An Elementary Approach to Scheduling in Generative Diffusion Models

Qiang Sun[1], H. Vincent Poor[2] and Wenyi Zhang[1]

[1]Department of Electronic Engineering and Information Science,

University of Science and Technology of China, Hefei, China

[2]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA

Email: qiangsun@mail.ustc.edu.cn, poor@princeton.edu and wenyizha@ustc.edu.cn

Abstract

An elementary approach to characterizing the impact of noise scheduling and time discretization in generative diffusion models is developed. We first utilize the Cramér-Rao bound to identify the Gaussian case as a fundamental performance limit, necessitating its study as a reference. Building on this insight, we consider a simplified model where the source distribution is multivariate Gaussian with a given covariance matrix in conjunction with the deterministic reverse sampling process. The explicit closed-form evolution trajectory of the distributions across reverse sampling steps is derived, and consequently, the Kullback-Leibler (KL) divergence between the source distribution and the reverse sampling output is obtained. The effect of the number of time discretization steps on the convergence of this KL divergence is studied via the Euler-Maclaurin expansion. An optimization problem is formulated, and its solution noise schedule is obtained via calculus of variations, shown to follow a tangent law whose coefficient is determined by the eigenvalues of the source covariance matrix. For an alternative scenario, more realistic in practice, where pretrained models have been obtained for some given noise schedules, the KL divergence also provides a measure to compare different time discretization strategies in reverse sampling. Experiments across different datasets and pretrained models demonstrate that the time discretization strategy selected by our approach consistently outperforms baseline and search-based strategies, particularly when the budget on the number of function evaluations is very tight.

I. Introduction

Diffusion models (DMs) [1]–[3] have established themselves as a dominant force in generative modeling, delivering state-of-the-art results across diverse applications. The fundamental mechanism of DMs involves the learning of the conditional distribution of a reverse sampling process through a predefined noise-adding forward process.

Research on the forward diffusion process typically focuses on optimizing the noise schedule, prediction target, and network architecture to yield a superior estimator. While recent studies have identified effective configurations via empirical testing or Fisher information analysis [4]–[8], these analyses generally fall short of directly assessing the impact of the noise schedule on the accuracy of the distribution generated by the reverse sampling process. Regarding this reverse sampling process, although the paradigm has shifted from stochastic differential equation (SDE) to more efficient deterministic ordinary differential equation (ODE) solvers [9]–[13], the generation procedure still necessitates multiple forward evaluations of a large network, with the cost measured by the number of function evaluations (NFEs), to substantiate its exceptional capabilities.

To improve generation quality under a fixed budget of NFEs, optimizing the time discretization strategy is crucial [5], [9], [10]. Recognizing this, a growing body of research has been conducted [14]–[20]. However, the majority of extant works rely on computationally expensive data-driven retraining or evaluation processes for searching the optimal steps, typically incurring a substantial overhead.

Aiming at understanding from first principles, the interplay between noise scheduling and sampling efficiency, in this work, we first utilize the Girsanov theorem and the Cramér-Rao bound to reveal that the KL divergence between the continuous and discretized reverse sampling processes in the Gaussian setting serves as a strict theoretical lower bound for that of general distributions. This finding validates the Gaussian setting as a necessary proxy for parameter optimization in DMs, providing a strong motivation for leveraging the analytical properties of Gaussians to analyze diffusion dynamics. Consequently, in the remainder of this paper, we adopt an elementary approach focusing on a multivariate Gaussian source with a given covariance matrix. Such a simple model induces a linear form of the optimal posterior estimator, and consequently an explicit closed-form characterization of the evolution trajectory of the distributions across the deterministic reverse sampling steps. Using this closed-form formulation, we can conveniently investigate the Kullback-Leibler (KL) divergence between the source distribution and the reverse sampling output.

We demonstrate two applications of this KL divergence-based analytical tool. In the first application, we study how fast the KL divergence asymptotically vanishes as the number of time discretization steps increases. Adopting the Euler-Maclaurin expansion [21], we obtain the precise asymptotic behavior of the reverse sampling process and identify the component that dominates the KL divergence. These allow us to establish a direct connection between the noise schedule and the sampling accuracy. Based on this, we propose and solve a variational optimization problem that yields a tangent law of the noise schedule, whose coefficient depends upon the eigenvalues of the source covariance matrix.

In the second application, we consider the practically relevant scenario where a DM has already been trained, under some prescribed noise schedule. We can then employ the KL divergence as a measure to compare different time discretization strategies in the reverse sampling process. This provides a principled low-cost approach to selecting efficient time discretization designs, without requiring any additional retraining or search. We test this approach over CIFAR-10 [22] and FFHQ-64 [23] datasets with different model configurations as outlined in [3], [5], and the experimental results confirm that our strategy consistently demonstrates superior performance in comparison to baseline and state-of-the-art search-based strategies, particularly when the budget on NFEs is very tight.

Notation: When there is no ambiguity, we omit the explicit time argument t for time-dependent functions (e.g., writing f for $f(t)$) and denote its discrete-time counterpart $f(t_j)$ by f_{t_j}. Derivatives with respect to t are written as $\dot{f} = \mathrm{d}f/\mathrm{d}t$ and $\ddot{f} = \mathrm{d}^2 f/\mathrm{d}t^2$. Throughout $\log(\cdot)$ represents the natural logarithm.

II. Preliminaries

In this section, we first review the mechanism of DMs, encompassing the forward diffusion process with its noise schedule design, and the reverse sampling process employed for generation.

The forward diffusion process of DMs is characterized by a linear Gaussian perturbation [2], [4]:

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $t \in [0, 1]$ denotes the diffusion time, $\mathbf{x}_0 \sim q_0(\mathbf{x}_0)$ represents the source data sample, and α_t, σ_t parameterize the noise schedule that directly controls the noise-adding process. Commonly adopted noise schedules in DMs are categorized into two types: variance-preserving (VP) and variance-exploding (VE) [3]. The VP schedule is constrained by $\alpha_t^2 + \sigma_t^2 = 1$, with boundary conditions $\lim_{t\to 0} \alpha_t = 1$ and $\lim_{t\to 1} \alpha_t = 0$. In contrast, the VE schedule maintains $\alpha_t \equiv 1$ while satisfying $\lim_{t\to 0} \sigma_t = 0$ and $\lim_{t\to 1} \sigma_t = \sigma_{\max} \gg 1$. Despite these differences, the forward transition kernel for both schedules unifies to:

$$q_{0t}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}), \tag{2}$$

which progressively diffuses the source distribution $q_0(\mathbf{x}_0)$ toward a tractable Gaussian prior as $t \to 1$. Moreover, the trajectory of the marginal distributions governed by (2) can be equivalently described by a SDE [3], [4]:

$$\mathrm{d}\mathbf{x}_t = f(t)\mathbf{x}_t \mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t, \quad \mathbf{x}_0 \sim q_0(\mathbf{x}_0) \tag{3}$$

where $\mathbf{w}_t$ denotes the standard Wiener process. The drift coefficient $f(t)$ and diffusion coefficient $g(t)$ are derived as:

$$f(t) = \frac{\dot{\alpha}}{\alpha}, \quad g(t) = \sqrt{2\sigma \left(\dot{\sigma} - \frac{\dot{\alpha}}{\alpha}\sigma \right)}. \tag{4}$$

According to the classical theory of diffusion processes [3], [24], the corresponding reverse-time SDE flowing from $t = 1$ to 0 is given by:

$$\mathrm{d}\mathbf{x}_t = \left(f(t)\mathbf{x}_t - g^2(t)\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t) \right) \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t, \tag{5}$$

where $\bar{\mathbf{w}}_t$ is a standard Wiener process in the reverse time. Theoretically, when initialized at $\mathbf{x}_1 \sim q_1(\mathbf{x}_1)$, this process reverses the diffusion trajectory and enables sampling from the source distribution $q_0(\mathbf{x}_0)$.

To render the reverse-time SDE tractable, DMs employ a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ to approximate the scaled score function, i.e., $-\sigma_t \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)$, which is optimized via the following mean squared error (MSE) objective:

$$\mathcal{L}_\theta = \mathbb{E}_{t\in[0,1], \mathbf{x}_0 \sim q(\mathbf{x}), \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right\|^2 \right], \tag{6}$$

which implies that the optimal estimator is the conditional expectation, i.e., $\boldsymbol{\epsilon}^*_\theta(\mathbf{x}_t, t) = \mathbb{E}[\boldsymbol{\epsilon}|\mathbf{x}_t]$. Inverting the linear forward relation in (1), then gives the identity:

$$\boldsymbol{\epsilon}^*_\theta(\mathbf{x}_t, t) = \mathbb{E}\left[\frac{\mathbf{x}_t - \alpha_t \mathbf{x}_0}{\sigma_t} \middle| \mathbf{x}_t \right] = \frac{\mathbf{x}_t - \alpha_t \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]}{\sigma_t} = -\sigma_t \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t). \tag{7}$$

where the last equality follows from Tweedie's formula [25]. Thus, learning to predict the noise $\boldsymbol{\epsilon}$ is mathematically equivalent to estimating the posterior mean of the source data $\mathbf{x}_0$.

To derive an implementable numerical scheme, we discretize the reverse-time SDE (5). Consider a time discretization sequence $\{t_i\}_{i=0}^N$ decreasing from $t_N = 1$ to $t_0 = 0$. For the integration step from t_j to t_{j-1} (i.e., the interval $t \in (t_{j-1}, t_j]$), we approximate the score component by freezing the neural estimator at the starting time t_j. This yields the following piecewise SDE approximation [26], [27]:

$$\mathrm{d}\mathbf{x}_t = \left(f(t)\mathbf{x}_t - h(t)\boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_j}, t_j) \right) \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t, \quad \mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_N^2 \mathbf{I}), \tag{8}$$

where the coefficient $h(t)$ is derived as:

$$h(t) = \frac{g^2(t)}{-\sigma(t)} = 2\left(\frac{\dot{\alpha}}{\alpha}\sigma - \dot{\sigma}\right). \tag{9}$$

In practice, the initialization parameter σ_N is set by the noise schedule, with $\sigma_N = 1$ for the VP setting and $\sigma_N = \sigma_{\max}$ for the VE setting. For subsequent theoretical analysis, we denote the path measures (probability laws) induced by the exact reverse-time SDE (5) and the approximate reverse-time SDE (8) as $\mathbb{Q}$ and $\mathbb{P}$, respectively.

While the stochastic scheme above recovers the data distribution, deterministic sampling is often preferred in practice for its stability and consistency. By converting the SDE into its corresponding probability flow ODE (PF-ODE) [3], we obtain the deterministic reverse sampling process [9], [10]. This process inverts the forward perturbation to generate samples from $q_0(\mathbf{x}_0)$ via the following iterative update rule for $j = N, N-1, \ldots, 1$:

$$\hat{\mathbf{x}}_{t_{j-1}} = \frac{\alpha_{t_{j-1}}}{\alpha_{t_j}}\hat{\mathbf{x}}_{t_j} + \left(\sigma_{t_{j-1}} - \frac{\alpha_{t_{j-1}}}{\alpha_{t_j}}\sigma_{t_j}\right)\boldsymbol{\epsilon}_\theta(\hat{\mathbf{x}}_{t_j}, t_j), \tag{10}$$

where the process is initialized with $\hat{\mathbf{x}}_{t_N} \sim \mathcal{N}(\mathbf{0}, \sigma_N^2\mathbf{I})$.

III. Gaussian Reference and Lower Bound

This section provides the theoretical motivation for our focus on the Gaussian setting. Specifically, by utilizing stochastic analysis on the reverse-time SDEs, we demonstrate that the Gaussian case serves as a fundamental lower bound for general distributions.

We first introduce a Gaussian reference process. Let $q_{0,G}(\mathbf{x}_0)$ be a Gaussian distribution constructed to share the identical covariance structure with the underlying data distribution $q_0(\mathbf{x})$. Specifically, denoting the data covariance as $\mathrm{Cov}_{q_0}[\mathbf{x}_0] = \boldsymbol{\Sigma}_{\mathbf{x}}$, we consider the reference distribution $q_{0,G}$ with covariance $\boldsymbol{\Sigma}_{\mathbf{x}}$. Due to the linearity of the forward diffusion governed by (1), this covariance matching is preserved throughout the forward process, i.e. $\mathrm{Cov}_{q_t}[\mathbf{x}_t] = \mathrm{Cov}_{q_{t,G}}[\mathbf{x}_t] = \boldsymbol{\Sigma}_t = \alpha_t^2\boldsymbol{\Sigma}_{\mathbf{x}} + \sigma_t^2\mathbf{I}$. Moreover, the Hessian of the log-density for this Gaussian reference is given directly by:

$$\nabla^2_{\mathbf{x}_t} \log q_{t,G}(\mathbf{x}_t) = -\boldsymbol{\Sigma}_t^{-1}. \tag{11}$$

Analogous to the general case, we define $\mathbb{Q}_G$ and $\mathbb{P}_G$ as the path measures induced by the exact reverse-time SDE and the approximate reverse-time SDE, respectively, operating under the Gaussian source assumption $q_{0,G}$.

Theorem 1. *Under mild regularity assumptions regarding the score matching error and the smoothness of the data distribution and the optimal estimation, the KL divergence of the Gaussian surrogate process serves as a lower bound for the true process:*

$$D_{KL}(\mathbb{Q}\|\mathbb{P}) \geq D_{KL}(\mathbb{Q}_G\|\mathbb{P}_G) \geq D_{KL}(q_{0,G}\|p_{0,G}). \tag{12}$$

Proof. We first aim to prove the inequality $D_{KL}(\mathbb{Q}\|\mathbb{P}) \geq D_{KL}(\mathbb{Q}_G\|\mathbb{P}_G)$. Let Δ denote the gap between the KL divergence of the general process and that of the Gaussian reference process.

