$)—alters the limiting posterior and regularizes the marginal energy.

Under a log-normal prior on the noise scale $b$, the prior density is:

$$p(b) = \frac{1}{b\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln b - \mu)^2}{2\sigma^2}\right). \tag{59}$$

From Appendix D, at a small orthogonal distance $r = \mathrm{dist}(\mathbf{u}, \mathcal{M})$ from a $d$-dimensional manifold with codimension $k = D - d$ (or $k = D$ for discrete data points), the marginal likelihood scales as $p(\mathbf{u}|b) \propto b^{-k} \exp\left(-\frac{r^2}{2b^2}\right)$. Combining the likelihood and prior via Bayes' rule yields the unnormalized posterior density:

$$p(b|\mathbf{u}) \propto \frac{1}{b} \exp\left(-k \ln b - \frac{r^2}{2b^2} - \frac{(\ln b - \mu)^2}{2\sigma^2}\right). \tag{60}$$

Unlike a uniform or polynomial prior, the quadratic term $-(\ln b - \mu)^2/(2\sigma^2)$ decays super-polynomially as $b \to 0$, dominating the polynomial likelihood singularity $-k \ln b$. Consequently, the marginal evidence integral $p(\mathbf{u}) = \int_0^\infty p(\mathbf{u}|b)p(b)\, db$ remains strictly finite even at $r = 0$, regularizing the marginal energy $E_{\text{marg}}(\mathbf{u}) = -\log p(\mathbf{u})$ on the data manifold.

Taking the limit $r \to 0$ (so that $\exp(-r^2/(2b^2)) \to 1$ pointwise for every $b > 0$) and completing the square in $\ln b$ in the exponent:

$$-k \ln b - \frac{(\ln b - \mu)^2}{2\sigma^2} = -\frac{\big(\ln b - (\mu - k\sigma^2)\big)^2}{2\sigma^2} + \text{const}, \tag{61}$$

we obtain the exact limiting posterior distribution at the data manifold:

$$p(b|\mathbf{u}) \xrightarrow{w} \mathcal{LN}\big(\mu - k\sigma^2,\, \sigma^2\big) \quad \text{as } r \to 0. \tag{62}$$

Note that for discrete points ($k = D$), this gives $\mathcal{LN}(\mu - D\sigma^2, \sigma^2)$, and at the critical codimension $k = 2$ where a flat prior fails to concentrate, the log-normal prior still yields a well-defined posterior $\mathcal{LN}(\mu - 2\sigma^2, \sigma^2)$. This result has two immediate implications: (1) because the limiting posterior retains a strictly positive variance $\sigma^2 > 0$, posterior uncertainty in autonomous noise prediction never vanishes even on the manifold; and (2) in high-dimensional settings where $k = D - d \gg 1$, the mean shift $-k\sigma^2$ pushes the characteristic floor scale $b_{\text{floor}} \triangleq \exp(\mu - k\sigma^2)$ exponentially close to zero, so that $p(b|\mathbf{u})$ behaves as an effective Dirac delta at zero across all practical solver tolerances.

**Impact on $\overline{\lambda}(\mathbf{u})$, the Transport Correction, and Overall Stability.** While replacing a uniform prior with a log-normal or logit-normal prior alters the exact posterior $p(t|\mathbf{u}) \propto p(\mathbf{u}|t)p(t)$—and therefore modifies the exact finite-dimensional values of the posterior-averaged gain $\overline{\lambda}(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}[\lambda(t)]$ and the Transport Correction $\mathrm{Cov}_{t|\mathbf{u}}(\lambda(t), \nabla E_t(\mathbf{u}))$—it does not change our fundamental stability conclusions:

