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. Considering a simplified model where the source distribution is multivariate Gaussian with a given covariance matrix, 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 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. This KL divergence is also adopted as a measure to compare different time discretization strategies for any given noise schedule in pretrained models. 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 adopt an elementary approach, considering the case where the source distribution is multivariate Gaussian 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 reverse sampling steps. Leveraging this closed-form formulation, we can conveniently investigate the Kullback-Leibler (KL) divergence between the source distribution and the 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 only 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. 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 recently proposed search-based strategies, particularly when the budget on NFEs is very tight.

Notation: For brevity, we abbreviate the function $f(t)$ as f 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 $\ddot{f} = \mathrm{d}^2 f/\mathrm{d}t^2$, while $\log(\cdot)$ 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 deterministic reverse sampling process employed for generation. Subsequently, we outline the multivariate Gaussian setting that enables the tractable theoretical analysis.

A. Diffusion Models

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, $\mathbf{x}_0 \sim q(\mathbf{x})$ 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 include two types: variance-preserving (VP) and variance-exploding (VE) [3]. The VP setting usually satisfies the constraint $\alpha^2(t) + \sigma^2(t) = 1$, where $\alpha(t)$ decays from 1 to 0. In contrast, the VE setting maintains $\alpha(t) \equiv 1$, while $\sigma(t)$ increases from 0 to a predefined maximum $\sigma_{\max} \gg 1$. In either case, the forward transition kernel is given by

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\big(\alpha_t \mathbf{x}_0,\, \sigma_t^2 \mathbf{I}\big)\,, \tag{2}$$

which progressively diffuses the source distribution $q(\mathbf{x}_0)$ towards a Gaussian noise as $t \to 1$.

The deterministic reverse sampling process, which reverses the forward diffusion process (1) to generate data samples from $q(\mathbf{x})$, is governed by the following update formula [9], [10], for each step $j \in [1, N]$:

$$\hat{\mathbf{x}}_{t_{j-1}} = \frac{\alpha_{t_{j-1}}}{\alpha_{t_j}} \hat{\mathbf{x}}_{t_j} + \left(\sigma_{t_{j-1}} - \frac{\alpha_{t_{j-1}}}{\alpha_{t_j}} \sigma_{t_j}\right) \boldsymbol{\epsilon}_\theta(\hat{\mathbf{x}}_{t_j}, t_j), \tag{3}$$

where $\{t_i\}_{i=0}^N$ denotes the time discretization sequence decreasing from $t_N = 1$ to $t_0 = 0$, and $\boldsymbol{\epsilon}_\theta$ is a learned noise estimator. The reverse sampling process is initialized with $\hat{\mathbf{x}}_{t_N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for the VP setting and $\hat{\mathbf{x}}_{t_N} \sim \mathcal{N}(\mathbf{0}, \sigma_{\max}^2 \mathbf{I})$ for the VE setting.

To obtain the noise estimator $\boldsymbol{\epsilon}_\theta$, the training objective is to minimize the following mean squared error [2]:

$$\mathcal{L}_\theta = \mathbb{E}_{t \in [0,1], \mathbf{x}_0 \sim q(\mathbf{x}), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right\|^2 \right], \tag{4}$$

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:

$$\boldsymbol{\epsilon}_\theta^*(\mathbf{x}_t, t) = \mathbb{E}\left[\frac{\mathbf{x}_t - \alpha_t \mathbf{x}_0}{\sigma_t} \middle| \mathbf{x}_t \right] = \frac{\mathbf{x}_t - \alpha_t \mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]}{\sigma_t}. \tag{5}$$

Thus, learning to predict the noise $\boldsymbol{\epsilon}$ is mathematically equivalent to estimating the posterior mean of the source data $\mathbf{x}_0$.