Under the assumption that the neural network approximates the optimal score, i.e., $\boldsymbol{\epsilon}^*_\theta(\mathbf{x}_t, t) = -\sigma_t\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)$ given in (7), its Jacobian is given by $\nabla_{\mathbf{x}}\boldsymbol{\epsilon}^*_\theta(\mathbf{x}_t, t) = -\sigma_t\nabla^2_{\mathbf{x}_t}\log q_t(\mathbf{x}_t)$. Specifically for the Gaussian reference

process, this simplifies to a constant Jacobian $\nabla_{\mathbf{x}}\boldsymbol{\epsilon}^*_{\theta,G}(\mathbf{x}_t,t) = \sigma_t\boldsymbol{\Sigma}_t^{-1}$ due to (11). Consequently, using Lemma 3 in Appendix E, the difference decomposes as:

$$\begin{aligned}\Delta &\triangleq D_{KL}(\mathbb{Q}\|\mathbb{P}) - D_{KL}(\mathbb{Q}_G\|\mathbb{P}_G)\\ &= \Delta_{init} + \sum_{j=1}^{N}\frac{1}{4}h_{t_j}^2\delta_j^2\left(\mathbb{E}_{q_{t_j}}\left[\|-\sigma_{t_j}\nabla_{\mathbf{x}}^2\log q_{t_j}(\mathbf{x})\|_F^2\right] - \|\sigma_{t_j}\boldsymbol{\Sigma}_{t_j}^{-1}\|_F^2\right) + \mathcal{R}(\delta).\end{aligned} \tag{13}$$

where the initial term Δ_{init} is defined as:

$$\Delta_{init} = D_{KL}(q_1(\mathbf{x}_1)||\mathcal{N}(\mathbf{0},\sigma_N^2\mathbf{I})) - D_{KL}(q_{1,G}(\mathbf{x}_1)||\mathcal{N}(\mathbf{0},\sigma_N^2\mathbf{I})). \tag{14}$$

For the initial term Δ_{init}, we utilize the decomposition of KL divergence: $D_{KL}(q\|p) = -H(q) + H(q,p)$, where $H(q,p) = -\mathbb{E}_q[\log p]$ is the cross-entropy. Since the prior $p_{\text{prior}} = \mathcal{N}(\mathbf{0},\sigma_N^2\mathbf{I})$ is Gaussian, the cross-entropy term depends solely on the first two moments of q. By construction, the Gaussian reference $q_{1,G}$ shares the identical mean and covariance with the true distribution q_1. Consequently, the cross-entropy terms cancel out:

$$\mathbb{E}_{q_1}[\log p_{\text{prior}}] = \mathbb{E}_{q_{1,G}}[\log p_{\text{prior}}]. \tag{15}$$

The difference Δ_{init} thus simplifies to the difference in differential entropies:

$$\Delta_{\text{init}} = (-H(q_1)) - (-H(q_{1,G})) = H(q_{1,G}) - H(q_1). \tag{16}$$

According to the Maximum Entropy Theorem [28], among all distributions with a fixed covariance matrix, the Gaussian distribution maximizes the differential entropy. Therefore, $H(q_{1,G}) \geq H(q_1)$, implying $\Delta_{\text{init}} \geq 0$.

To determine the sign of Δ, we analyze the Jacobian terms. The difference can be expanded as:

$$\begin{aligned}&\mathbb{E}_{q_{t_j}}\left[\|-\sigma_{t_j}\nabla_{\mathbf{x}}^2\log q_{t_j}(\mathbf{x})\|_F^2\right] - \|\sigma_{t_j}\boldsymbol{\Sigma}_{t_j}^{-1}\|_F^2\\ &= \sigma_{t_j}^2\left(\mathbb{E}_{q_{t_j}}\left[\|\nabla_{\mathbf{x}}^2\log q_{t_j}(\mathbf{x})\|_F^2\right] - \|\boldsymbol{\Sigma}_{t_j}^{-1}\|_F^2\right) &(17)\\ &\geq \sigma_{t_j}^2\left(\left\|\mathbb{E}_{q_{t_j}}[\nabla_{\mathbf{x}}^2\log q_{t_j}(\mathbf{x})]\right\|_F^2 - \|\boldsymbol{\Sigma}_{t_j}^{-1}\|_F^2\right) &(18)\\ &= \sigma_{t_j}^2\left(\|-\mathcal{I}(q_{t_j})\|_F^2 - \|\boldsymbol{\Sigma}_{t_j}^{-1}\|_F^2\right) &(19)\\ &= \sigma_{t_j}^2\left(\|\mathcal{I}(q_{t_j})\|_F^2 - \|\boldsymbol{\Sigma}_{t_j}^{-1}\|_F^2\right)\\ &\geq 0, &(20)\end{aligned}$$

where $\mathcal{I}(q_t) \triangleq \mathbb{E}_{q_t}[-\nabla_{\mathbf{x}}^2\log q_t(\mathbf{x})]$ denotes the Fisher Information Matrix. The inequality in (18) follows from Jensen's inequality applied to the convex function $\|\cdot\|_F^2$. Equality (19) utilizes the identity that the expected Hessian of the log-likelihood is the negative of the Fisher Information Matrix. Finally, the inequality in (20) is guaranteed by the multivariate Cramér-Rao Bound, which states that $\mathcal{I}(q_t) \succeq \boldsymbol{\Sigma}_t^{-1}$ [28], and all the equalities hold if and only if $q_t(\mathbf{x})$ is a also Gaussian distribution.

Substituting this result and $\Delta_{init} \geq 0$ into (13), we obtain $\Delta \geq 0$ for a sufficiently small step size. Furthermore, the second inequality $D_{KL}(\mathbb{Q}_G\|\mathbb{P}_G) \geq D_{KL}(q_{0,G}\|p_{0,G})$ follows directly from the Data Processing Inequality (DPI), as the marginal distribution at $t=0$ is a projection of the path measure. □

Theorem 1 reveals that the Gaussian setting characterizes a theoretical lower bound of the KL divergence for general distributions. This motivates our subsequent analysis of the Gaussian source, serving as a tractable proxy for general distributions, to derive optimal parameter designs for DMs.

IV. KL-Based Analysis on Gaussian setting and Applications

In this section, we present the results derived from our theoretical analysis under the Gaussian setting and the widely used deterministic reverse sampling process given in (10), and demonstrate their applications. For clarity of exposition, we focus on the VP setting; analogous results for the VE setting are briefly described in discussions and remarks.

As the setup given in Section III, we suppose $\mathbf{x}_0 \sim q_{0,G}(\mathbf{x}_0) = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_x})$, so the marginal distribution $q_t(\mathbf{x}_t)$ remains Gaussian:

$$q_{t,G}(\mathbf{x}_t) = \mathcal{N}\left(\mathbf{0}, \alpha_t^2 \mathbf{\Sigma_x} + \sigma_t^2 \mathbf{I}\right). \tag{21}$$

Moreover, in the Gaussian setting, the optimal posterior estimator admits a closed-form solution [29]:

$$\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] = \alpha_t \mathbf{\Sigma_x} \left(\alpha_t^2 \mathbf{\Sigma_x} + \sigma_t^2 \mathbf{I}\right)^{-1} \mathbf{x}_t. \tag{22}$$

Without loss of generality, we assume that the covariance matrix $\mathbf{\Sigma_x}$ is positive definite. Let the eigendecomposition of the covariance matrix be $\mathbf{\Sigma_x} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{\Lambda} = \mathrm{diag}(\mu_1, \ldots, \mu_k)$ denotes the diagonal matrix of eigenvalues.

A. KL Divergence Analysis

We begin by characterizing the evolution trajectory of the distributions across reverse sampling steps, as given by the following lemma.

Lemma 1. *With the initialization $\hat{\mathbf{x}}_{t_N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, time discretization sequence $\{t_i\}_{i=0}^N$, and the optimal estimator given by (7) and (22), the generated sample $\hat{\mathbf{x}}_{t_0}$ produced by the reverse sampling process (10) is distributed as*

$$\hat{p}_{0,G}\left(\hat{\mathbf{x}}_{t_0}\right) = \mathcal{N}\big(\mathbf{0},\, \mathbf{U}\mathbf{M}\mathbf{U}^{-1}\big), \quad \mathbf{M} = \mathrm{diag}(m_1, \ldots, m_k), \tag{23}$$

where the ℓ-th eigenvalue is given by

$$m_\ell = \prod_{j=1}^{N} \left(\frac{\alpha_{t_{j-1}} \alpha_{t_j} \mu_\ell + \sigma_{t_{j-1}} \sigma_{t_j}}{\alpha_{t_j}^2 \mu_\ell + \sigma_{t_j}^2} \right)^2, \quad \ell \in \{1, \ldots, k\}. \tag{24}$$

Proof. See Appendix A. □

By (21), we can rewrite $q_{0,G}(\mathbf{x}_{t_0}) = \mathcal{N}(\mathbf{0}, \mathbf{U}\mathbf{N}\mathbf{U}^{-1})$, where $\mathbf{N} = \mathrm{diag}(n_1, \ldots, n_k)$ with

$$n_\ell \triangleq \alpha_{t_0}^2 \mu_\ell + \sigma_{t_0}^2. \tag{25}$$

According to Lemma 1, the KL divergence is expressible in closed form:

$$D_{\mathrm{KL}}(\hat{p}_{0,G}(\hat{\mathbf{x}}_{t_0}) \,\|\, q_{0,G}(\mathbf{x}_{t_0})) = \frac{1}{2} \sum_{\ell=1}^{k} \left(\frac{m_\ell}{n_\ell} - \log \frac{m_\ell}{n_\ell} - 1 \right), \tag{26}$$

by the standard KL divergence expression between two Gaussian distributions.

For the sake of the subsequent analysis, for each μ_ℓ, we define

$$S_\ell^N \triangleq \sum_{j=1}^{N} \log\left(\frac{\alpha_{t_{j-1}}\alpha_{t_j}\mu_\ell + \sigma_{t_{j-1}}\sigma_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\right), \tag{27}$$

so that, from (24) we have $S_\ell^N = \frac{1}{2}\log m_\ell$.

To further investigate the asymptotic convergence behavior of the KL divergence (26) with respect to the number of time discretization steps N, we present the following lemma derived by the Euler-Maclaurin expansion.

Lemma 2. *Consider S_ℓ^N defined in (27) with the uniform discretization $t_j \triangleq j/N$ on $[0,1]$. Assume the schedule functions α, σ are sufficiently smooth. Define*

$$I_\ell \triangleq \int_0^1 F_\ell(t)\,\mathrm{d}t, \quad F_\ell(t) \triangleq -\frac{\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma}}{\alpha^2\mu_\ell + \sigma^2}. \tag{28}$$

Then, as $N \to \infty$, we obtain

$$S_\ell^N = I_\ell + \frac{E_\ell^1}{N} + \mathcal{O}\left(\frac{1}{N^2}\right), \tag{29}$$

where E_ℓ^1 is independent of N, specifically,

$$E_\ell^1 = -\frac{\mu_\ell}{2}\int_0^1 \frac{(\alpha\dot{\sigma} - \sigma\dot{\alpha})^2}{(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t. \tag{30}$$

Moreover, it holds that

$$I_\ell = \frac{1}{2}\log n_\ell. \tag{31}$$

Proof. See Appendix B. □

Combining the definition (27) with the closed-form solution (31) and the expansion (29) in Lemma 2, we can explicitly express the residual r_ℓ as

$$r_\ell \triangleq S_\ell^N - I_\ell = \frac{1}{2}\log\frac{m_\ell}{n_\ell} = \frac{E_\ell^1}{N} + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{32}$$

This implies that $r_\ell = \mathcal{O}(1/N)$, which justifies the use of the Taylor expansion for large N. Substituting (32) into (26), the KL divergence can be expanded as follows:

$$\begin{aligned} D_{\mathrm{KL}}(\hat{p}_{0,G}(\hat{\mathbf{x}}_{t_0}) \,\|\, q_{0,G}(\mathbf{x}_{t_0})) &= \frac{1}{2}\sum_{\ell=1}^{k}\left(e^{2r_\ell} - 2r_\ell - 1\right) \\ &= \sum_{\ell=1}^{k} r_\ell^2 + \mathcal{O}\left(\sum_{\ell=1}^{k}|r_\ell|^3\right) \\ &= \frac{1}{N^2}\sum_{\ell=1}^{k}(E_\ell^1)^2 + \mathcal{O}\left(\frac{1}{N^3}\right). \end{aligned} \tag{33}$$

Here, the second equality follows from the Taylor expansion $e^x - x - 1 = \frac{1}{2}x^2 + \mathcal{O}(|x|^3)$ with $x = 2r_\ell$, and the last equality results from substituting the leading order term of r_ℓ.

Since the higher-order terms vanish faster as $N \to \infty$, minimizing the KL divergence at a fixed N asymptotically boils down to minimizing the coefficient of the leading $1/N^2$ term. This yields the following objective functional subject to corresponding boundary conditions:

$$\min_{\alpha,\,\sigma}\ \mathcal{L}[\alpha,\sigma] \triangleq \sum_{\ell=1}^{k}(E_\ell^1)^2. \tag{34}$$

Remark (Convergence and Consistency)**.** From (33), we observe that the KL divergence scales as $\mathcal{O}(1/N^2)$. As $N \to \infty$, the eigenvalue mismatch vanishes (i.e., $m_\ell/n_\ell \to 1$), leading to $D_{\mathrm{KL}}(p_{0,G}(\hat{\mathbf{x}}_{t_0})\|q_{0,G}(\mathbf{x}_{t_0})) \to 0$. This verifies that the reverse sampling process in (10) is asymptotically consistent, recovering the exact source distribution in the continuous-time limit, in accordance with the classical theory well known for DMs (see, e.g., [24]).

B. Derivation of Tangent Law

The preceding analysis of the KL divergence reveals a direct connection between the noise schedule and sampling accuracy, culminating in the formulation of the optimization problem (34). However, directly minimizing the objective functional in (34) over the schedule functions $\alpha(t)$ and $\sigma(t)$ is generally intractable. To gain analytical tractability, we decompose the problem and first analyze the optimality condition for a fixed μ_ℓ. This leads us to introduce a specific law of noise scheduling, defined as follows.

Definition 1. For a fixed mode $\mu_\ell > 0$, we define the tangent law schedule via the ratio $\eta(t) \triangleq \sigma(t)/\alpha(t)$ as

$$\eta_\ell(t) = \sqrt{\mu_\ell}\tan\left(\frac{\pi}{2}t\right), \quad t \in [0,1). \tag{35}$$

The following theorem establishes that this tangent law schedule is, in fact, the unique minimizer of the reverse sampling mismatch for the ℓ-th mode.

Theorem 2. *The minimization of the term $(E_\ell^1)^2$ in (34), with the expression of E_ℓ^1 given in (30), is equivalent to minimizing*

$$\min_{\eta} \mathcal{J}_\ell[\eta] \triangleq \int_0^1 \frac{\dot{\eta}^2}{\left(\mu_\ell + \eta^2\right)^2}\,\mathrm{d}t,$$
$$\textit{subject to } \lim_{t\to 0}\eta(t) = 0 \textit{ and } \lim_{t\to 1}\eta(t) = \infty. \tag{36}$$

The unique solution $\eta(\cdot)$ that minimizes the functional $\mathcal{J}_\ell$ is given by the tangent law schedule $\eta_\ell(t)$ in Definition 1.

Proof. Using the quotient rule $\dot{\eta} = (\alpha\dot{\sigma} - \sigma\dot{\alpha})/\alpha^2$ and the relation $\alpha^2\mu_\ell + \sigma^2 = \alpha^2(\mu_\ell + \eta^2)$, (30) can be simplified to

$$\frac{(\alpha\dot{\sigma} - \sigma\dot{\alpha})^2}{(\alpha^2\mu_\ell + \sigma^2)^2} = \frac{\dot{\eta}^2}{(\mu_\ell + \eta^2)^2},$$

and with the VP setting, we have the boundary conditions (36). Thus, minimizing $(E_\ell^1)^2$ is equivalent to minimizing $\mathcal{J}_\ell[\eta]$.