1. **Strict Boundedness at $r \to 0$:** Under the limiting distribution $\mathcal{LN}(\mu - k\sigma^2, \sigma^2)$, all positive and negative polynomial moments exist and are strictly finite:
$$\mathop{\mathbb{E}}_{b|\mathbf{u}}[b^p] = \exp\left(p(\mu - k\sigma^2) + \frac{1}{2}p^2\sigma^2\right) < \infty \qquad \forall\, p \in \mathbb{R}. \tag{63}$$
Because both $\lambda(t) \propto b(t)^m$ and $\nabla E_t(\mathbf{u}) \approx \frac{\boldsymbol{\delta} + (1-a(t))\mathbf{x}_j}{b(t)^2}$ are polynomial in $b$ and $b^{-1}$, the posterior-averaged gain $\overline{\lambda}(\mathbf{u})$, the marginal gradient $\nabla E_{\text{marg}}(\mathbf{u})$, and the Transport Correction $\mathrm{Cov}_{t|\mathbf{u}}(\lambda(t), \nabla E_t(\mathbf{u}))$ are guaranteed to remain finite and bounded all the way to $r = 0$.
2. **Invariance of High-Dimensional Concentration ($b(t) \gg b_{\text{floor}}, k \gg 1$):** Away from the floor scale $b_{\text{floor}} = \exp(\mu - k\sigma^2)$, maximizing the log-posterior $\ln p(b|\mathbf{u}) = -k \ln b - \frac{r^2}{2b^2} - \frac{(\ln b - \mu)^2}{2\sigma^2} + \text{const}$ yields the stationarity condition $\frac{r^2}{b^2} = k + \frac{\ln b - \mu}{\sigma^2}$. Whenever $b(t) \gg b_{\text{floor}}$ and the codimension dominates the prior precision ($k \gg \sigma^{-2}$), the likelihood curvature $\mathcal{O}(k)$ overwhelms the prior, so $p(b|\mathbf{u})$ concentrates sharply around $\hat{b}^2 \approx r^2/k$ with $\mathrm{Var}(\ln b|\mathbf{u}) = \mathcal{O}(k^{-1})$, leaving the interior asymptotic geometry completely unchanged.
3. **Parameterization Stability in the Floor Limit ($b(t) \ll b_{\text{floor}}$):** One might wonder whether the finiteness of $\lim_{r\to 0} \mathbb{E}_{\tau|\mathbf{u}}[1/b(\tau)] = \exp(-(\mu - k\sigma^2) + \frac{1}{2}\sigma^2) \triangleq C_{\text{LN}} < \infty$ under a log-normal prior prevents autonomous noise prediction from breaking down as $r \to 0$, or conversely whether the non-degenerate limiting posterior destabilizes signal prediction. Neither occurs:
   - *Noise prediction still fails:* When $b(t) \ll b_{\text{floor}}$, the posterior freezes at $b(\tau) \sim b_{\text{floor}}$, causing the autonomous noise prediction to vanish as $f^{*,\text{noise}}(\mathbf{u}) \approx (\mathbf{u} - \mathbf{x})C_{\text{LN}} = \mathcal{O}(b(t)/b_{\text{floor}}) \to \mathbf{0}$ while the true oracle noise remains $f_t^{*,\text{noise}}(\mathbf{u}) \approx (\mathbf{u} - \mathbf{x})/b(t) = \boldsymbol{\epsilon} \sim \mathcal{O}(1)$. Thus $b(t)\,\mathbb{E}_{\tau|\mathbf{u}}[1/b(\tau)] \to 0$, driving the discrepancy factor in Eq. (75) to $\left|b(t)\,\mathbb{E}_{\tau|\mathbf{u}}[1/b(\tau)] - 1\right| \to |0 - 1| = 1$ (a 100% relative error) so that $\Delta\mathbf{v}_{\text{noise}} \propto \mathcal{O}(\dot{b}(t))$ still diverges for square-root schedules.
   - *Signal prediction remains a stable attractor:* On a discrete dataset $\mathcal{X} = \{\mathbf{x}_j\}_{j=1}^N$, the frozen posterior at $r = 0$ induces a tiny constant smoothing shift $\boldsymbol{\eta}_k \triangleq \hat{\mathbf{x}}(\mathbf{x}_k) - \mathbf{x}_k = \mathbb{E}_{b\sim\mathcal{LN}(\mu - k\sigma^2, \sigma^2)}[D_b^*(\mathbf{x}_k) - \mathbf{x}_k]$ of doubly-exponentially small norm $\|\boldsymbol{\eta}_k\| \sim \exp(-\frac{1}{2}\Delta_{\min}^2 e^{2(k\sigma^2 - \mu)})$. Although comparing $\mathbf{v}_{\text{aut}}$ against the unsmoothed oracle at $\mathbf{u} = \mathbf{x}_k$ for $b(t) \ll b_{\text{floor}}$ multiplies $\|\boldsymbol{\eta}_k\|$ by $|\dot{b}(t)/b(t)|$, the actual autonomous signal ODE $\frac{d\mathbf{u}}{dt} = \frac{\dot{b}(t)}{b(t)}(\mathbf{u} - \hat{\mathbf{x}}(\mathbf{u}))$ is a strictly contractive attractor toward the slightly smoothed fixed point $\mathbf{u}^* \approx \mathbf{x}_k + \boldsymbol{\eta}_k$, where $\mathbf{u}^* - \hat{\mathbf{x}}(\mathbf{u}^*) = \mathbf{0}$ and the velocity vanishes cleanly.