B. Gaussian Setup

To facilitate theoretical analysis, we consider the case where the source distribution is multivariate Gaussian with a given covariance matrix, i.e., $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{x})$. Without loss of generality, we assume that the covariance matrix $\boldsymbol{\Sigma}_\mathbf{x}$ is positive definite. Let the eigendecomposition of the covariance matrix be $\boldsymbol{\Sigma}_\mathbf{x} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^{-1}$, where $\mathbf{U}$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \mathrm{diag}(\mu_1, \ldots, \mu_k)$ denotes the diagonal matrix of eigenvalues. Furthermore, due to the linearity of the forward diffusion process (1), the marginal distribution $q(\mathbf{x}_t)$ remains Gaussian:

$$q(\mathbf{x}_t) = \mathcal{N}\left(\mathbf{0}, \alpha_t^2 \boldsymbol{\Sigma}_\mathbf{x} + \sigma_t^2 \mathbf{I}\right). \tag{6}$$

Moreover, the optimal posterior estimator admits a closed-form solution [24]:

$$\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t] = \alpha_t \boldsymbol{\Sigma}_\mathbf{x} \left(\alpha_t^2 \boldsymbol{\Sigma}_\mathbf{x} + \sigma_t^2 \mathbf{I}\right)^{-1} \mathbf{x}_t. \tag{7}$$

III. KL-Based Analysis and Applications

In this section, we present our main results derived from an exact analysis under the Gaussian setting, 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.

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 (5) and (7), the generated sample $\hat{\mathbf{x}}_{t_0}$ produced by the reverse sampling process (3) is distributed as*

$$p\left(\hat{\mathbf{x}}_{t_0}\right) = \mathcal{N}\!\left(\mathbf{0},\, \mathbf{U}\mathbf{M}\mathbf{U}^{-1}\right), \quad \mathbf{M} = \operatorname{diag}(m_1, \ldots, m_k), \tag{8}$$

where the ℓ-th eigenvalue is given by

$$m_\ell = \prod_{j=1}^{N} \left(\frac{\alpha_{t_{j-1}} \alpha_{t_j} \mu_\ell + \sigma_{t_{j-1}} \sigma_{t_j}}{\alpha_{t_j}^2 \mu_\ell + \sigma_{t_j}^2} \right)^2, \quad \ell \in \{1, \ldots, k\}. \tag{9}$$

Proof. See Appendix A. $\square$

By (6), we can rewrite $q(\mathbf{x}_{t_0}) = \mathcal{N}(\mathbf{0}, \mathbf{U}\mathbf{N}\mathbf{U}^{-1})$, where $\mathbf{N} = \operatorname{diag}(n_1, \ldots, n_k)$ with

$$n_\ell \triangleq \alpha_{t_0}^2 \mu_\ell + \sigma_{t_0}^2. \tag{10}$$

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

$$D_{\mathrm{KL}}(p(\hat{\mathbf{x}}_{t_0}) \,\|\, q(\mathbf{x}_{t_0})) = \frac{1}{2} \sum_{\ell=1}^{k} \left(\frac{m_\ell}{n_\ell} - \log \frac{m_\ell}{n_\ell} - 1 \right), \tag{11}$$

by the standard KL divergence expression between two zero-mean Gaussians.

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

$$S_\ell^N \triangleq \sum_{j=1}^{N} \log\left(\frac{\alpha_{t_{j-1}} \alpha_{t_j} \mu_\ell + \sigma_{t_{j-1}} \sigma_{t_j}}{\alpha_{t_j}^2 \mu_\ell + \sigma_{t_j}^2} \right), \tag{12}$$

so that, from (9) we have $S_\ell^N = \frac{1}{2} \log m_\ell$.

To further investigate the asymptotic convergence behavior of the KL divergence (11) 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 (12) with the uniform discretization $t_j \triangleq j/N$ on $[0, 1]$. Assume the schedule functions α, σ are sufficiently smooth. Define*

$$I_\ell \triangleq \int_0^1 F_\ell(t)\, \mathrm{d}t, \quad F_\ell(t) \triangleq -\frac{\alpha \dot{\alpha} \mu_\ell + \sigma \dot{\sigma}}{\alpha^2 \mu_\ell + \sigma^2}. \tag{13}$$