To solve this variational problem, we introduce a change of variables:

$$Q_\ell(t) \triangleq \frac{1}{\sqrt{\mu_\ell}}\arctan\left(\frac{\eta(t)}{\sqrt{\mu_\ell}}\right). \tag{37}$$

Differentiating $Q_\ell(t)$ with respect to t yields

$$\dot{Q}_\ell(t) = \frac{1}{\sqrt{\mu_\ell}} \cdot \frac{1}{1 + (\eta/\sqrt{\mu_\ell})^2} \cdot \frac{\dot{\eta}}{\sqrt{\mu_\ell}} = \frac{\dot{\eta}}{\mu_\ell + \eta^2}. \tag{38}$$

Consequently, the objective functional simplifies to $\mathcal{J}_\ell[\eta] = \int_0^1 \left(\dot{Q}_\ell\right)^2 \mathrm{d}t$. The Euler–Lagrange equation [21] for the Lagrangian function $L(t, Q_\ell, \dot{Q}_\ell) = \dot{Q}_\ell^2$ is

$$\frac{\partial L}{\partial Q_\ell} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{Q}_\ell}\right) = -2\ddot{Q}_\ell = 0, \tag{39}$$

implying that the optimal $Q_\ell(t)$ is linear in t: $Q_\ell(t) = ct + d$.

Given the boundary conditions in (36), the transformed boundary values are $Q_\ell(0) = 0$ and $Q_\ell(1) = \frac{\pi}{2\sqrt{\mu_\ell}}$. Therefore, the optimal trajectory is $Q_\ell(t) = \frac{\pi t}{2\sqrt{\mu_\ell}}$. Finally, combining the definition of $Q_\ell(t)$ (37) yields the tangent law schedule $\eta_\ell(t) = \sqrt{\mu_\ell}\tan(\frac{\pi}{2}t)$. □

Theorem 2 establishes that the optimal schedule $\eta_\ell(t)$ depends explicitly on the specific eigenvalue μ_ℓ. However, practical diffusion models typically enforce a unified schedule shared across all data dimensions. To reconcile this theoretical optimality with practical scenarios, we introduce a global schedule parameterized by a scalar $\gamma > 0$, adopting the form of tangent law derived in Theorem 2:

$$\eta_\gamma(t) \triangleq \sqrt{\gamma}\,\tan(\frac{\pi}{2}t). \tag{40}$$

Substituting this parameterized tangent law schedule into the leading order coefficient (30) and the global objective functional (34) yields the following choice of γ.

Theorem 3. *Under the parameterized tangent law schedule* (40)*, the global objective functional* (34) *is strictly convex with respect to* γ *and admits a unique closed-form minimizer,*

$$\gamma^\star = \sqrt{\frac{\sum_{\ell=1}^k \mu_\ell}{\sum_{\ell=1}^k \mu_\ell^{-1}}} = \sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma_x})}{\mathrm{tr}(\mathbf{\Sigma_x}^{-1})}}. \tag{41}$$

Proof. As derived in Appendix D-A, the leading order coefficient (30) for the parameterized tangent law schedule (40) is given by

$$E_\ell^1(\gamma) = -\frac{\pi^2}{16}\left(\sqrt{\frac{\mu_\ell}{\gamma}} + \sqrt{\frac{\gamma}{\mu_\ell}}\right). \tag{42}$$

Substituting (42) into $\mathcal{L}(\gamma) = \sum_{\ell=1}^k \left(E_\ell^1(\gamma)\right)^2$ and omitting scaling factors and constant terms, we have that minimizing $\mathcal{L}(\gamma)$ is equivalent to minimizing the function

$$J(\gamma) = \gamma \sum_{\ell=1}^k \mu_\ell^{-1} + \gamma^{-1} \sum_{\ell=1}^k \mu_\ell. \tag{43}$$

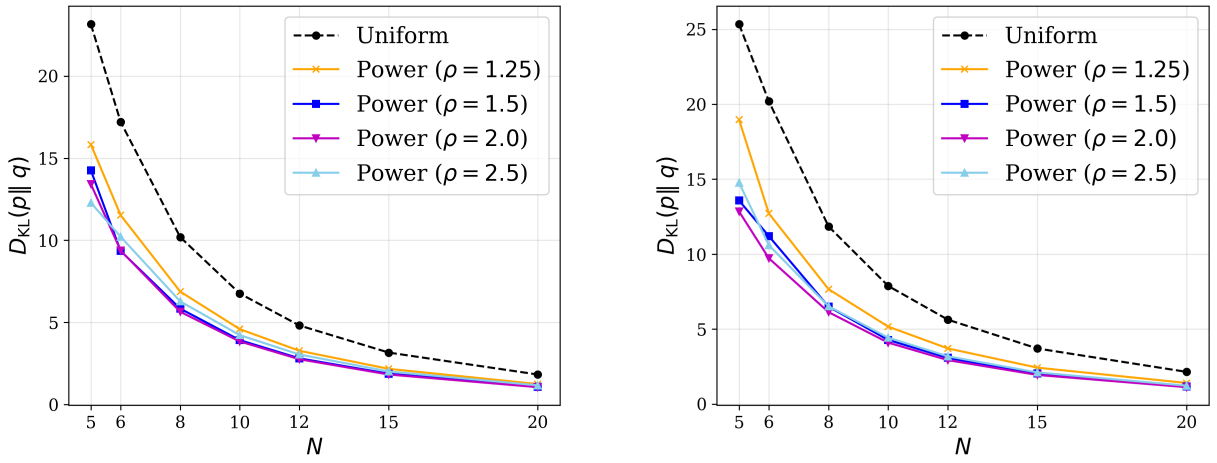
The first and second derivatives of $J(\gamma)$ are

$$J'(\gamma) = \sum_{\ell=1}^k \mu_\ell^{-1} - \gamma^{-2} \sum_{\ell=1}^k \mu_\ell,\; J''(\gamma) = 2\gamma^{-3} \sum_{\ell=1}^k \mu_\ell. \tag{44}$$

Since $\mu_\ell > 0$, we have $J''(\gamma) > 0$ for all $\gamma > 0$, which implies $J(\gamma)$ is strictly convex. Finally, setting $J'(\gamma) = 0$ yields the unique minimizer in (41). □

Remark (Optimization for the VE setting)**.** The derivation for the VE setting parallels the VP setting, but differs in the boundary condition at $t = 1$. This introduces a dependency on the terminal noise level $\sigma_{\max}$ into the objective, which generally precludes a simple closed-form solution for $\gamma^\star$. Nevertheless, in the practical regime of $\sigma_{\max}^2 \gg \mu_\ell$, the boundary effect becomes negligible. Thus, the VP optimizer $\gamma^\star$ in (41) can still provide an accurate approximation (details in Appendix D-A).

Additional numerical results demonstrating that the parameterized tangent law schedule outperforms other noise schedules for the Gaussian setting are presented in Appendix D-B.



(a) VP setting.

(b) VE setting.

Fig. 1: KL divergence versus steps N for the time discretization strategy (45) under (a) VP and (b) VE settings.

C. Time Discretization Strategy Evaluation

In this subsection, we demonstrate another more direct application of the KL divergence-based analysis. Given that training a new DM is prohibitively computationally expensive for large datasets, in practice, users are typically constrained to pretrained models with prescribed noise schedules. Consequently, our focus shifts to designing time discretization strategies, as they crucially influence the generation quality of existing pretrained models. Specifically, we leverage the closed-form KL divergence (26) as a measure to assess different time discretization strategies. This enables us to identify effective time discretization designs without incurring significant computational overhead.

To be specific, let us first parameterize the steps in terms of the half-logSNR $\lambda(t) \triangleq \log(\alpha(t)/\sigma(t))$ and consider a *power-uniform* discretization companded with ρ (using sign-preserving powers $sign(x)|x|^\rho$ when x is negative):

$$\lambda_i^{(\rho)} = \left(\lambda_{t_N}^{1/\rho} + \tfrac{i}{N}\left(\lambda_{t_0}^{1/\rho} - \lambda_{t_N}^{1/\rho}\right)\right)^\rho, \quad i = 0, \ldots, N, \tag{45}$$

which degenerates to the *uniform-*λ rule introduced in [10] when $\rho = 1$. With this, the time discretization steps used in the reverse sampling process (10) are recovered via $t_i = \lambda^{-1}(\lambda_i^{(\rho)})$, which only depends on the prescribed $\alpha(t), \sigma(t)$.

To apply (26), we model the source distribution using a synthetic power-law covariance spectrum, which mimics the heavy-tailed eigenvalue decay characteristic of real image data [30]. Figure 1 presents the quantitative evaluation of the closed-form KL divergence (26) across varying step budgets N (details in Appendix F). This comparison covers both VP and VE settings as considered in [3] with the same time discretization strategy given in (45). The results demonstrate that in both settings, the discretization strategy (45) with an appropriate choice of ρ can achieve notably faster convergence compared to the widely used *uniform-*λ baseline, significantly reducing the reverse sampling mismatch under limited budgets. These preliminary observations motivate further experimentation on real-world image generation tasks, as presented in the next section.

TABLE I: Quantitative comparison of time discretization strategies in terms of FID scores (↓). **Bold** highlights the best result, while underlining marks the second best.

Sampler	Step strategy	NFEs														
		3	4	5	6	7	8	9	10	12	15	20	25	30	50	100
Model: EDM — Dataset: CIFAR-10																
DPM-Solver++	Uniform-λ	120.54	72.33	53.15	37.78	29.43	23.78	19.29	16.29	12.16	8.74	6.01	4.69	3.96	2.86	2.32
	Xue'24	89.59	56.29	37.94	**26.07**	21.65	18.26	18.99	15.87	11.89	8.67	6.00	4.70	3.98	2.86	2.33
	Ours ($\rho = 1.5$)	**82.52**	54.95	**35.93**	26.76	**20.01**	**16.31**	**13.26**	**11.47**	**8.70**	**6.48**	**4.67**	**3.83**	**3.34**	**2.60**	**2.24**
	Ours ($\rho = 2.0$)	86.27	**53.82**	37.26	27.04	20.74	16.75	13.75	11.78	9.04	6.66	4.81	3.92	3.42	2.63	2.25
UniPC	Uniform-λ	119.07	69.04	48.41	32.47	23.77	18.14	13.83	11.13	7.65	5.07	3.38	2.76	2.48	2.16	2.08
	Xue'24	83.91	50.00	29.89	**18.70**	14.49	11.63	13.23	10.48	7.44	5.02	3.38	2.76	2.49	2.16	2.08
	Ours ($\rho = 1.5$)	79.62	49.44	29.50	20.16	14.11	10.78	**8.37**	**6.94**	**5.09**	**3.76**	**2.84**	**2.49**	**2.32**	**2.13**	2.08
	Ours ($\rho = 2.0$)	**78.88**	**45.79**	**28.27**	18.94	**13.84**	**10.72**	8.67	7.22	5.45	4.08	3.06	2.63	2.42	2.16	2.08
Model: EDM — Dataset: FFHQ-64																
DPM-Solver++	Uniform-λ	88.98	63.98	50.69	40.48	32.92	27.63	23.69	20.61	16.24	12.28	8.79	6.97	5.90	4.10	3.11
	Xue'24	89.62	55.70	41.09	**29.80**	25.77	22.13	22.55	20.33	15.99	12.14	8.75	6.98	5.89	4.10	3.12
	Ours ($\rho = 1.5$)	**78.71**	**50.79**	**37.54**	30.46	**24.68**	**20.95**	**17.64**	**15.58**	**12.34**	**9.46**	**6.95**	**5.64**	**4.89**	**3.61**	2.91
	Ours ($\rho = 2.0$)	82.31	54.54	38.60	29.96	24.75	21.18	18.07	15.89	12.63	9.67	7.09	5.75	4.95	3.63	**2.90**
UniPC	Uniform-λ	88.15	61.85	47.19	35.91	27.71	22.04	17.93	14.85	10.77	7.47	5.02	3.98	3.46	2.78	2.53
	Xue'24	83.65	49.87	33.37	22.35	18.38	14.99	16.26	14.25	10.51	7.36	4.99	3.99	3.47	2.79	2.53
	Ours ($\rho = 1.5$)	75.76	**46.25**	31.69	23.76	18.14	14.41	**11.62**	**9.80**	**7.32**	**5.35**	**3.91**	**3.29**	**2.99**	**2.61**	2.49
	Ours ($\rho = 2.0$)	**75.49**	46.84	**30.18**	**21.66**	**17.13**	**14.05**	11.73	10.01	7.68	5.69	4.12	3.42	3.05	**2.61**	**2.48**
Model: VP-SDE — Dataset: CIFAR-10																
DPM-Solver++	Uniform-λ	93.66	71.75	49.70	38.16	30.38	24.91	20.93	17.91	13.95	10.39	7.54	6.07	5.29	3.96	3.24
	Xue'24	52.80	38.45	**24.85**	**19.13**	**16.61**	18.02	20.47	15.11	12.70	10.28	7.51	6.05	5.28	3.95	3.26
	Ours ($\rho = 1.5$)	53.56	41.23	33.83	25.29	20.69	16.59	14.19	12.25	9.75	7.54	5.77	4.86	4.32	3.50	3.06
	Ours ($\rho = 2.0$)	**51.87**	**32.27**	28.98	23.09	19.82	**16.28**	**14.12**	**12.08**	**9.57**	**7.51**	**5.70**	**4.79**	**4.29**	**3.46**	**3.03**
UniPC	Uniform-λ	92.24	68.91	45.57	33.01	24.71	19.12	15.23	12.42	9.03	6.35	4.56	3.80	3.46	3.01	2.85
	Xue'24	49.77	36.07	**21.33**	**14.94**	**12.39**	12.21	14.55	9.51	7.98	6.28	4.55	3.80	3.47	3.01	2.85
	Ours ($\rho = 1.5$)	50.91	38.36	28.49	19.75	14.77	11.14	9.00	7.55	**5.76**	**4.45**	**3.56**	3.19	3.04	2.86	2.81
	Ours ($\rho = 2.0$)	**43.55**	**28.43**	22.24	17.02	13.48	**10.70**	**8.89**	**7.46**	**5.76**	4.51	**3.56**	**3.17**	**3.01**	**2.82**	**2.80**

V. Experiments

Focusing on the practical image generation tasks using pretrained DMs, we evaluate the performance of our selected time discretization strategies in conjunction with off-the-shelf samplers. Experiments are conducted on CIFAR-10 and FFHQ-64 datasets using official model checkpoints from both EDM [5] and VP-SDE [3] architectures. We integrate our time discretization strategies with representative ODE samplers, specifically DPM-Solver++ [11] and UniPC [12], whose first-order update formulas correspond exactly to (10), and report the Fréchet Inception Distance (FID) score [31], a universally adopted metric in image generation tasks that quantifies the Wasserstein distance between the feature distributions of real images and 50k generated samples.