# E Analysis of Specific Architectures: Exact Low-Noise Asymptotics

In this section, we show that near the data manifold, the effective gain $\overline{\lambda}(\mathbf{u})$ in the vector field vanishes at a rate that perfectly counteracts the divergence of the gradient of the marginal energy. We assume discrete data to analyze the asymptotic behavior of the learned autonomous field $f^*(\mathbf{u})$ near the data manifold ($\mathbf{u} \to \mathcal{X}$). We derive the exact form using the limit of the marginal energy gradient:

$$\nabla_{\mathbf{u}} E_{marg}(\mathbf{u}) \approx \frac{\mathbf{u} - a(t)\mathbf{x}}{b(t)^2} \tag{64}$$

Substituting this into the General Energy-Aligned Decomposition ($f^* = \overline{\lambda}\nabla E + \text{Drift}$), we examine the properties of the target learned by each parameterization as $t \to 0$.

**1. DDPM (Noise Prediction).** For this model, we have $a(t) \approx 1$. The effective gain is $\lambda(t) \approx b(t)$ and the drift is zero.

$$f^*(\mathbf{u}) \approx b(t) \cdot \frac{\mathbf{u} - \mathbf{x}}{b(t)^2} = \frac{\mathbf{u} - \mathbf{x}}{b(t)} \sim \mathcal{O}(1) \tag{65}$$

**Result (Bounded Target):** The learned target contains a removable singularity of order $\mathcal{O}(1)$. While the target itself is bounded, it does not inherently define the flow dynamics. Standard diffusion ODEs typically scale this target by $1/b(t)$ (i.e., $d\mathbf{u}/dt \propto f^*/b(t)$). Thus, while the network learns a stable quantity, the implied autonomous dynamics may still diverge without careful parameterization.

**2. EDM (Signal Prediction).** For EDM, $a(t) = 1$ and $b(t) = \sigma_t$. For the signal target ($c = 1, d = 0$) we get an effective gain of $\lambda(t) = -b(t)^2$ and a linear drift of $+\mathbf{u}$.

$$f^*(\mathbf{u}) \approx -b(t)^2 \cdot \frac{\mathbf{u} - \mathbf{x}}{b(t)^2} + \mathbf{u} = -(\mathbf{u} - \mathbf{x}) + \mathbf{u} = \mathbf{x} \sim \mathcal{O}(1). \tag{66}$$

**Result (Bounded Target):** The learned target algebraically simplifies directly to the clean data $\mathbf{x}$, which is $\mathcal{O}(1)$ and bounded. Similar to DDPM, the implied ODE dynamics require further stability analysis (see Section 6).

**3. Flow Matching (Velocity Prediction).** For Flow Matching, $a(t) = 1 - t$ and $b(t) = t$. The exact gain is $\lambda(t) = \frac{t}{1-t}$ and the linear drift coefficient is $-\frac{1}{1-t}$. Instead of an asymptotic approximation, the General Energy-Aligned Decomposition reveals a perfect algebraic cancellation for all $t \in (0, 1)$ when using the exact conditional energy gradient $\nabla E_t(\mathbf{u}) = \frac{\mathbf{u} - (1-t)D_t^*(\mathbf{u})}{t^2}$:

$$f_t^*(\mathbf{u}) = \underbrace{\frac{t}{1-t}}_{\text{Gain}} \cdot \underbrace{\frac{\mathbf{u} - (1-t)D_t^*(\mathbf{u})}{t^2}}_{\text{Gradient}} \underbrace{-\frac{1}{1-t}\mathbf{u}}_{\text{Drift}} = \frac{\mathbf{u} - t\mathbf{u} - (1-t)D_t^*(\mathbf{u})}{t(1-t)} = \frac{\mathbf{u} - D_t^*(\mathbf{u})}{t}. \tag{67}$$

Taking the posterior expectation yields the autonomous field $f^*(\mathbf{u}) = \mathbb{E}_{t|\mathbf{u}}\left[\frac{\mathbf{u} - D_t^*(\mathbf{u})}{t}\right] \sim \mathcal{O}(1)$.

**Result (Stable Transport):** The learned target algebraically simplifies to a finite velocity vector everywhere, not just asymptotically. For Flow Matching, this target *is* the ODE velocity ($d\mathbf{u}/dt = f^*(\mathbf{u})$). Since the target is $\mathcal{O}(1)$, the resulting generation trajectories approach the manifold with finite speed, ensuring stable transport without numerical explosion.

**4. Equilibrium Matching (Stabilized Transport).** For EqM, $a(t) = 1 - t, b(t) = t, c(t) = -t$, and $d(t) = t$. The exact gain is $\lambda(t) = \frac{t^2}{1-t}$ and the linear drift coefficient is $-\frac{t}{1-t}$. Again, applying the exact decomposition yields a perfect cancellation for all $t$:

$$f_t^*(\mathbf{u}) = \underbrace{\frac{t^2}{1-t}}_{\text{Gain}} \cdot \underbrace{\frac{\mathbf{u} - (1-t)D_t^*(\mathbf{u})}{t^2}}_{\text{Gradient}} \underbrace{-\frac{t}{1-t}\mathbf{u}}_{\text{Drift}} = \frac{\mathbf{u} - (1-t)D_t^*(\mathbf{u}) - t\mathbf{u}}{1-t} = \mathbf{u} - D_t^*(\mathbf{u}). \tag{68}$$

Taking the posterior expectation gives $f^*(\mathbf{u}) = \mathbf{u} - \mathbb{E}_{t|\mathbf{u}}[D_t^*(\mathbf{u})]$. As $\mathbf{u}$ approaches the clean data $\mathbf{x}$, $D_t^*(\mathbf{u}) \to \mathbf{x}$, yielding $f^*(\mathbf{u}) \to \mathbf{u} - \mathbf{x}$.