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

$$S_\ell^N = I_\ell + \frac{E_\ell^1}{N} + \mathcal{O}\!\left(\frac{1}{N^2}\right), \tag{14}$$

where E_ℓ^1 is independent of N, specifically,

$$E_\ell^1 = -\frac{\mu_\ell}{2} \int_0^1 \frac{(\alpha \dot{\sigma} - \sigma \dot{\alpha})^2}{(\alpha^2 \mu_\ell + \sigma^2)^2}\, \mathrm{d}t. \tag{15}$$

Moreover, it holds that

$$I_\ell = \frac{1}{2} \log n_\ell. \tag{16}$$

Proof. See Appendix B. $\square$

Combining the definition (12) with the closed-form solution (16) and the expansion (14) in Lemma 2, we can explicitly express the residual r_ℓ as

$$r_\ell \triangleq S_\ell^N - I_\ell = \frac{1}{2}\log\frac{m_\ell}{n_\ell} = \frac{E_\ell^1}{N} + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{17}$$

This implies that $r_\ell = \mathcal{O}(1/N)$, which justifies the use of the Taylor expansion for large N. Substituting (17) into (11), the KL divergence can be expanded as follows:

$$\begin{aligned} D_{\mathrm{KL}}(p(\hat{\mathbf{x}}_{t_0}) \,\|\, q(\mathbf{x}_{t_0})) &= \frac{1}{2}\sum_{\ell=1}^{k}\left(e^{2r_\ell} - 2r_\ell - 1\right) \\ &= \sum_{\ell=1}^{k} r_\ell^2 + \mathcal{O}\left(\sum_{\ell=1}^{k}|r_\ell|^3\right) \\ &= \frac{1}{N^2}\sum_{\ell=1}^{k}(E_\ell^1)^2 + \mathcal{O}\left(\frac{1}{N^3}\right). \end{aligned} \tag{18}$$

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 is asymptotically equivalent to minimizing the coefficient of the leading $1/N^2$ term. This yields the following objective functional subject to corresponding boundary conditions:

$$\min_{\alpha,\,\sigma}\ \mathcal{L}[\alpha,\sigma] \triangleq \sum_{\ell=1}^{k}(E_\ell^1)^2. \tag{19}$$

Remark (Convergence and Consistency)**.** From (18), 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 $D_{\mathrm{KL}}(p(\hat{\mathbf{x}}_{t_0})\|q(\mathbf{x}_{t_0})) \to 0$. This verifies that the reverse sampling process in (3) 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., [25]).

B. *Derivation of the 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 (19). However, directly minimizing the objective functional in (19) 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 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

$$\eta_\ell(t) = \sqrt{\mu_\ell}\tan\left(\frac{\pi}{2}t\right), \quad t \in [0,1). \tag{20}$$

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

Theorem 1. *The minimization of the term $(E_\ell^1)^2$ in (19), with the expression of E_ℓ^1 given in (15), is equivalent to minimizing*

$$\min_\eta \mathcal{J}_\ell[\eta] \triangleq \int_0^1 \frac{\dot{\eta}^2}{\left(\mu_\ell + \eta^2\right)^2} \,\mathrm{d}t,$$
$$\textit{subject to } \eta(0) = 0 \textit{ and } \lim_{t\to 1} \eta(t) = \infty. \tag{21}$$

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 $\alpha^2\mu_\ell + \sigma^2 = \alpha^2(\mu_\ell + \eta^2)$, (15) 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 (21). 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:

$$Q_\ell(t) \triangleq \frac{1}{\sqrt{\mu_\ell}} \arctan\left(\frac{\eta(t)}{\sqrt{\mu_\ell}}\right). \tag{22}$$

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

$$\dot{Q}_\ell(t) = \frac{1}{\sqrt{\mu_\ell}} \cdot \frac{1}{1 + (\eta/\sqrt{\mu_\ell})^2} \cdot \frac{\dot{\eta}}{\sqrt{\mu_\ell}} = \frac{\dot{\eta}}{\mu_\ell + \eta^2}. \tag{23}$$

Consequently, the objective functional simplifies to

$$\mathcal{J}_\ell[\eta] = \int_0^1 \left(\dot{Q}_\ell\right)^2 \mathrm{d}t. \tag{24}$$

The Euler–Lagrange equation [21] for the Lagrangian function $L(t, Q_\ell, \dot{Q}_\ell) = \dot{Q}_\ell^2$ is

$$\frac{\partial L}{\partial Q_\ell} - \frac{\mathrm{d}}{\mathrm{dt}}\left(\frac{\partial L}{\partial \dot{Q}_\ell}\right) = 2\ddot{Q}_\ell = 0, \tag{25}$$

which implies that the optimal $Q_\ell(t)$ must be a linear function of t: $Q_\ell(t) = ct + d$.