Comparison Schemes. We benchmark against two representative strategies with their officially available implementations: (i) *Uniform-λ* [10], the widely adopted heuristic default for ODE samplers; (ii) *Xue'24* [17], a

state-of-the-art approach that attempts to balance between optimization overhead and generation performance for searching optimal steps. (iii) *Ours*, the *power-uniform* schemes (45) with $\rho = 1.5$ and 2.0.

The quantitative results are summarized in Table I. As shown, our approach consistently outperforms the default baseline by a significant margin, particularly under limited NFE budgets. Furthermore, in the majority of cases, it surpasses the competing search-based method. Notably, while the search-based method occasionally achieves competitive results in few-step regimes, such an edge diminishes significantly as the number of steps increases. Conversely, our approach maintains consistent superiority across all budgets without incurring any costly overhead, validating its effectiveness across varying model architectures and datasets.

VI. Conclusion

This work has presented a first-principles analysis of diffusion sampling, establishing an explicit connection between the noise schedule and sampling accuracy. Furthermore, it has demonstrated that the optimal schedule can be derived analytically for the Gaussian case, exhibiting a tangent law whose coefficient is determined by the eigenvalues of the source covariance matrix. Bridging theory and practice, we have shown the practical merit of our theoretical framework through relevant applications, illustrating how it serves as a useful tool for efficiently identifying improved time discretization strategies to enhance generation quality, compared to existing baselines. Of interest for future work in this area is to extend our theoretical framework to support guidance sampling and high-resolution generation tasks.

References

[1] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37. PMLR, Jul. 2015, pp. 2256–2265.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[3] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[4] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 21 696–21 707.

[5] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 26 565–26 577.

[6] T. Hang, S. Gu, J. Bao, F. Wei, D. Chen, X. Geng, and B. Guo, "Improved noise schedule for diffusion training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2025, pp. 4796–4806.

[7] J. E. Santos and Y. T. Lin, "Using Ornstein-Uhlenbeck process to understand denoising diffusion probabilistic model and its noise schedules," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2311.17673

[8] L. Zhang and S. Syed, "The cosine schedule is Fisher-Rao-optimal for masked discrete diffusion models," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2508.04884

[9] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[10] C. Lu, Y. Zhou, F. Bao, J. Chen, C. LI, and J. Zhu, "DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 5775–5787.

[11] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models," *Machine Intelligence Research*, vol. 22, no. 4, pp. 730–751, 2025.

[12] W. Zhao, L. Bai, Y. Rao, J. Zhou, and J. Lu, "UniPC: A unified predictor-corrector framework for fast sampling of diffusion models," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 49 842–49 869.

[13] K. Zheng, C. Lu, J. Chen, and J. Zhu, "DPM-Solver-v3: Improved diffusion ODE solver with empirical model statistics," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 55 502–55 542.

[14] Y. Wang, X. Wang, A.-D. Dinh, B. Du, and C. Xu, "Learning to schedule in diffusion probabilistic models," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 2478–2488.

[15] L. Li, H. Li, X. Zheng, J. Wu, X. Xiao, R. Wang, M. Zheng, X. Pan, F. Chao, and R. Ji, "Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 7105–7114.

[16] A. Sabour, S. Fidler, and K. Kreis, "Align your steps: Optimizing sampling schedules in diffusion models," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235. PMLR, Jul. 2024, pp. 42 947–42 975.

[17] S. Xue, Z. Liu, F. Chen, S. Zhang, T. Hu, E. Xie, and Z. Li, "Accelerating diffusion sampling with optimized time steps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 8292–8301.

[18] C. Williams, A. Campbell, A. Doucet, and S. Syed, "Score-optimal diffusion schedules," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 107 960–107 983.

[19] S. Xu, Y. Liu, and A. W.-K. Kong, "Variance-reduction guidance: Sampling trajectory optimization for diffusion models," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2025, pp. 1–6.

[20] A. Zhu, R. Su, Q. Zhao, L. Feng, M. Shen, and S. He, "Hierarchical schedule optimization for fast and robust diffusion model sampling," arXiv, 2025. [Online]. Available: https://arxiv.org/abs/2511.11688

[21] D. Zwillinger, *CRC Standard Mathematical Tables and Formulae*. Chapman and Hall/CRC, 2002.

[22] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Toronto, Ontario, Tech. Rep., 2009.

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[24] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[25] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011.

[26] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang, "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions," *arXiv preprint arXiv:2209.11215*, 2022.

[27] X. Liu, L. Wu, M. Ye, and Q. Liu, "Let us build bridges: Understanding and extending diffusion generative models," *arXiv preprint arXiv:2208.14699*, 2022.

[28] T. M. Cover, *Elements of Information Theory*. John Wiley & Sons, 1999.

[29] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer Science & Business Media, 2013.

[30] A. Torralba and A. Oliva, "Statistics of natural image categories," *Network: Computation in Neural Systems*, vol. 14, no. 3, pp. 391–412, 2003.

[31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6626–6637.

[32] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. Academic Press, 2014.

[33] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *Proceedings of the 11th International Conference on Learning Representations*, 2023.

[34] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. PMLR, Jul. 2021, pp. 8162–8171.

[35] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., ser. Universitext. Berlin, Heidelberg: Springer Berlin, Heidelberg, 2003.

APPENDIX A
PROOF OF LEMMA 1

This proof shows that, under the optimal estimator given by (7) and (22), the explicit closed-form evolution trajectory of the distributions across reverse sampling steps can be derived.

Firstly, substituting (7) into the deterministic reverse sampling process (10) yields

$$\begin{aligned}\hat{\mathbf{x}}_{t_{j-1}} &= \frac{\alpha_{t_{j-1}}}{\alpha_{t_j}}\hat{\mathbf{x}}_{t_j} + \left(\sigma_{t_{j-1}} - \frac{\alpha_{t_{j-1}}}{\alpha_{t_j}}\sigma_{t_j}\right)\frac{\hat{\mathbf{x}}_{t_j} - \alpha_{t_j}\mathbb{E}[\mathbf{x}_0|\hat{\mathbf{x}}_{t_j}]}{\sigma_{t_i}} \\ &= \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\hat{\mathbf{x}}_{t_j} + \left(\alpha_{t_{j-1}} - \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\alpha_{t_j}\right)\mathbb{E}[\mathbf{x}_0|\hat{\mathbf{x}}_{t_j}].\end{aligned} \tag{46}$$

Under the Gaussian setup, the optimal posterior estimator (22) gives

$$\mathbb{E}[\mathbf{x}_0|\hat{\mathbf{x}}_{t_j}] = \alpha_{t_j}\boldsymbol{\Sigma}_{\mathbf{x}}\left(\alpha_{t_j}^2\boldsymbol{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right)^{-1}\hat{\mathbf{x}}_{t_j}. \tag{47}$$

Plugging (47) into (46) yields the linear update

$$\begin{aligned}\hat{\mathbf{x}}_{t_{j-1}} &= \left(\frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\mathbf{I} + \left(\alpha_{t_{j-1}} - \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\alpha_{t_j}\right)\alpha_{t_j}\boldsymbol{\Sigma}_{\mathbf{x}}\left(\alpha_{t_j}^2\boldsymbol{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right)^{-1}\right)\hat{\mathbf{x}}_{t_j} \\ &= \left(\frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\left(\alpha_{t_j}^2\boldsymbol{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right) + \left(\alpha_{t_{j-1}}\alpha_{t_j}\boldsymbol{\Sigma}_{\mathbf{x}} - \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\alpha_{t_j}^2\boldsymbol{\Sigma}_{\mathbf{x}}\right)\right)\left(\alpha_{t_j}^2\boldsymbol{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right)^{-1}\hat{\mathbf{x}}_{t_j} \\ &= \left(\sigma_{t_{j-1}}\sigma_{t_j}\mathbf{I} + \alpha_{t_{j-1}}\alpha_{t_j}\boldsymbol{\Sigma}_{\mathbf{x}}\right)\left(\alpha_{t_j}^2\boldsymbol{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right)^{-1}\hat{\mathbf{x}}_{t_j}.\end{aligned} \tag{48}$$

Since $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$ with $\boldsymbol{\Lambda} = \mathrm{diag}(\mu_1, \ldots, \mu_k)$, we have $(\alpha_{t_j}^2\boldsymbol{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I})^{-1} = \mathbf{U}(\alpha_{t_j}^2\boldsymbol{\Lambda} + \sigma_{t_j}^2\mathbf{I})^{-1}\mathbf{U}^{-1}$. Then the updata formula (48) becomes

$$\begin{aligned}\hat{\mathbf{x}}_{t_{j-1}} &= \mathbf{U}\left(\sigma_{t_{j-1}}\sigma_{t_j}\mathbf{I} + \alpha_{t_{j-1}}\alpha_{t_j}\boldsymbol{\Lambda}\right)\mathbf{U}^{-1}\mathbf{U}(\alpha_{t_j}^2\boldsymbol{\Lambda} + \sigma_{t_j}^2\mathbf{I})^{-1}\mathbf{U}^{-1}\hat{\mathbf{x}}_{t_j} \\ &= \mathbf{U}\left(\sigma_{t_{j-1}}\sigma_{t_j}\mathbf{I} + \alpha_{t_{j-1}}\alpha_{t_j}\boldsymbol{\Lambda}\right)(\alpha_{t_j}^2\boldsymbol{\Lambda} + \sigma_{t_j}^2\mathbf{I})^{-1}\mathbf{U}^{-1}\hat{\mathbf{x}}_{t_j} \\ &\triangleq \mathbf{U}\mathbf{D}_j\mathbf{U}^{-1}\hat{\mathbf{x}}_{t_j},\end{aligned} \tag{49}$$

where the diagonal matrix $\mathbf{D}_{t_j}$ is

$$\mathbf{D}_j = \left(\alpha_{t_{j-1}}\alpha_{t_j}\boldsymbol{\Lambda} + \sigma_{t_{j-1}}\sigma_{t_j}\mathbf{I}\right)\left(\alpha_{t_j}^2\boldsymbol{\Lambda} + \sigma_{t_j}^2\mathbf{I}\right)^{-1} = \mathrm{diag}(d_{j,1}, \ldots, d_{j,k}),$$

with the diagonal entries

$$d_{j,\ell} = \frac{\alpha_{t_{j-1}}\alpha_{t_j}\mu_\ell + \sigma_{t_{j-1}}\sigma_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}, \qquad \ell = 1, \ldots, k. \tag{50}$$

Then, using the diagonalized linear update (49) repeatedly with the time discretization sequence $\{t_i\}_{i=0}^N$ yields

$$\hat{\mathbf{x}}_{t_0} = \mathbf{U}\left(\prod_{j=1}^{N}\mathbf{D}_j\right)\mathbf{U}^{-1}\hat{\mathbf{x}}_{t_N}, \tag{51}$$

By (21), in practical reverse sampling, the reverse process is initialized from the terminal prior:

$$\hat{\mathbf{x}}_{t_N} \sim \mathcal{N}(\mathbf{0}, \sigma_N^2\mathbf{I}), \quad \text{where} \quad \sigma_N = \begin{cases} 1 & \text{VP setting,} \\ \sigma_{\max} & \text{VE setting.} \end{cases} \tag{52}$$

Since the update (51) is affine transformation on a Gaussian distribution, $\hat{\mathbf{x}}_{t_0}$ remains Gaussian:

$$\hat{p}_{0,G}(\hat{\mathbf{x}}_{t_0}) = \mathcal{N}\big(\mathbf{0},\, \mathbf{U}\mathbf{M}\mathbf{U}^{-1}\big)\,, \tag{53}$$

where the diagonal matrix $\mathbf{M}$ is

$$\mathbf{M} \triangleq \sigma_N^2 \left(\prod_{j=1}^{N} \mathbf{D}_j\right) \left(\prod_{j=1}^{N} \mathbf{D}_j\right)^{\top} = \mathrm{diag}(m_1, \ldots, m_k), \tag{54}$$

with its diagonal entries

$$m_\ell = \sigma_N^2 \prod_{j=1}^{N} d_{j,\ell}^2, \qquad \ell = 1, \ldots, k, \tag{55}$$

where $d_{j,\ell}$ is given in (50). This yields the result stated in Lemma 1.

The KL divergence between two zero-mean Gaussians is

$$D_{\mathrm{KL}}(\mathcal{N}(0, \Sigma_p) \| \mathcal{N}(0, \Sigma_q)) = \frac{1}{2}\left(\mathrm{tr}(\mathbf{S}) - \log\det(\mathbf{S}) - k\right),$$

where $\mathbf{S} = \Sigma_q^{-1}\Sigma_p$. Finally, combining the distribution $q(\mathbf{x}_{t_0}) = \mathcal{N}(\mathbf{0}, \mathbf{U}\mathbf{N}\mathbf{U}^{-1})$ with $\mathbf{N} = \mathrm{diag}(n_1, \ldots, n_k)$, we can immediately obtain

$$\begin{aligned} D_{\mathrm{KL}}(\hat{p}_{0,G}(\hat{\mathbf{x}}_{t_0}) \,\|\, q_{0,G}(\mathbf{x}_{t_0})) &= D_{\mathrm{KL}}\big(\mathcal{N}\big(\mathbf{0},\, \mathbf{U}\mathbf{M}\mathbf{U}^{-1}\big) \,\|\, \mathcal{N}\big(\mathbf{0},\, \mathbf{U}\mathbf{N}\mathbf{U}^{-1}\big)\big) \\ &= \frac{1}{2}\left(\mathrm{tr}(\mathbf{N}^{-1}\mathbf{M}) - \log\det(\mathbf{N}^{-1}\mathbf{M}) - k\right) \\ &= \frac{1}{2}\sum_{\ell=1}^{k}\left(\frac{m_\ell}{n_\ell} - \log\frac{m_\ell}{n_\ell} - 1\right), \end{aligned} \tag{56}$$

where m_ℓ and n_ℓ are respectively given in (55) and (25).

Appendix B
Proof of Lemma 2

For a fixed mode $\ell \in \{1, \ldots, k\}$, recall

$$S_\ell^N = \sum_{j=1}^{N} \log\left(\frac{\alpha_{t_{j-1}}\alpha_{t_j}\mu_\ell + \sigma_{t_{j-1}}\sigma_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\right). \tag{57}$$

Proposition 1. *Let $t_j \triangleq j/N$ be a uniform grid on $[0,1]$ with step size $h \triangleq 1/N$. Assume that α, σ are sufficiently smooth. Then, the sum S_ℓ^N admits the following expansion as $h \to 0$:*

$$S_\ell^N = \sum_{j=1}^{N}\Big(F_\ell(t_j)\, h + G_\ell(t_j)\, h^2\Big) + \mathcal{O}(h^2), \tag{58}$$

where the coefficients are given by

$$F_\ell(t_j) = -\frac{\alpha_{t_j}\dot{\alpha}_{t_j}\mu_\ell + \sigma_{t_j}\dot{\sigma}_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}, \tag{59}$$

$$G_\ell(t_j) = \frac{\alpha_{t_j}\ddot{\alpha}_{t_j}\mu_\ell + \sigma_{t_j}\ddot{\sigma}_{t_j}}{2\big(\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2\big)} - \frac{\big(\alpha_{t_j}\dot{\alpha}_{t_j}\mu_\ell + \sigma_{t_j}\dot{\sigma}_{t_j}\big)^2}{2\big(\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2\big)^2}. \tag{60}$$

Proof. Using the Taylor expansion at t_j (with $t_{j-1} = t_j - h$), we have

$$\alpha_{t_{j-1}} = \alpha(t_j - h) = \alpha_{t_j} - h\,\dot{\alpha}_{t_j} + \frac{h^2}{2}\,\ddot{\alpha}_{t_j} + \mathcal{O}(h^3),$$
$$\sigma_{t_{j-1}} = \sigma(t_j - h) = \sigma_{t_j} - h\,\dot{\sigma}_{t_j} + \frac{h^2}{2}\,\ddot{\sigma}_{t_j} + \mathcal{O}(h^3), \tag{61}$$

where $\dot{\alpha}_{t_j} \triangleq (\mathrm{d}\alpha/\mathrm{d}t)(t_j)$ and $\ddot{\alpha}_{t_j} \triangleq (\mathrm{d}^2\alpha/\mathrm{d}t^2)(t_j)$, and similarly for σ.