**Result (Vanishing Equilibrium):** The learned target scales exactly with the distance to the manifold, vanishing as $\mathcal{O}(\|\mathbf{u} - \mathbf{x}\|)$. Since EqM uses this target directly as the ODE velocity, the dynamics naturally slow down and stop at the data ($d\mathbf{u}/dt \to 0$). This creates a stable fixed point at the manifold, contrasting with the constant-velocity transport of Flow Matching.

# F Derivation of Stability Conditions

In this appendix, we provide the derivation of the Unified Sampler Dynamics and the rigorous proofs for the stability limits of the three parameterizations discussed in Section 6.

## F.1 Unified Sampler Dynamics

For general affine noise schedules defined by $\mathbf{u}_t = a(t)\mathbf{x} + b(t)\boldsymbol{\epsilon}$, the generation process involves integrating a differential equation. We derive the exact ODE by inverting the linear system relating the data, noise, and observation.

The flow of the process is given by differentiating the forward process:

$$\dot{\mathbf{u}} = \dot{a}(t)\mathbf{x} + \dot{b}(t)\boldsymbol{\epsilon}. \tag{69}$$

An autonomous generative model predicts a target $f^*(\mathbf{u}) = c(t)\mathbf{x} + d(t)\boldsymbol{\epsilon}$. Combining this with the observation identity $\mathbf{u} = a(t)\mathbf{x} + b(t)\boldsymbol{\epsilon}$, we form the linear system:

$$\begin{bmatrix} \mathbf{u} \\ f^*(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} a(t) & b(t) \\ c(t) & d(t) \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\epsilon} \end{bmatrix}. \tag{70}$$

Solving for $\mathbf{x}$ and $\boldsymbol{\epsilon}$ and substituting them into the flow equation $\dot{\mathbf{u}}$, we obtain the general sampler ODE:

$$\frac{d\mathbf{u}}{dt} = \underbrace{\left(\frac{\dot{a}d - \dot{b}c}{ad - bc}\right)}_{\mu(t)} \mathbf{u} + \underbrace{\left(\frac{\dot{b}a - \dot{a}b}{ad - bc}\right)}_{\nu(t)} f^*(\mathbf{u}). \tag{71}$$

We identify $\mu(t)$ as the schedule drift coefficient and $\nu(t)$ as the effective gain of the parameterization.

In order to analyze the stability of generation, we also consider generation using the optimal conditional model $f_t^*(\mathbf{u})$ that has access to the noise level. We define the Drift Perturbation Error $\Delta\mathbf{v}$ as the norm difference between the autonomous drift (using $f^*(\mathbf{u})$) and the oracle drift. Since the linear term $\mu(t)\mathbf{u}$ is identical for both, it cancels out giving us:

$$\Delta\mathbf{v}(\mathbf{u}, t) = |\nu(t)| \cdot \|f^*(\mathbf{u}) - f_t^*(\mathbf{u})\|. \tag{72}$$

If this Drift Perturbation Error has singularities in it, this results in unstable generation dynamics.

### F.2 Stability Analysis by Parameterization

We evaluate the limit of $\Delta\mathbf{v}$ as $t \to 0$ (near the data manifold) for standard models. We assume standard boundary conditions $a(t) \to 1$ and $b(t) \to 0$.

**Case 1: Noise Prediction (DDPM/DDIM)** Target: $\boldsymbol{\epsilon}$ ($c = 0, d = 1$). Using the unified coefficients ($c = 0, d = 1$), the effective gain simplifies exactly to $\nu(t) = \frac{\dot{b}a - \dot{a}b}{a} = \dot{b}(t) - \dot{a}(t)\frac{b(t)}{a(t)}$. Near the data manifold ($t \to 0$), the boundary conditions dictate $a(t) \to 1$ and $b(t) \to 0$. Provided $\dot{a}(t)$ remains bounded (e.g., $\dot{a}(t) = -1$ for Flow Matching), the second term vanishes entirely. For Variance Preserving (VP) schedules, this stability is guaranteed algebraically by the identity $a(t)^2 + b(t)^2 = 1$, which gives $\dot{a}(t) = -\frac{b(t)}{a(t)}\dot{b}(t)$. Substituting this yields exactly $\nu(t) = \frac{\dot{b}(t)}{a(t)^2}$. In either formulation, as $a(t) \to 1$, the gain asymptotically behaves as the noise derivative: $\nu(t) \approx \dot{b}(t)$.