Given the boundary conditions $\eta(0) = 0$ and $\lim_{t\to 1} \eta(t) = \infty$, 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 $Q_\ell(t)$ (22) yields the tangent law schedule $\eta_\ell(t) = \sqrt{\mu_\ell}\tan(\frac{\pi}{2}t)$. □

Theorem 1 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 1:

$$\eta_\gamma(t) \triangleq \sqrt{\gamma}\,\tan\left(\frac{\pi}{2}t\right). \tag{26}$$

Substituting this parameterized tangent law schedule into the leading order coefficient (15) and the global objective functional (19) yields the following global optimization result.

Theorem 2. *Under the parameterized tangent law schedule (26), the global objective functional (19) is strictly convex with respect to γ and admits a unique closed-form minimizer,*

$$\gamma^\star = \sqrt{\frac{\sum_{\ell=1}^{k} \mu_\ell}{\sum_{\ell=1}^{k} \mu_\ell^{-1}}} = \sqrt{\frac{\mathrm{tr}(\mathbf{\Sigma_x})}{\mathrm{tr}(\mathbf{\Sigma_x}^{-1})}}. \tag{27}$$

Proof. As derived in Appendix D, the leading order coefficient (15) for the parameterized tangent law schedule (26) is given by

$$E_\ell^1(\gamma) = -\frac{\pi^2}{16}\left(\sqrt{\frac{\mu_\ell}{\gamma}} + \sqrt{\frac{\gamma}{\mu_\ell}}\right). \tag{28}$$

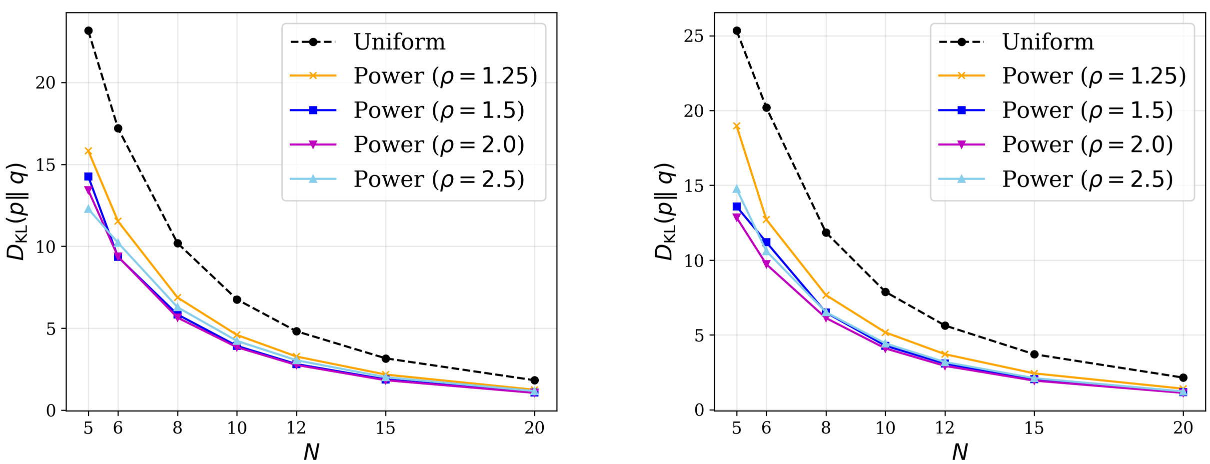
Substituting (28) into $\mathcal{L}(\gamma) = \sum_{\ell=1}^{k}\left(E_\ell^1(\gamma)\right)^2$ and omitting scaling factors and constant terms, we have that minimizing $\mathcal{L}(\gamma)$ is equivalent to minimizing the function

$$J(\gamma) = \gamma \sum_{\ell=1}^{k} \mu_\ell^{-1} + \gamma^{-1} \sum_{\ell=1}^{k} \mu_\ell. \tag{29}$$