Substituting (61) into the numerator term of the summand in (57), we obtain

$$\begin{aligned}\alpha_{t_{j-1}}\alpha_{t_j}\mu_\ell + \sigma_{t_{j-1}}\sigma_{t_j} &= \left(\alpha_{t_j} - h\,\dot{\alpha}_{t_j} + \frac{h^2}{2}\,\ddot{\alpha}_{t_j}\right)\alpha_{t_j}\mu_\ell + \left(\sigma_{t_j} - h\,\dot{\sigma}_{t_j} + \frac{h^2}{2}\,\ddot{\sigma}_{t_j}\right)\sigma_{t_j} + \mathcal{O}(h^3)\\ &= \alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2 - h\big(\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}\big) + \frac{h^2}{2}\big(\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}\big) + \mathcal{O}(h^3).\end{aligned} \tag{62}$$

Then, using (62), the sum S_ℓ^N becomes

$$\begin{aligned}S_\ell^N &= \sum_{j=1}^{N}\log\left(\frac{\alpha_{t_{j-1}}\alpha_{t_j}\mu_\ell + \sigma_{t_{j-1}}\sigma_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\right)\\ &= \sum_{j=1}^{N}\log\left(1 + \frac{-h\big(\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}\big) + \frac{h^2}{2}\big(\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}\big)}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2} + \mathcal{O}(h^3)\right)\\ &\triangleq \sum_{j=1}^{N}\log\left(1 + \delta_j\right),\end{aligned} \tag{63}$$

where the term δ_j is expanded as

$$\begin{aligned}\delta_j &= \frac{-h\big(\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}\big) + \frac{h^2}{2}\big(\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}\big)}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2} + \mathcal{O}(h^3)\\ &= -\frac{\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\,h + \frac{\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}}{2(\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2)}\,h^2 + \mathcal{O}(h^3)\\ &\triangleq a_j h + b_j h^2 + \mathcal{O}(h^3).\end{aligned} \tag{64}$$

Since $\alpha(\cdot)$ and $\sigma(\cdot)$ are smooth, we have $\delta_j = \mathcal{O}(h)$ as $h \to 0$. Consequently, squaring this expression yields

$$\delta_j^2 = a_j^2h^2 + \mathcal{O}(h^3) = \left(\frac{\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\right)^2 h^2 + \mathcal{O}(h^3). \tag{65}$$

Applying the Taylor expansion $\log(1+x) = x - \frac{x^2}{2} + \mathcal{O}(x^3)$, we get

$$\begin{aligned}\log(1+\delta_j) &= \delta_j - \frac{1}{2}\delta_j^2 + \mathcal{O}(\delta_j^3) = \big(a_jh + b_jh^2\big) - \frac{1}{2}\left(a_j^2h^2\right) + \mathcal{O}(h^3)\\ &= -\frac{\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\,h + \left[\frac{\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}}{2(\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2)} - \frac{\big(\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}\big)^2}{2(\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2)^2}\right]h^2 + \mathcal{O}(h^3)\\ &\triangleq F_\ell(t_j)h + G_\ell(t_j)h^2 + \mathcal{O}(h^3),\end{aligned} \tag{66}$$

where we used the fact that $\mathcal{O}(\delta_j^3) = \mathcal{O}(h^3)$. Finally, substituting (66) into $S_\ell^N = \sum_{j=1}^N \log(1+\delta_j)$, and noting that the summation of local $\mathcal{O}(h^3)$ error terms over $N \propto h^{-1}$ steps results in a global error of $\mathcal{O}(h^2)$, we obtain

$$S_\ell^N = \sum_{j=1}^{N}\Big(F_\ell(t_j)\,h + G_\ell(t_j)\,h^2\Big) + \mathcal{O}(h^2), \tag{67}$$

which concludes the proof. □

A sum of the form $\sum_{j=1}^{N} f(jh)$ with $h = 1/N$ can be approximated by the Euler–Maclaurin formula [21]:

$$\sum_{j=1}^{N} f(jh)\, h = \int_0^1 f(s)\, \mathrm{d}s + \frac{h}{2}\big(f(1) - f(0)\big) + \frac{h^2}{12}\big(f'(1) - f'(0)\big) + R(f,h), \tag{68}$$

where $R(f,h) = \mathcal{O}(h^3)$ as $h \to 0$.

Applying (68) to the Riemann sums in Proposition 1 yields, as $h \to 0$,

$$\begin{aligned} \sum_{j=1}^{N} F_\ell(t_j)\, h &= \int_0^1 F_\ell(t)\, \mathrm{d}t + \frac{h}{2}\big(F_\ell(1) - F_\ell(0)\big) + \mathcal{O}(h^2), \\ \sum_{j=1}^{N} G_\ell(t_j)\, h &= \int_0^1 G_\ell(t)\, \mathrm{d}t + \mathcal{O}(h), \end{aligned} \tag{69}$$

where the function $F_\ell(t)$ and $G_\ell(t)$ are

$$F_\ell(t) = -\frac{\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma}}{\alpha^2\mu_\ell + \sigma^2}, \tag{70}$$

$$G_\ell(t) = \frac{\alpha\ddot{\alpha}\mu_\ell + \sigma\ddot{\sigma}}{2\,(\alpha^2\mu_\ell + \sigma^2)} - \frac{\big(\alpha\dot{\alpha}\mu_\ell + \sigma_{t_j}\dot{\sigma}\big)^2}{2\,(\alpha^2\mu_\ell + \sigma^2)^2}. \tag{71}$$

Recalling the expansion in (58) and substituting the approximations from (69), we obtain

$$\begin{aligned} S_\ell^N &= \sum_{j=1}^{N}\Big(F_\ell(t_j)\, h + G_\ell(t_j)\, h^2\Big) + \mathcal{O}(h^2) \\ &= \sum_{j=1}^{N} F_\ell(t_j)\, h + h\sum_{j=1}^{N} G_\ell(t_j)\, h + \mathcal{O}(h^2) \\ &= \int_0^1 F_\ell(t)\, \mathrm{d}t + \frac{h}{2}\big(F_\ell(1) - F_\ell(0)\big) + h\left(\int_0^1 G_\ell(t)\, \mathrm{d}t + \mathcal{O}(h)\right) + \mathcal{O}(h^2) \\ &= \int_0^1 F_\ell(t)\, \mathrm{d}t + E_\ell^1\, h + \mathcal{O}(h^2), \end{aligned} \tag{72}$$

where the leading order coefficient is given by

$$E_\ell^1 = \frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) + \int_0^1 G_\ell(t)\, \mathrm{d}t. \tag{73}$$

Define the continuous counterpart

$$I_\ell \triangleq \int_0^1 F_\ell(t)\, \mathrm{d}t. \tag{74}$$

Then by (72) and (74), the discretization error $r_\ell \triangleq S_\ell^N - I_\ell$ satisfies

$$r_\ell = \frac{E_\ell^1}{N} + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{75}$$

To explicitly quantify the error, we proceed to derive the analytical forms for the integral term I_ℓ and the leading-order coefficient E_ℓ^1.

A. Closed Form of Integral I_ℓ

Substituting the definition of $F_\ell(t)$ (70) and using the differential relation $\mathrm{d}(\alpha^2\mu_\ell+\sigma^2) = 2(\alpha\dot{\alpha}\mu_\ell+\sigma\dot{\sigma})\,\mathrm{d}t$, the integral admits an analytic closed-form solution:

$$\begin{aligned} I_\ell &= \int_0^1 F_\ell(t)\,\mathrm{d}t \\ &= \int_0^1 -\frac{\alpha\dot{\alpha}\mu_\ell+\sigma\dot{\sigma}}{\alpha^2\mu_\ell+\sigma^2}\,\mathrm{d}t \\ &= -\frac{1}{2}\int_0^1 \frac{\mathrm{d}(\alpha^2\mu_\ell+\sigma^2)}{\alpha^2\mu_\ell+\sigma^2} \\ &= -\frac{1}{2}\log(\alpha^2\mu_\ell+\sigma^2)\Big|_{t=0}^{t=1} \\ &= \frac{1}{2}\log\left(\frac{\alpha^2(0)\mu_\ell+\sigma^2(0)}{\alpha^2(1)\mu_\ell+\sigma^2(1)}\right). \end{aligned} \tag{76}$$

To interpret this result, we first recall that $n_\ell \triangleq \alpha^2(0)\mu_\ell+\sigma^2(0)$, consistent with the notation in (25). We now consider the following two boundary conditions, as we mentioned before in Section II.

- Variance Preserving (VP): The boundary conditions are specified as $\lim_{t\to 0}\alpha(t)=1$, $\lim_{t\to 0}\sigma(t)=0$, and typically $\lim_{t\to 1}\alpha(t)=0$, $\lim_{t\to 1}\sigma(t)=1$. Substituting these into (76) yields
$$I_\ell\big|_{\mathrm{VP}} = \frac{1}{2}\log\left(\frac{n_\ell}{0\cdot\mu_\ell+1}\right) = \frac{1}{2}\log n_\ell. \tag{77}$$
- Variance Exploding (VE): The boundary conditions are $\alpha(t)\equiv 1$, $\lim_{t\to 0}\sigma(t)=0$, and $\lim_{t\to 1}\sigma(t)=\sigma_{\max}$. The integral (76) becomes
$$I_\ell\big|_{\mathrm{VE}} = \frac{1}{2}\log\left(\frac{n_\ell}{\mu_\ell+\sigma_{\max}^2}\right). \tag{78}$$
In the regime where $\sigma_{\max}^2 \gg \mu_\ell$ (i.e., $\mu_\ell/\sigma_{\max}^2 \to 0$), using the expansion $\log(A+x)\approx\log A + x/A$, we recover the asymptotic behavior
$$I_\ell\big|_{\mathrm{VE}} + \log\sigma_{\max} = \frac{1}{2}\log n_\ell + \mathcal{O}\left(\frac{\mu_\ell}{\sigma_{\max}^2}\right). \tag{79}$$

We thus establish the unified relationship

$$\frac{1}{2}\log n_\ell \approx I_\ell + \log\sigma_N, \tag{80}$$

where σ_N is defined in (52). Note that the relation in (80) is exact for the VP setting (where $\sigma_N=1$), and holds asymptotically for the VE setting (where $\sigma_N=\sigma_{\max}$) in the high-noise setting ($\sigma_{\max}^2\gg\mu_\ell$).

Recalling the discrete counterpart from (55) and (57), we have the exact identity

$$\frac{1}{2}\log m_\ell = S_\ell^N + \log\sigma_N. \tag{81}$$

Subtracting (80) from (81), we obtain

$$\frac{1}{2}\log\frac{m_\ell}{n_\ell} \approx S_\ell^N - I_\ell.$$

Consequently, to minimize the KL divergence given in (56), which is determined by the discrepancy between m_ℓ and n_ℓ, it suffices to minimize the discretization error between the Riemann sum S_ℓ^N and the integral I_ℓ.

B. Simplification of Leading Order Coefficient E_ℓ^1

We decompose the integrand $G_\ell(t)$ defined in (71) into two parts, denoted as $A_\ell(t)$ and $B_\ell(t)$:

$$G_\ell(t) = \underbrace{\frac{\alpha\ddot{\alpha}\mu_\ell + \sigma\ddot{\sigma}}{2(\alpha^2\mu_\ell + \sigma^2)}}_{A_\ell(t)} - \underbrace{\frac{(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{2(\alpha^2\mu_\ell + \sigma^2)^2}}_{B_\ell(t)}. \tag{82}$$

Substituting this decomposition into the expression for E_ℓ^1 (73), we obtain

$$\begin{aligned} E_\ell^1 &= \frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) + \int_0^1 G_\ell(t)\,\mathrm{d}t \\ &= \frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) + \int_0^1 A_\ell(t)\,\mathrm{d}t - \int_0^1 B_\ell(t)\,\mathrm{d}t. \end{aligned} \tag{83}$$

Applying integration by parts to $\int A_\ell(t)\,\mathrm{d}t$ yields

$$\begin{aligned} \int_0^1 A_\ell(t)\,\mathrm{d}t &= \int_0^1 \frac{1}{2}\left(\frac{\alpha\mu_\ell}{\alpha^2\mu_\ell + \sigma^2}\ddot{\alpha} + \frac{\sigma}{\alpha^2\mu_\ell + \sigma^2}\ddot{\sigma}\right)\mathrm{d}t \\ &= \frac{1}{2}\int_0^1 \frac{\alpha\mu_\ell}{\alpha^2\mu_\ell + \sigma^2}\,\mathrm{d}(\dot{\alpha}) + \frac{1}{2}\int_0^1 \frac{\sigma}{\alpha^2\mu_\ell + \sigma^2}\,\mathrm{d}(\dot{\sigma}) \\ &= \frac{1}{2}\left[\frac{\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma}}{\alpha^2\mu_\ell + \sigma^2}\right]_0^1 - \frac{1}{2}\int_0^1 \left(\dot{\alpha}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\alpha\mu_\ell}{\alpha^2\mu_\ell + \sigma^2} + \dot{\sigma}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\sigma}{\alpha^2\mu_\ell + \sigma^2}\right)\mathrm{d}t \\ &= -\frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) - \frac{1}{2}\int_0^1 \frac{(\dot{\alpha}^2\mu_\ell + \dot{\sigma}^2)(\alpha^2\mu_\ell + \sigma^2) - 2(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})}{(\alpha^2\mu_\ell + \sigma^2)^2}\mathrm{d}t \\ &= -\frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) - \frac{1}{2}\int_0^1 \frac{(\dot{\alpha}^2\mu_\ell + \dot{\sigma}^2)(\alpha^2\mu_\ell + \sigma^2) - 2(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t, \end{aligned} \tag{84}$$

where we identified the boundary term as $-\frac{1}{2}(F_\ell(1) - F_\ell(0))$ based on the definition of $F_\ell(t)$ in (70).