For a general noise level $\tau$, the optimal conditional noise estimator is $f_\tau^*(\mathbf{u}) = \boldsymbol{\epsilon}_\tau^*(\mathbf{u}) = \frac{\mathbf{u} - a(\tau)D_\tau^*(\mathbf{u})}{b(\tau)} \approx \frac{\mathbf{u} - D_\tau^*(\mathbf{u})}{b(\tau)}$ near the manifold (note that any $(1 - a(\tau))\mathbf{x}/b(\tau)$ contribution either vanishes as $\mathcal{O}(b(\tau))$ for EDM/VP schedules or equals the constant $\mathbf{x}$ for linear schedules $1 - a(\tau) = b(\tau) = \tau$, canceling identically in the difference below). Because the conditional denoiser $D_\tau^*(\mathbf{u}) = \mathbb{E}[\mathbf{x}|\mathbf{u}, \tau]$ depends on the integration variable $\tau$, it cannot be replaced by $\mathbf{x}$ and factored directly out of $\mathbb{E}_{\tau|\mathbf{u}}[\cdot]$. To isolate the $\tau$-independent geometric vector cleanly, we add and subtract the nearest clean data point $\mathbf{x}$ in the numerator:

$$f_\tau^*(\mathbf{u}) \approx \frac{\mathbf{u} - \mathbf{x}}{b(\tau)} + \frac{\mathbf{x} - D_\tau^*(\mathbf{u})}{b(\tau)}. \tag{73}$$

Substituting this decomposition into the drift error $\Delta\mathbf{v}_{\text{noise}} \approx |\dot{b}(t)|\|\mathbb{E}_{\tau|\mathbf{u}}[f_\tau^*(\mathbf{u})] - f_t^*(\mathbf{u})\|$ separates the error into two distinct terms:

$$\Delta\mathbf{v}_{\text{noise}} \approx \left\| \underbrace{\dot{b}(t)(\mathbf{u} - \mathbf{x})\left(\mathbb{E}_{\tau|\mathbf{u}}\left[\frac{1}{b(\tau)}\right] - \frac{1}{b(t)}\right)}_{\text{Dominant Geometric Error (Jensen Gap)}} + \underbrace{\dot{b}(t)\left(\mathbb{E}_{\tau|\mathbf{u}}\left[\frac{\mathbf{x} - D_\tau^*(\mathbf{u})}{b(\tau)}\right] - \frac{\mathbf{x} - D_t^*(\mathbf{u})}{b(t)}\right)}_{\text{Vanishing Denoiser Residual}} \right\|. \tag{74}$$

In the first term, $(\mathbf{u} - \mathbf{x})$ depends only on the fixed evaluation point $\mathbf{u}$ and its nearest clean neighbor $\mathbf{x}$, and is therefore strictly independent of $\tau$, allowing it to be factored outside the expectation $\mathbb{E}_{\tau|\mathbf{u}}[\cdot]$. In the second term, splitting the posterior integral over $\tau \in [0, \delta] \cup [\delta, 1]$ shows that on a discrete data support ($D > 2$), the local mass $\tau \in [0, \delta]$ decays exponentially fast as $\|\mathbf{x} - D_\tau^*(\mathbf{u})\| \propto$

$\exp(-C/b(\tau)^2)$ (Eq. (78)), while the distant high-noise tail $\tau \in [\delta, 1]$ is suppressed by the diverging nearest-neighbor evidence $1/Z_k(\mathbf{u}) = \mathcal{O}(b(t)^{D-2}) \to 0$ (Lemma 4), so the residual vanishes relative to the $\mathcal{O}(1)$ first term. Factoring out the geometric direction $\|\mathbf{u} - \mathbf{x}\|$ from the dominant term isolates the scaling behavior:

$$\Delta \mathbf{v}_{noise} \approx \|\mathbf{u} - \mathbf{x}\| \left| \frac{\dot{b}(t)}{b(t)} \underbrace{\left( b(t)\mathbb{E}_{\tau|\mathbf{u}} \left[ \frac{1}{b(\tau)} \right] - 1 \right)}_{\text{Jensen Gap}} \right|. \tag{75}$$