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

$$J'(\gamma) = \sum_{\ell=1}^{k} \mu_\ell^{-1} - \gamma^{-2} \sum_{\ell=1}^{k} \mu_\ell, J''(\gamma) = 2\gamma^{-3} \sum_{\ell=1}^{k} \mu_\ell. \tag{30}$$

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 (27). □



(a) VP setting.

(b) VE setting.

Fig. 1: KL divergence versus the number of time discretization steps N for the time discretization strategy (31) under (a) VP and (b) VE settings.

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 $\sigma^2_{\max} \gg \mu_\ell$, the boundary effect becomes negligible. Thus, the VP optimizer $\gamma^\star$ in (27) can still provide an accurate approximation.

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 with the tangent law schedule is prohibitively computationally expensive, and in practice, we 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 (11) 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 with ρ (using sign-preserving powers $sign(x)|x|^\rho$ when x is negative):

$$\lambda_i^{(\rho)} = \left(\lambda_{t_N}^{1/\rho} + \tfrac{i}{N}\big(\lambda_{t_0}^{1/\rho} - \lambda_{t_N}^{1/\rho}\big)\right)^\rho, \quad i = 0, \ldots, N, \tag{31}$$

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 (3) are recovered via $t_i = \lambda^{-1}(\lambda_i^{(\rho)})$, which only depends on the prescribed $\alpha(t), \sigma(t)$.

We model the data distribution using a synthetic power-law covariance spectrum, which mimics the heavy-tailed eigenvalue decay characteristic of real image data. Figure 1 presents the quantitative evaluation of the closed-form KL divergence (11) across varying step budgets N. This comparison covers both VP and VE settings as considered in [3] with the same time discretization strategy defined in (31). The results demonstrate that in both settings, the time discretization with an appropriate choice of ρ can achieve notably faster convergence compared to the widely used *uniform-*λ baseline, significantly reducing the discretization error under limited budgets. These preliminary observations motivate further experimentation on real-world image generation tasks, as presented in the next section.

IV. Experiments

Focusing on the practical generation tasks with pretrained DMs, we evaluate the performance of our selected time discretization strategies in conjunction with general 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 (3), and report Fréchet Inception Distance (FID) scores, a universally accepted 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 strikes a balance between optimization overhead and generation performance for searching optimal steps. (iii) *Ours*, the *power-uniform* schemes with $\rho = 1.5$ and 2.0.