Substituting the result from (84) back into the expression of E_ℓ^1 given in (83), we obtain

$$\begin{aligned} E_\ell^1 &= \frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) + \int_0^1 A_\ell(t)\,\mathrm{d}t - \int_0^1 B_\ell(t)\,\mathrm{d}t \\ &= -\frac{1}{2}\int_0^1 \frac{(\dot{\alpha}^2\mu_\ell + \dot{\sigma}^2)(\alpha^2\mu_\ell + \sigma^2) - 2(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t - \int_0^1 \frac{(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{2(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t \\ &= -\frac{1}{2}\int_0^1 \frac{(\dot{\alpha}^2\mu_\ell + \dot{\sigma}^2)(\alpha^2\mu_\ell + \sigma^2) - (\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t \\ &= -\frac{\mu_\ell}{2}\int_0^1 \frac{(\alpha\dot{\sigma} - \sigma\dot{\alpha})^2}{(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t. \end{aligned} \tag{85}$$

This final expression confirms that the leading order coefficient E_ℓ^1 is non-positive for any choice of noise schedule $\alpha(t), \sigma(t)$.

APPENDIX C
SUPPLEMENTARY DETAILS OF THEOREM 2

In this appendix, we complement the result of Theorem 2 to the VE setting. The objective is identical to the VP setting shown in (36); the only difference lies in the boundary condition at $t = 1$.

Considering the noise schedule $\alpha(t), \sigma(t)$ for VE settings and the definition $\eta(t) = \sigma(t)/\alpha(t)$, we get the boundary conditions as follows:

$$\eta(0) = 0 \quad \text{and} \quad \lim_{t\to 1}\eta(1) = \sigma_{\max}. \tag{86}$$

Based on the definition $Q_\ell(t) \triangleq \frac{1}{\sqrt{\mu_\ell}} \arctan\left(\frac{\eta(t)}{\sqrt{\mu_\ell}}\right)$ from (37), the transformed boundary conditions become

$$Q_\ell(0) = 0 \quad \text{and} \quad Q_\ell(1) = \frac{1}{\sqrt{\mu_\ell}} \arctan\left(\frac{\sigma_{\max}}{\sqrt{\mu_\ell}}\right).$$

Recall that the Euler–Lagrange equation for the Lagrangian $L = \dot{Q}_\ell^2$ (39) implies that the optimal trajectory is linear $Q_\ell(t) = ct + d$.

By imposing the boundary conditions on this linear solution, we obtain the explicit optimal trajectory in the transformed space:

$$Q_\ell(t) = \frac{t}{\sqrt{\mu_\ell}} \arctan\left(\frac{\sigma_{\max}}{\sqrt{\mu_\ell}}\right). \tag{87}$$

Subsequently, combining (37) and (87), we recover the corresponding tangent law schedule for the VE setting $\eta_\ell(t)\big|_{\text{VE}}$:

$$\eta_\ell(t)\big|_{\text{VE}} = \sqrt{\mu_\ell} \tan\left(t \cdot \arctan\left(\frac{\sigma_{\max}}{\sqrt{\mu_\ell}}\right)\right), \tag{88}$$

so when $\sigma_{\max}^2 \gg \mu_\ell$, we have $\arctan\left(\frac{\sigma_{\max}}{\sqrt{\mu_\ell}}\right) \approx \frac{\pi}{2}$, implying that $\eta_\ell(t)\big|_{\text{VE}}$ approaches the VP solution $\eta_\ell(t)\big|_{\text{VP}}$ (35) given in Definition 1.

APPENDIX D
SUPPLEMENTARY DETAILS OF THEOREM 3

In this appendix, we derive an explicit expression for the leading order coefficient E_ℓ^1 under the parameterized tangent law schedule and present additional numerical results.

A. Derivation of Leading-Order Coefficient

Firstly, by the tangent law (35) and (88) derived in Theorem 2, we redefine the parameterized tangent law schedule for both VP and VE setting

$$\eta_\gamma(t) \triangleq \frac{\sigma_\gamma(t)}{\alpha_\gamma(t)} = \sqrt{\gamma}\, \tan\big(\Theta_\gamma t\big), \quad \gamma > 0, \tag{89}$$

where the factor Θ_γ is determined by the endpoint condition $\eta_\gamma(1) = \eta_1$:

$$\Theta_\gamma \triangleq \arctan\left(\frac{\eta_1}{\sqrt{\gamma}}\right). \tag{90}$$

We specify the boundary values η_1 for the following two settings:

- Variance Preserving (VP): Since $\alpha(1) = 0$ and $\sigma(1) = 1$, we have $\eta_1 \to \infty$. Consequently, the angle becomes:

$$\Theta_\gamma\big|_{\text{VP}} = \frac{\pi}{2}.$$

- Variance Exploding (VE): Since $\alpha(t) \equiv 1$ and $\sigma(1) = \sigma_{\max}$, we have $\eta_1 = \sigma_{\max}$,

$$\Theta_\gamma\big|_{\text{VE}} = \arctan\left(\frac{\sigma_{\max}}{\sqrt{\gamma}}\right) = \frac{\pi}{2} - \arctan\left(\frac{\sqrt{\gamma}}{\sigma_{\max}}\right).$$

Same as we mentioned before in Appendix B-A and Appendix C, under the condition of $\sigma_{\max}^2 \gg \gamma$, the VP setting solution can be regarded as an accurate approximation of the VE setting.

Substituting the parameterized tangent law schedule $\eta_\gamma(t) = \sqrt{\gamma}\tan(\Theta_\gamma t)$ into the expression of E_ℓ^1 given in (85), we proceed as follows:

$$\begin{aligned}
E_\ell^1(\gamma) &= -\frac{\mu_\ell}{2}\int_0^1 \frac{(\alpha\dot{\sigma}-\sigma\dot{\alpha})^2}{(\alpha^2\mu_\ell+\sigma^2)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell}{2}\int_0^1 \frac{\dot{\eta}(t)^2}{(\mu_\ell+\eta(t)^2)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell}{2}\int_0^1 \frac{\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{\gamma}\tan\left(\Theta_\gamma t\right)\right)\right]^2}{\left(\mu_\ell+\gamma\tan^2\left(\Theta_\gamma t\right)\right)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell}{2}\int_0^1 \frac{\gamma\Theta_\gamma^2\sec^4(\Theta_\gamma t)}{\left(\frac{\mu_\ell\cos^2(\Theta_\gamma t)+\gamma\sin^2(\Theta_\gamma t)}{\cos^2(\Theta_\gamma t)}\right)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell\gamma\Theta_\gamma^2}{2}\int_0^1 \frac{\sec^4(\Theta_\gamma t)\cdot\cos^4(\Theta_\gamma t)}{\left(\mu_\ell\cos^2(\Theta_\gamma t)+\gamma\sin^2(\Theta_\gamma t)\right)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell\gamma\Theta_\gamma}{2}\int_0^{\Theta_\gamma} \frac{1}{\left(\mu_\ell\cos^2 x+\gamma\sin^2 x\right)^2}\,\mathrm{d}x \\
&= -\frac{\mu_\ell\gamma\Theta_\gamma}{2}\int_0^{\Theta_\gamma} \frac{1}{\left(\mu_\ell+(\gamma-\mu_\ell)\sin^2 x\right)^2}\,\mathrm{d}x.
\end{aligned} \tag{91}$$

The standard integral result from [32, formula 2.563.1 and 2.562.1 in pp.177] is

$$\begin{aligned}
\int \frac{\mathrm{d}x}{(a+b\sin^2 x)^2} &= \frac{1}{2a(a+b)}\left[(2a+b)\int\frac{\mathrm{d}x}{a+b\sin^2 x}+\frac{b\sin x\cos x}{a+b\sin^2 x}\right] \\
&= \frac{1}{2a(a+b)}\left[\frac{(2a+b)\,\mathrm{sign}(a)}{\sqrt{a(a+b)}}\arctan\left(\sqrt{\frac{a+b}{a}}\tan x\right)+\frac{b\sin x\cos x}{a+b\sin^2 x}\right](b>-a),
\end{aligned} \tag{92}$$

where sign denotes the sign function. Applying this to $E_\ell^1(\gamma)$ (91) with $a=\mu_\ell$ and $b=\gamma-\mu_\ell$, we obtain

$$\begin{aligned}
E_\ell^1(\gamma) &= -\frac{\mu_\ell\gamma\Theta_\gamma}{2}\int_0^{\Theta_\gamma}\frac{1}{\left(\mu_\ell+(\gamma-\mu_\ell)\sin^2 x\right)^2}\,\mathrm{d}x \\
&= -\frac{\mu_\ell\gamma\Theta_\gamma}{2}\cdot\frac{1}{2\mu_\ell\gamma}\left[\frac{\mu_\ell+\gamma}{\sqrt{\mu_\ell\gamma}}\arctan\left(\sqrt{\frac{\gamma}{\mu_\ell}}\tan\Theta_\gamma\right)+\frac{(\gamma-\mu_\ell)\sin\Theta_\gamma\cos\Theta_\gamma}{\mu_\ell\cos^2\Theta_\gamma+\gamma\sin^2\Theta_\gamma}\right] \\
&= -\frac{\Theta_\gamma}{4}\left[\frac{\mu_\ell+\gamma}{\sqrt{\mu_\ell\gamma}}\arctan\left(\sqrt{\frac{\gamma}{\mu_\ell}}\tan\Theta_\gamma\right)+\frac{(\gamma-\mu_\ell)\tan\Theta_\gamma}{\mu_\ell+\gamma\tan^2\Theta_\gamma}\right].
\end{aligned} \tag{93}$$

Finally, in the VP setting where $\Theta_\gamma=\frac{\pi}{2}$, we observe the following limiting behaviors

$$\arctan\left(\sqrt{\frac{\gamma}{\mu_\ell}}\tan\Theta_\gamma\right)\to\frac{\pi}{2},\qquad\frac{(\gamma-\mu_\ell)\tan\Theta_\gamma}{\mu_\ell+\gamma\tan^2\Theta_\gamma}\to 0.$$

Substituting these limits yields the explicit expression

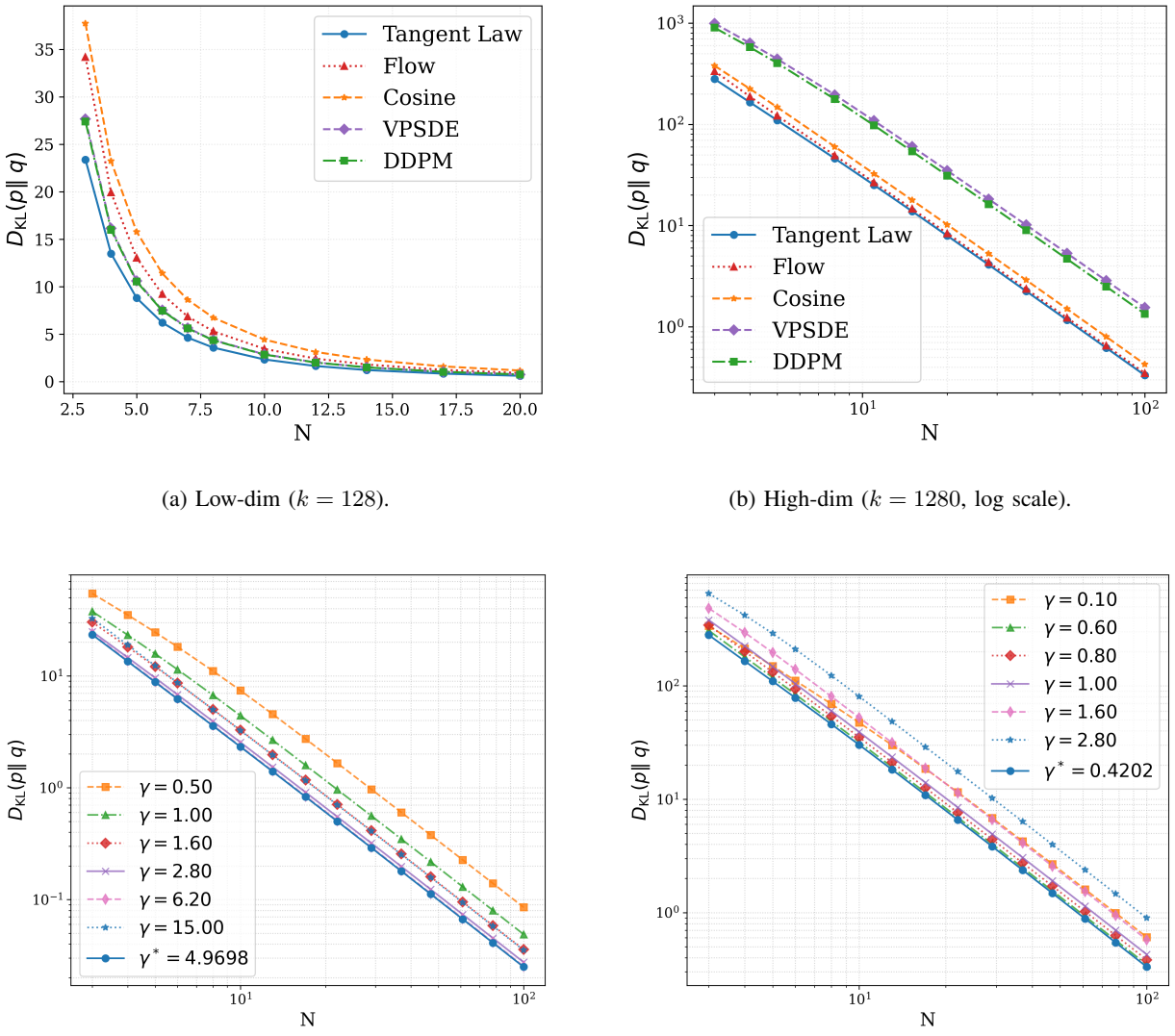
$$E_\ell^1(\gamma)\big|_{\mathrm{VP}} = -\frac{\pi/2}{4}\left(\frac{\mu_\ell+\gamma}{\sqrt{\mu_\ell\gamma}}\cdot\frac{\pi}{2}\right) = -\frac{\pi^2}{16}\left(\sqrt{\frac{\mu_\ell}{\gamma}}+\sqrt{\frac{\gamma}{\mu_\ell}}\right), \tag{94}$$

thereby recovering the expression presented in (42).

B. Numerical Results on Theorem 3

To assess the performance of the parameterized tangent law noise schedule derived in Theorem 3, we conduct a two-fold evaluation based on the closed-form KL divergence (26), with results summarized in Figure 2. First, we compare our parameterized tangent law schedule (using $\gamma^\star$) against representative noise schedules widely employed

in DMs (Figure 2 top row). The results indicate that our method achieves faster convergence in both low-dimensional and high-dimensional synthetic Gaussian sources. Second, we perform an ablation study on the hyperparameter γ (Figure 2 bottom row). The empirical results align with our theoretical analysis, showing that the derived $\gamma^\star$ (41) leads to the best performance compared to other choices.