Near the data manifold, the observation is generated via $\mathbf{u} = a(t)\mathbf{x} + b(t)\boldsymbol{\epsilon}$. As $t \to 0$ and $a(t) \to 1$, the distance shrinks proportionally to the noise level: $\|\mathbf{u} - \mathbf{x}\| = \mathcal{O}(b(t))$, which cancels the $1/b(t)$ factor in Eq. (75) so that the target estimation error $\|f^*(\mathbf{u}) - f_t^*(\mathbf{u})\|$ remains $\mathcal{O}(1)$ in finite dimensions (vanishing only as $\mathcal{O}(k^{-1})$ as $k \to \infty$; note that under a log-normal prior in the floor limit $b(t) \ll b_{\text{floor}}$, $b(t)\mathbb{E}_{\tau|\mathbf{u}}[1/b(\tau)] \to 0$ so the bracket evaluates to 1). Substituting $\|\mathbf{u} - \mathbf{x}\| = \mathcal{O}(b(t))$ reduces the drift error to a product of the schedule derivative and the Jensen Gap:

$$\Delta \mathbf{v}_{noise} \approx \mathcal{O}(\dot{b}(t)) \cdot \left| b(t)\mathbb{E}_{\tau|\mathbf{u}} \left[ \frac{1}{b(\tau)} \right] - 1 \right|. \tag{76}$$

Consequently, the structural stability of autonomous noise prediction is dictated entirely by the schedule derivative $\dot{b}(t)$:

- For a linear schedule ($b(t) \propto t$, as in Flow Matching), $\dot{b}(t)$ is constant. The drift error evaluates to a bounded $\mathcal{O}(1)$ value, meaning autonomous noise prediction on a linear schedule does *not* diverge as $t \to 0$.
- For Variance Preserving SDEs ($b(t) \propto \sqrt{t}$, as in standard DDPM), $\dot{b}(t) \propto t^{-0.5}$. The error diverges as $\mathcal{O}(t^{-0.5})$, rendering the dynamics structurally unstable ($\lim_{t\to 0} \Delta \mathbf{v}_{noise} = \infty$).

**Case 2: Signal Prediction (EDM)**

Target: $\mathbf{x}$ ($c = 1, d = 0$). Using the unified coefficients ($a = 1, \dot{a} = 0, c = 1, d = 0$), the effective gain simplifies exactly to $\nu(t) = -\frac{\dot{b}(t)}{b(t)}$. The drift error is determined by the denoising error:

$$\Delta \mathbf{v}_{signal} \approx \left| \frac{\dot{b}(t)}{b(t)} \right| \|\hat{\mathbf{x}}(\mathbf{u}) - \mathbf{x}_t^*(\mathbf{u})\|. \tag{77}$$

As $t \to 0$, both $b(t)$ and $\|\hat{\mathbf{x}}(\mathbf{u}) - \mathbf{x}_t^*(\mathbf{u})\|$ go to 0. To resolve the $0/0$ indeterminacy, we assume the data manifold is discrete, $\mathcal{X} = \{\mathbf{x}_k\}_{k=1}^N$. Near a specific data point $\mathbf{x}_k$, the optimal conditional denoiser $\mathbf{x}_t^*(\mathbf{u})$ is a softmax-weighted average of the dataset. The error is dominated by the distance to the nearest neighbor $\mathbf{x}_j$, with a weight proportional to the Gaussian likelihood ratio:

$$\|\mathbf{x}_t^*(\mathbf{u}) - \mathbf{x}_k\| \propto \exp\left( -\frac{\|\mathbf{x}_j - \mathbf{x}_k\|^2}{2b(t)^2} \right). \tag{78}$$

This conditional error decays *exponentially* with respect to the inverse variance $1/b(t)^2$. For the noise-agnostic estimator $\hat{\mathbf{x}}(\mathbf{u}) = \int_0^1 \mathbf{x}_\tau^*(\mathbf{u}) p(\tau|\mathbf{u})\, d\tau$ under a non-vanishing prior ($p(0) > 0$), splitting the integral into a local neighborhood $\tau \in [0, \delta]$ and a distant high-noise tail $\tau \in [\delta, 1]$ shows that the local posterior mass inherits this $\mathcal{O}(\exp(-C/b(t)^2))$ exponential decay, while the distant tail $\tau \in [\delta, 1]$ is suppressed by the diverging nearest-neighbor evidence $1/Z_k(\mathbf{u}) = \mathcal{O}(b(t)^{D-1})$ for $b(t) \propto t$ (or $\mathcal{O}(b(t)^{D-2})$ for a flat prior on $v = b^2$; Lemma 4). For $D > 2$ (and especially in high dimensions $D \gg 1$, where $b(t)^{D-1}$ vanishes at an extreme high-degree rate), both contributions decay strictly faster than the $\mathcal{O}(1/t)$ divergence of the sampler gain:

$$\lim_{t\to 0} \left| \frac{\dot{b}(t)}{b(t)} \right| \|\hat{\mathbf{x}}(\mathbf{u}) - \mathbf{x}_t^*(\mathbf{u})\| = 0. \tag{79}$$

Thus, $\lim_{t\to 0} \Delta \mathbf{v}_{signal} = 0$, proving that signal prediction is asymptotically stable for autonomous models on discrete data manifolds (and as shown in Appendix D.4, under a log-normal prior the autonomous ODE $\frac{d\mathbf{u}}{dt} = \frac{\dot{b}(t)}{b(t)}(\mathbf{u} - \hat{\mathbf{x}}(\mathbf{u}))$ remains a contractive attractor to the slightly smoothed fixed point $\mathbf{u}^* \approx \mathbf{x}_k + \boldsymbol{\eta}_k$).