(a) Low-dim ($k = 128$).

(b) High-dim ($k = 1280$, log scale).

(c) Low-dim ($k = 128$, log scale).

(d) High-dim ($k = 1280$, log scale).

Fig. 2: Performance analysis of the parameterized tangent law schedule in Gaussian settings. We report the KL divergence versus the number of discretization steps N. Top Row: Comparison of our parameterized tangent law schedule (using $\gamma^\star$) against existing noise schedules [2], [3], [33], [34]. Bottom Row: Ablation study of different γ values, verifying the optimality of our derived $\gamma^\star$. The left and right columns correspond to low-dimensional ($k = 128$) and high-dimensional ($k = 1280$) cases, respectively. The results demonstrate the effectiveness of our derived $\gamma^\star$ in Theorem 3.

APPENDIX E
LEMMA OF GIRSANOV'S THEOREM

Lemma 3. *Let $\mathbb{P}_\theta$ denote the path measure induced by the continuous-time SDE parameterized by the neural network $\boldsymbol{\epsilon}_\theta$, and let $\tilde{\mathbb{P}}_\theta$ denote the measure induced by its discretization scheme:*

$$\mathbb{P}_\theta: \quad \mathrm{d}\mathbf{x}_t = \left(f_t\mathbf{x}_t - h_t\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)\mathrm{d}t + g_t\mathrm{d}\bar{\mathbf{w}}_t, \quad \mathbf{x}_1 \sim q_1(\mathbf{x}_1), \tag{95a}$$

$$\tilde{\mathbb{P}}_\theta: \quad \mathrm{d}\mathbf{x}_t = \left(f_t\mathbf{x}_t - h_t\boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_j}, t_j)\right)\mathrm{d}t + g_t\mathrm{d}\bar{\mathbf{w}}_t, \quad \mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_N^2\mathbf{I}). \tag{95b}$$

The KL divergence between these two processes satisfies:

$$D_{KL}(\mathbb{P}_\theta||\tilde{\mathbb{P}}_\theta) = D_{KL}(q_1(\mathbf{x}_1)||\mathcal{N}(\mathbf{0}, \sigma_N^2\mathbf{I})) + \frac{1}{2}\int_0^1 \frac{h_t^2}{g_t^2}\,\mathbb{E}_{\mathbf{x}_t\sim p_t}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_j}, t_j)\|^2\right]\mathrm{d}t \tag{96}$$

$$= D_{KL}(q_1(\mathbf{x}_1)||\mathcal{N}(\mathbf{0}, \sigma_N^2\mathbf{I})) + \sum_{j=1}^{N}\frac{1}{4}h_{t_j}^2\,\delta_j^2\,\mathbb{E}_{\mathbf{x}_{t_j}\sim p_{t_j}}\left[\|\mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_{t_j}, t_j)\|_F^2\right] + \mathcal{R}(\delta), \tag{97}$$

where $\delta_j = t_j - t_{j-1}$ is the step size, $\mathbf{J}_{\boldsymbol{\epsilon}} \triangleq \nabla_{\mathbf{x}}\boldsymbol{\epsilon}_\theta(\mathbf{x}, t)$ denotes the Jacobian matrix of $\boldsymbol{\epsilon}_\theta$, and $\mathcal{R}(\delta)$ represents the cumulative higher-order error terms satisfying $\mathcal{R}(\delta) = \mathcal{O}(\delta^2)$ as $\delta \triangleq \max_j \delta_j \to 0$.

Proof. The proof proceeds in two steps. First, we apply Girsanov's theorem to derive the exact integral form of the KL divergence. Second, we analyze the local discretization error using Itô's calculus to obtain the leading-order expansion.

To facilitate the rigorous derivation of the error bound, we first rely on the following regularity conditions. These assumptions are standard in the theoretical analysis of score-based diffusion models:

- Condition 1 (Coefficients): The SDE coefficients f_t and g_t are Lipschitz continuous with respect to time t and bounded on $[0, 1]$.
- Condition 2 (Neural Network): The score estimator $\boldsymbol{\epsilon}_\theta(\mathbf{x}, t)$ is twice continuously differentiable (C^2) with respect to $\mathbf{x}$ and C^1 with respect to t. Moreover, both the network output and its Jacobian are bounded.
- Condition 3 (Bounded Moments): The drift term of the score evolution (defined later in Eq. (101)) possesses a bounded second moment, i.e., $\sup_{t\in[0,1]}\mathbb{E}[\|\boldsymbol{\phi}(\mathbf{x}_t, t)\|^2] < \infty$.

Let u_t denote the difference between the drift terms of $\mathbb{P}_\theta$ and $\hat{\mathbb{P}}_\theta$, normalized by the diffusion coefficient g_t. We define:

$$u_t \triangleq \frac{-h_t\left(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_j}, t_j)\right)}{g_t}. \tag{98}$$

To ensure the validity of the measure change, we assume Novikov's condition holds, which requires that the expectation of the stochastic exponential is finite:

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^1 \|u_t\|^2\mathrm{d}t\right)\right] = \mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^1 \frac{h_t^2}{g_t^2}\left\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_j}, t_j)\right\|^2\mathrm{d}t\right)\right] < \infty. \tag{99}$$

Under conditions 1 and 2, Girsanov's theorem [35] applies. The KL divergence decomposes into the divergence of the initial distributions and the expected path integral of the squared norm of γ_t:

$$D_{KL}(\mathbb{P}_\theta||\tilde{\mathbb{P}}_\theta) = D_{KL}(q_1||\mathcal{N}(\mathbf{0}, \sigma_N^2\mathbf{I})) + \frac{1}{2}\int_0^1 \mathbb{E}\left[\|u_t\|^2\right]\mathrm{d}t. \tag{100}$$

Substituting the definition of γ_t back into Eq. (100) yields the first equality in (96).

We focus on the expectation term within the interval $t \in (t_{j-1}, t_j]$. Applying Itô's Lemma [35] to $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, its dynamics follow:

$$\begin{aligned}\mathrm{d}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) &= \left(\frac{\partial \boldsymbol{\epsilon}_\theta}{\partial t} + \mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_t, t)\left(f_t \mathbf{x}_t - h_t \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \frac{1}{2} g_t^2 \nabla_{\mathbf{x}}^2 \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) \mathrm{d}t + g_t \mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_t, t) \mathrm{d}\bar{\mathbf{w}}_t \\ &\triangleq \boldsymbol{\phi}(\mathbf{x}_t, t)\mathrm{d}t + g_t \mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_t, t)\mathrm{d}\bar{\mathbf{w}}_t, \end{aligned} \tag{101}$$

where $\mathbf{J}_{\boldsymbol{\epsilon}} = \nabla_{\mathbf{x}} \boldsymbol{\epsilon}_\theta \in \mathbb{R}^{d \times d}$ is the Jacobian matrix and $\nabla_{\mathbf{x}}^2$ denotes the vector Laplacian operator. Integrating from t to t_j:

$$\boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_j}, t_j) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) = \int_t^{t_j} \boldsymbol{\phi}(\mathbf{x}_\tau, \tau)\mathrm{d}\tau + \int_t^{t_j} g_\tau \mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_\tau, \tau)\mathrm{d}\bar{\mathbf{w}}_\tau. \tag{102}$$

Squaring and taking the expectation, we utilize the Itô Isometry property [35], which eliminates the cross-term between the bounded variation drift integral and the martingale diffusion integral:

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_j}, t_j)\|^2\right] = \int_t^{t_j} g_\tau^2 \mathbb{E}\left[\|\mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_\tau, \tau)\|_F^2\right] \mathrm{d}\tau + \mathbb{E}\left[\left\|\int_t^{t_j} \boldsymbol{\phi}(\mathbf{x}_\tau, \tau)\mathrm{d}\tau\right\|^2\right]. \tag{103}$$

Assuming the time discretization is sufficiently fine (i.e., the step size $\delta_j \triangleq t_j - t_{j-1}$ is small enough), we can apply local expansions based on Conditions 1, 2, and 3:

- Using the Lipschitz continuity of the coefficients (Condition 1), we approximate the ratio $\frac{h_t^2}{g_t^2}$ by its value at t_j: $\frac{h_t^2}{g_t^2} = \frac{h_{t_j}^2}{g_{t_j}^2} + \mathcal{O}(t_j - t)$.
- Let $\Phi(\tau) \triangleq g_\tau^2 \mathbb{E}[\|\mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_\tau, \tau)\|_F^2]$. Under Conditions 1 and 2, $\Phi(\tau)$ is Lipschitz continuous. Expanding around t_j: $\Phi(\tau) = \Phi(t_j) + \mathcal{O}(t_j - \tau)$.
- Since the drift $\boldsymbol{\phi}$ has a bounded second moment (Condition 3), applying the Cauchy-Schwarz inequality yields:

$$\mathbb{E}\left[\left\|\int_t^{t_j} \boldsymbol{\phi}(\mathbf{x}_\tau, \tau)\mathrm{d}\tau\right\|^2\right] \leq (t_j - t)\int_t^{t_j} \mathbb{E}\left[\|\boldsymbol{\phi}\|^2\right] \mathrm{d}\tau = \mathcal{O}((t_j - t)^2). \tag{104}$$

Substituting these estimates back into the local error expectation (103):

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_j}, t_j)\|^2\right] &= \int_t^{t_j} \left(\Phi(t_j) + \mathcal{O}(t_j - \tau)\right) \mathrm{d}\tau + \mathcal{O}((t_j - t)^2) \\ &= \Phi(t_j)(t_j - t) + \underbrace{\int_t^{t_j} \mathcal{O}(t_j - \tau)\mathrm{d}\tau}_{\mathcal{O}((t_j - t)^2)} + \mathcal{O}((t_j - t)^2) \\ &= g_{t_j}^2 \mathbb{E}_{\mathbf{x}_{t_j}}[\|\mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_{t_j}, t_j)\|_F^2](t_j - t) + \mathcal{O}((t_j - t)^2). \end{aligned} \tag{105}$$

Now, we substitute Eq. (105) back into the Girsanov integral derived in Step 1. Focusing on the contribution from the j-th interval $(t_{j-1}, t_j]$, the term involving u_t becomes:

$$\begin{aligned}\int_{t_{j-1}}^{t_j} \mathbb{E}\left[\|u_t\|^2\right] \mathrm{d}t &= \int_{t_{j-1}}^{t_j} \frac{h_t^2}{g_t^2} \mathbb{E}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t_j}, t_j)\|^2\right] \mathrm{d}t \\ &= \int_{t_{j-1}}^{t_j} \left(\frac{h_{t_j}^2}{g_{t_j}^2} + \mathcal{O}(t_j - t)\right) \left(g_{t_j}^2 \mathbb{E}[\|\mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_{t_j}, t_j)\|_F^2](t_j - t) + \mathcal{O}((t_j - t)^2)\right) \mathrm{d}t \\ &= \int_{t_{j-1}}^{t_j} \left(h_{t_j}^2 \mathbb{E}[\|\mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_{t_j}, t_j)\|_F^2](t_j - t) + \underbrace{\mathcal{O}((t_j - t)^2)}_{\text{Cross terms \& Higher orders}}\right) \mathrm{d}t. \end{aligned} \tag{106}$$

So we have the integral of the squared drift difference:

$$\begin{aligned}\int_0^1 \mathbb{E}\left[\|u_t\|^2\right] \mathrm{d}t &= \sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} \mathbb{E}\left[\|u_t\|^2\right] \mathrm{d}t \\ &= \sum_{j=1}^{N} \left(h_{t_j}^2 \mathbb{E}[\|\mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_{t_j}, t_j)\|_F^2] \int_{t_{j-1}}^{t_j} (t_j - t)\mathrm{d}t + \int_{t_{j-1}}^{t_j} \mathcal{O}((t_j - t)^2)\mathrm{d}t \right) \\ &= \sum_{j=1}^{N} \left(\frac{1}{2} h_{t_j}^2 \delta_j^2 \mathbb{E}[\|\mathbf{J}_{\boldsymbol{\epsilon}}(\mathbf{x}_{t_j}, t_j)\|_F^2] + \mathcal{O}(\delta_j^3) \right). \end{aligned} \tag{107}$$

Finally, substituting this back into the Girsanov formula (100), and summing over all intervals $j = 1, \ldots, N$:

$$\begin{aligned} D_{KL}(\mathbb{P}_\theta || \tilde{\mathbb{P}}_\theta) &= D_{KL}(q_1 || \mathcal{N}(\mathbf{0}, \sigma_N^2 \mathbf{I})) + \frac{1}{2} \sum_{j=1}^{N} \left(\frac{1}{2} h_{t_j}^2 \delta_j^2 \mathbb{E}[\|\mathbf{J}_{\boldsymbol{\epsilon}}\|_F^2] + \mathcal{O}(\delta_j^3) \right) \\ &= D_{KL}(q_1 || \mathcal{N}(\mathbf{0}, \sigma_N^2 \mathbf{I})) + \sum_{j=1}^{N} \frac{1}{4} h_{t_j}^2 \delta_j^2 \mathbb{E}[\|\mathbf{J}_{\boldsymbol{\epsilon}}\|_F^2] + \mathcal{R}(\delta), \end{aligned} \tag{108}$$

where $\mathcal{R}(\delta)$ represents the higher-order terms, which satisfies $\mathcal{R}(\delta) \triangleq \mathcal{O}((\max_j \delta_j)^2)$ as $\delta = \max_j \delta_j \to 0$. This recovers the exact form stated in (96). □

APPENDIX F
IMPLEMENTATION DETAILS

In this appendix, we provide implementation details for the experiments discussed in Section IV-C and Section V, along with more qualitative results and visual samples.

a) Time Discretization Strategy: The *power-uniform* discretization strategy is formulated as:

$$\lambda_i^{(\rho)} = \left(\lambda_{t_N}^{1/\rho} + \frac{i}{N} \left(\lambda_{t_0}^{1/\rho} - \lambda_{t_N}^{1/\rho} \right) \right)^{\rho}, \quad i = 0, \ldots, N, \tag{109}$$

where $\lambda(t) = \log(\alpha(t)/\sigma(t))$ represents the half-logSNR. The boundary values λ_{t_0} and λ_{t_N} correspond to the start and end points of the sampling trajectory, determined by the noise schedules employed during training (6).

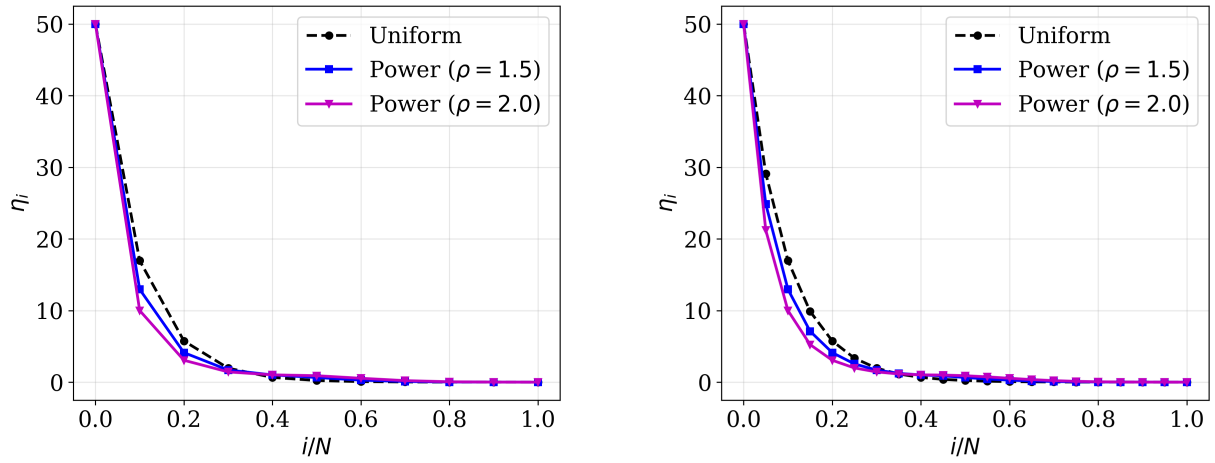
To accommodate negative values (specifically when $\lambda < 0$), we adopt the sign-preserving power function, defined as:

$$x^\rho \triangleq \mathrm{sign}(x)|x|^\rho,$$

where $\mathrm{sign}(x)$ denotes the sign function. As illustrated in Figure 3, the *power-uniform* discretization with $\rho > 1$ induces a more rapid decay of η in the high-noise regime (where $\lambda < 0$), which corresponds to the early stages of the reverse sampling process.

Given the discretization in terms of λ, the discrete time steps t_i for the reverse sampling process (10) are obtained via the inverse mapping $t_i = \lambda^{-1}(\lambda_i^{(\rho)})$. This mapping relies on the specific definitions of $\alpha(t)$ and $\sigma(t)$. Specifically, for the cosine schedule defined as $\alpha(t) = \cos\left(\frac{\pi}{2}t\right)$ and $\sigma(t) = \sin\left(\frac{\pi}{2}t\right)$, this relationship yields the closed-form solution:

$$t_i = \frac{2}{\pi} \arctan\left(e^{-\lambda_i^{(\rho)}} \right). \tag{110}$$



(a) Steps $N = 10$.

(b) Steps $N = 20$.

Fig. 3: Visualization of the $\eta_i = e^{-\lambda_i}$ under different time discretization strategies.

b) Synthetic Source Distribution: We model the source distribution using a synthetic covariance spectrum constructed to mimic the heavy-tailed eigenvalue profiles observed in natural image datasets. Formally, the eigenvalues $\{\mu_\ell\}_{\ell=1}^k$ are generated using a shifted power-law model:

$$\mu_\ell = C \cdot (\ell + i_0)^{-p} + \varepsilon, \qquad \ell = 1, \dots, k, \tag{111}$$

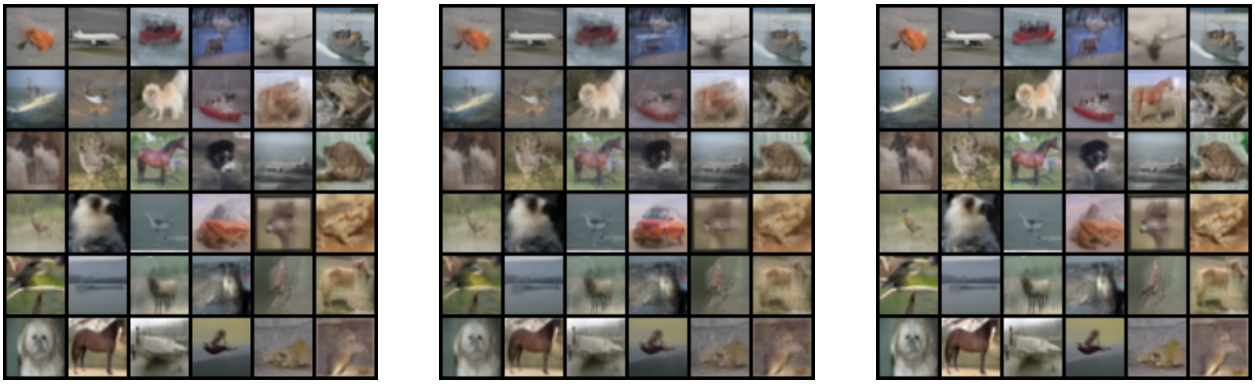
where $p > 0$ governs the decay rate, $i_0 > 0$ smooths the spectral head. The parameter C is calibrated to satisfy the dominant eigenvalue condition $\mu_1 = \mu_{\max}$, while the spectral floor $\varepsilon > 0$ is tuned to ensure the minimal eigenvalue level $\mu_k \geq \mu_{\min}$.

c) Visualization of Generated Images: In this part, we present additional qualitative comparisons of samples generated by different pretrained models, employing representative ODE samplers across various time discretization strategies. The visual results are The visual results are provided on the following pages.

Our implementation builds upon the open-source frameworks listed in Table II. We gratefully acknowledge the authors for making their code and pretrained models publicly available.

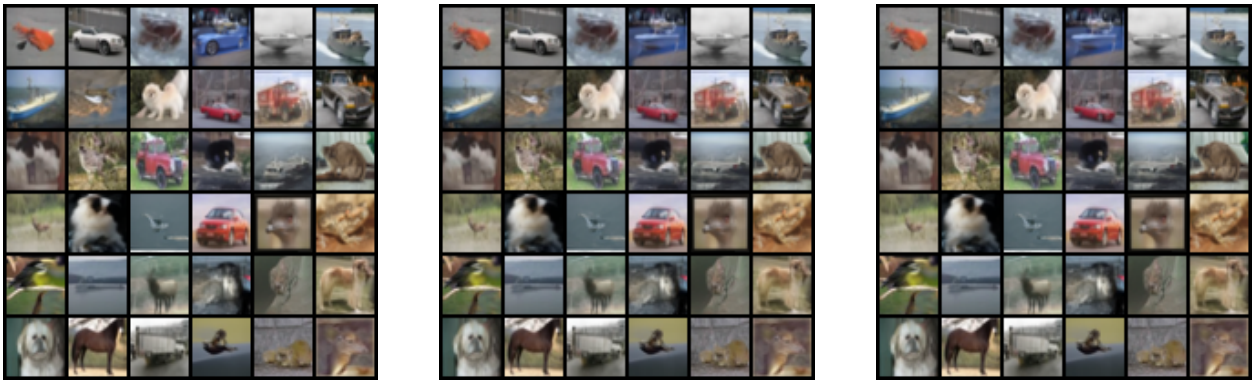
TABLE II: List of open-source repositories referenced in this work.

Method	**Source Code URL**
Score SDE	https://github.com/yang-song/score_sde_pytorch
EDM	https://github.com/NVlabs/edm
DPM-Solver	https://github.com/LuChengTHU/dpm-solver
UniPC	https://github.com/wl-zhao/UniPC
DPM-Solver-v3	https://github.com/thu-ml/DPM-Solver-v3
DM-NonUniform	https://github.com/scxue/DM-NonUniform



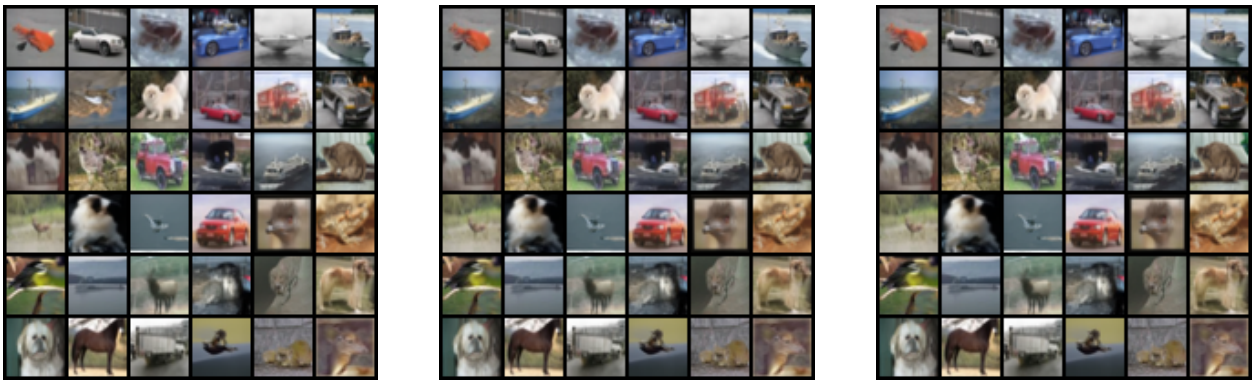
(a) $\rho = 1.5$. (b) Uniform-λ. (c) Xue'24.

Fig. 4: EDM checkpoint with DPM-Solver++ on CIFAR-10 (NFEs=6).



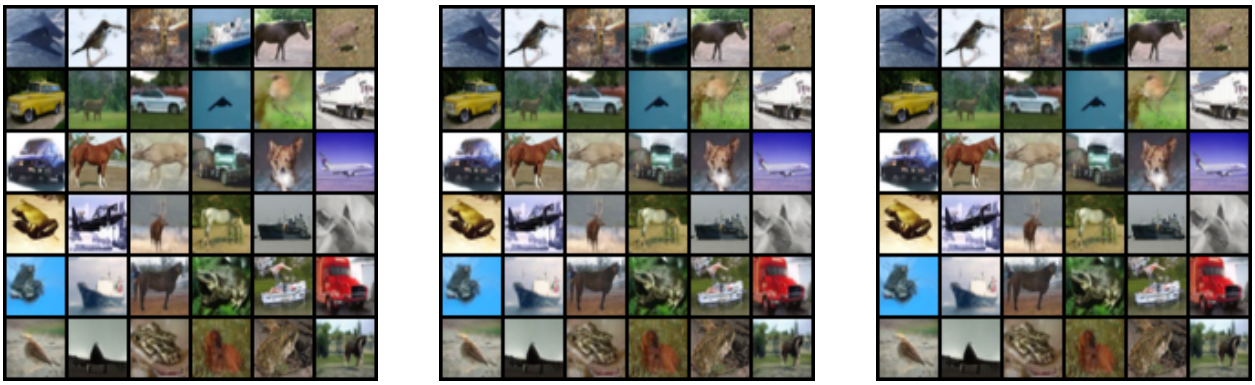
(a) $\rho = 1.5$. (b) Uniform-λ. (c) Xue'24.

Fig. 5: EDM checkpoint with DPM-Solver++ on CIFAR-10 (NFEs=12).



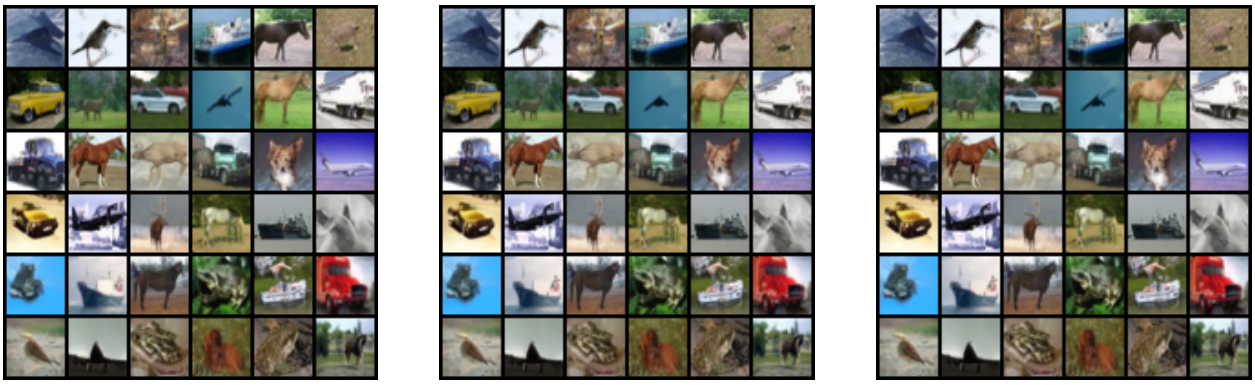
(a) $\rho = 1.5$. (b) Uniform-λ. (c) Xue'24.

Fig. 6: EDM checkpoint with DPM-Solver++ on CIFAR-10 (NFEs=15).



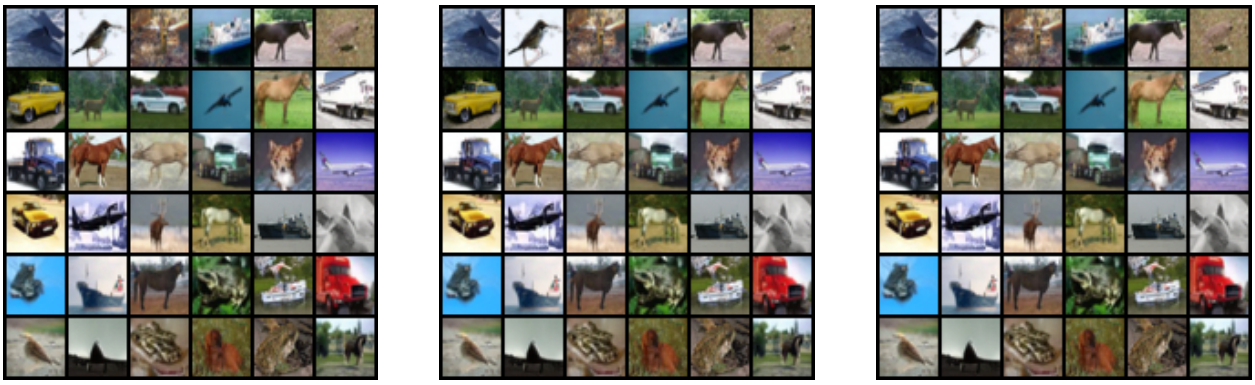
(a) $\rho = 2.0$. (b) Uniform-λ. (c) Xue'24.

Fig. 7: VP-SDE checkpoint with UniPC on CIFAR-10 (NFEs=20).



(a) $\rho = 2.0$. (b) Uniform-λ. (c) Xue'24.

Fig. 8: VP-SDE checkpoint with UniPC on CIFAR-10 (NFEs=25).



(a) $\rho = 2.0$. (b) Uniform-λ. (c) Xue'24.

Fig. 9: VP-SDE checkpoint with UniPC on CIFAR-10 (NFEs=50).

(a) $\rho = 1.5$. (b) Uniform-λ. (c) Xue'24.

Fig. 10: EDM checkpoint with DPM-Solver++ on FFHQ-64 (NFEs=8).

(a) $\rho = 1.5$. (b) Uniform-λ. (c) Xue'24.

Fig. 11: EDM checkpoint with DPM-Solver++ on FFHQ-64 (NFEs=10).

(a) $\rho = 1.5$. (b) Uniform-λ. (c) Xue'24.

Fig. 12: EDM checkpoint with DPM-Solver++ on FFHQ-64 (NFEs=12).