**Case 3: Velocity Prediction (Flow Matching)** Target: $\mathbf{v} = \dot{\mathbf{u}}$ ($c = -1, d = 1$). The denominator of the unified coefficients is $ad - bc = 1$, resulting in a constant gain $\nu(t) = 1$. The error norm is simply the posterior deviation of the velocity field:

$$\Delta\mathbf{v}_{FM} = \|\mathbb{E}_{\tau|\mathbf{u}}[f_\tau^*(\mathbf{u})] - f_t^*(\mathbf{u})\|. \tag{80}$$

Applying the same decomposition $f_\tau^{*,\text{FM}}(\mathbf{u}) = \frac{\mathbf{u} - D_\tau^*(\mathbf{u})}{\tau} = \frac{\mathbf{u}-\mathbf{x}}{\tau} + \frac{\mathbf{x} - D_\tau^*(\mathbf{u})}{\tau}$ as in Case 1 shows that as $t \to 0$, $\|\mathbf{u} - \mathbf{x}\| = \mathcal{O}(t)$ cancels the $\mathcal{O}(1/t)$ scaling of $\mathbb{E}_{\tau|\mathbf{u}}[1/\tau]$, so that the target estimation error evaluates to a bounded $\mathcal{O}(1)$ vector in finite dimensions (matching the $\mathcal{O}(1)$ Transport Correction in Appendix D.3, and vanishing as $\mathcal{O}(k^{-1})$ in high dimensions). Because the sampler gain $\nu(t) = 1$ is unity, this bounded estimation error is never amplified by a singular schedule factor, and the Drift Perturbation Error $\Delta\mathbf{v}_{FM}$ remains strictly $\mathcal{O}(1)$, confirming that velocity parameterization is inherently stable for autonomous generation.

**Synthesis of the Race Condition** To intuitively summarize the mathematical proofs above, the limiting stability of an autonomous parameterization effectively acts as a race condition between the sampler gain and the target estimation error as $t \to 0$:

- **Noise Prediction (DDPM/DDIM):** The conditional target scales inversely with the noise standard deviation ($\boldsymbol{\epsilon}_t^*(\mathbf{u}) \propto 1/b(t)$), while the sampler gain scales with the schedule derivative ($\nu(t) \approx \dot{b}(t)$). Because the distance to the manifold also shrinks at the rate of $\mathcal{O}(b(t))$, the $b(t)$ terms cancel in the target estimation error, leaving $\|f^*(\mathbf{u}) - f_t^*(\mathbf{u})\| = \mathcal{O}(1)$. Multiplying by $\nu(t) \approx \dot{b}(t)$ causes the drift error to scale directly with the schedule derivative $\mathcal{O}(\dot{b}(t))$, amplifying the finite *"Jensen Gap"* defined in Eq. (75). This results in an $\mathcal{O}(t^{-0.5})$ divergence for VP schedules, causing generation to fail.
- **Signal Prediction (EDM):** The gain contains a singularity ($\nu(t) \propto \dot{b}(t)/b(t)$). However, the error in the signal estimator vanishes rapidly near the discrete data manifold (exponentially on the local posterior mass and as $\mathcal{O}(b(t)^{D-1})$ on the distant tail), counteracting the $\mathcal{O}(1/t)$ divergence of the gain and resulting in a stable contractive flow ($\lim \Delta\mathbf{v} \to 0$).
- **Velocity Prediction (Flow Matching):** The update is identity-mapped with a unit gain ($\nu(t) = 1$). There are no singular sampler coefficients to amplify the $\mathcal{O}(1)$ target estimation error. The dynamics absorb posterior uncertainty into a bounded effective drift, making this parameterization inherently stable.

# G Experimental Details and Hyperparameters

To ensure exact reproducibility of our empirical results, we detail the training and sampling configurations in Table 3. The CIFAR-10 and ImageNet-64 experiments were implemented using the Hackable Diffusion library, while the Toy 2D experiments were implemented independently using JAX and Flax to allow for custom geometric conditioning.

**Network Architecture:** For the Toy 2D experiments, we used a custom Wide-ResNet architecture consisting of 12 residual blocks with a hidden dimension of 2048. Each block applies LayerNorm, followed by a FiLM projection and Swish activations. For both CIFAR-10 and ImageNet-64, we utilized a U-Net backbone with RMSNorm and GELU activations. The CIFAR-10 model employs 128 base channels with channel multipliers of (1, 2, 2, 2) and 2 residual blocks per resolution. The ImageNet-64 model scales this up to 128 base channels, multipliers of (1, 2, 3, 4), and 3 residual blocks per resolution.

**Conditioning Mechanisms:** To isolate the effect of time conditioning, we maintained identical network architectures across all runs. Conditional models utilized standard sinusoidal time embeddings, whereas for the unconditional ("blind") models, this global time embedding was explicitly overridden and fixed to a zero vector. In the Toy 2D experiments, because standard MLP/ResNet layers do not naturally compute global Euclidean norms, the network was additionally supplied with a purely input-derived "geometric clock"—computed strictly as the scaled vector norm $\|\mathbf{u}\|$ of the noisy observation with zero access to the true time $t$—which was concatenated with the (zeroed) time embedding prior to the FiLM projection. As established in Section 5.2, the input norm in high-dimensional regimes serves as an implicit spatial proxy for the noise level; isolating and providing this input feature directly allowed us to empirically validate our concentration theory. Finally, class labels for the image datasets were integrated via an adaptive normalization mechanism.

Table 3: Hyperparameters for empirical evaluations on Toy 2D, CIFAR-10, and ImageNet-64.

| **Hyperparameter** | **Toy 2D** | **CIFAR-10** | **ImageNet-64** |
|---|---|---|---|
| Batch Size | 1024 | 1024 | 512 |
| Learning Rate | $4 \times 10^{-4}$ (Cosine) | $1 \times 10^{-4}$ | $3 \times 10^{-5}$ |
| Optimizer | AdamW | AdamW | AdamW |
| Training Steps | 75k | 200k | 1.5m |
| EMA Decay | 0.9999 | 0.9999 | 0.9999 |
| Integrator | DDIM | DDIM | DDIM |
| NFE (Sampling) | 100 | 50 | 100 |
| $t_{min}$ cutoff | 0.0 | $10^{-4}$ | $10^{-4}$ |
| $t_{max}$ cutoff | 1.0 | 1.0 | 1.0 |

**Training and Optimization:** All models were trained using the AdamW optimizer with a global gradient clipping norm of 1.0. The Toy 2D models were trained for 75,000 steps using a cosine decay learning rate schedule peaking at $4 \times 10^{-4}$ (with 1,000 warmup steps) and a weight decay of $10^{-4}$. The CIFAR-10 models were trained for 200,000 steps with a constant learning rate of $10^{-4}$, while the ImageNet-64 models were trained for 1,500,000 steps with a learning rate of $3 \times 10^{-5}$. During training, we maintained an exponential moving average (EMA) of the network weights with a decay rate of 0.9999 across all experiments.

As experiments are run on small datasets with small models, we do not need a lot of compute resources to train the models. All experiments in this paper are done using 4 Google TPUv4's.

**Sampling:** For the Toy 2D Flow Matching (FM) models, we used a standard Euler ODE solver integrating from $t = 1.0$ to $t = 0.0$ over 100 steps. For the image datasets, we integrated the probability flow ODE using a deterministic DDIM step (stochastic coefficient set to 0.0), acting as a standard Euler ODE solver. We applied a uniform time schedule with a minimum time cutoff ($\epsilon = 10^{-4}$) to prevent numerical instability at the boundary near the clean data manifold. FID metrics are computed over 50,000 generated samples.

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are discussed in detail and supported by the findings of the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Main limitations such as generalization vs memorization, or instability analysis being simply a necessary but not sufficient condition are clearly mentioned and discussed.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All the assumptions and proofs of the reuslts are provided in detail in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: While the main contribution of the paper is theoretical, for the experiments, the libraries used and the main parameters to reproduce the results are clearly reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [No]

Justification: The main contribution of the paper is theoretical and experiments are all run on open standard datasets such as CIFAR-10 and ImageNet-64. While all the details of the implementation are reported, the code is not yet released, but it should be very easy to implement.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: The experiments that are run are small experiments for which all the hyperparameters needed to reproduce the results are reported.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No]

Justification: While the paper does report some metrics, those are shown to validate the theoretical contribution of the paper. The metrics reported are orders of magnitude different which clearly already support the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: As experiments are run on very small networks and small datasets such as ImageNet-64, we used very small compute resources and have reported them in the Appendix.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conform to NeurIPS's code of ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The paper's subject has no societal impact concerns.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper looks at a purely theoretical problem.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All libraries and datasets are clearly mentioned and cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release any new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Not appliable to this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: Not applicable to this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were used for formatting, verification and help with sections of the paper as it is standard practice.

    Guidelines:

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