TABLE I: Quantitative comparison of time discretization strategies in terms of FID scores (↓). **Bold** highlights the best result, while $\underline{\text{underlining}}$ 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**	$\underline{54.95}$	**35.93**	$\underline{26.76}$	**20.01**	**16.31**	**13.26**	**11.47**	**8.70**	**6.48**	**4.67**	**3.83**	**3.34**	**2.60**	**2.24**
	Ours ($\rho = 2.0$)	$\underline{86.27}$	**53.82**	$\underline{37.26}$	27.04	$\underline{20.74}$	$\underline{16.75}$	$\underline{13.75}$	$\underline{11.78}$	$\underline{9.04}$	$\underline{6.66}$	$\underline{4.81}$	$\underline{3.92}$	$\underline{3.42}$	$\underline{2.63}$	$\underline{2.25}$
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 ($\rho = 1.5$)	$\underline{79.62}$	$\underline{49.44}$	$\underline{29.50}$	20.16	$\underline{14.11}$	$\underline{10.78}$	**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**	$\underline{18.94}$	**13.84**	**10.72**	$\underline{8.67}$	$\underline{7.22}$	$\underline{5.45}$	$\underline{4.08}$	$\underline{3.06}$	$\underline{2.63}$	$\underline{2.42}$	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**	$\underline{2.91}$
	Ours ($\rho = 2.0$)	$\underline{82.31}$	$\underline{54.54}$	$\underline{38.60}$	$\underline{29.96}$	$\underline{24.75}$	$\underline{21.18}$	$\underline{18.07}$	$\underline{15.89}$	$\underline{12.63}$	$\underline{9.67}$	$\underline{7.09}$	$\underline{5.75}$	$\underline{4.95}$	$\underline{3.63}$	**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	$\underline{22.35}$	18.38	14.99	16.26	14.25	10.51	7.36	4.99	3.99	3.47	2.79	2.53
	Ours ($\rho = 1.5$)	$\underline{75.76}$	**46.25**	$\underline{31.69}$	23.76	$\underline{18.14}$	$\underline{14.41}$	**11.62**	**9.80**	**7.32**	**5.35**	**3.91**	**3.29**	**2.99**	**2.61**	$\underline{2.49}$
	Ours ($\rho = 2.0$)	**75.49**	$\underline{46.84}$	**30.18**	**21.66**	**17.13**	**14.05**	$\underline{11.73}$	$\underline{10.01}$	$\underline{7.68}$	$\underline{5.69}$	$\underline{4.12}$	$\underline{3.42}$	$\underline{3.05}$	**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	$\underline{52.80}$	$\underline{38.45}$	**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	$\underline{16.59}$	$\underline{14.19}$	$\underline{12.25}$	$\underline{9.75}$	$\underline{7.54}$	$\underline{5.77}$	$\underline{4.86}$	$\underline{4.32}$	$\underline{3.50}$	$\underline{3.06}$
	Ours ($\rho = 2.0$)	**51.87**	**32.27**	$\underline{28.98}$	$\underline{23.09}$	$\underline{19.82}$	**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	$\underline{49.77}$	$\underline{36.07}$	**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	$\underline{11.14}$	$\underline{9.00}$	$\underline{7.55}$	**5.76**	**4.45**	**3.56**	$\underline{3.19}$	$\underline{3.04}$	$\underline{2.86}$	$\underline{2.81}$
	Ours ($\rho = 2.0$)	**43.55**	**28.43**	$\underline{22.24}$	$\underline{17.02}$	$\underline{13.48}$	**10.70**	**8.89**	**7.46**	**5.76**	$\underline{4.51}$	**3.56**	**3.17**	**3.01**	**2.82**	**2.80**

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.

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

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

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

APPENDIX A
PROOF OF LEMMA 1

This proof shows that, under the optimal estimator given by (5) and (7), the explicit closed-form evolution trajectory of the distributions across reverse sampling steps can be derived.

Firstly, substituting (5) into the deterministic reverse sampling process (3) yields

$$\begin{aligned}\hat{\mathbf{x}}_{t_{j-1}} &= \frac{\alpha_{t_{j-1}}}{\alpha_{t_j}}\hat{\mathbf{x}}_{t_j} + \left(\sigma_{t_{j-1}} - \frac{\alpha_{t_{j-1}}}{\alpha_{t_j}}\sigma_{t_j}\right)\frac{\hat{\mathbf{x}}_{t_j} - \alpha_{t_j}\mathbb{E}[\mathbf{x}_0|\hat{\mathbf{x}}_{t_j}]}{\sigma_{t_i}} \\ &= \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\hat{\mathbf{x}}_{t_j} + \left(\alpha_{t_{j-1}} - \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\alpha_{t_j}\right)\mathbb{E}[\mathbf{x}_0|\hat{\mathbf{x}}_{t_j}].\end{aligned} \tag{32}$$

Under the Gaussian setup, the optimal posterior estimator (7) gives

$$\mathbb{E}[\mathbf{x}_0|\hat{\mathbf{x}}_{t_j}] = \alpha_{t_j}\mathbf{\Sigma}_{\mathbf{x}}\left(\alpha_{t_j}^2\mathbf{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right)^{-1}\hat{\mathbf{x}}_{t_j}. \tag{33}$$

Plugging (33) into (32) yields the linear update

$$\begin{aligned}\hat{\mathbf{x}}_{t_{j-1}} &= \left(\frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\mathbf{I} + \left(\alpha_{t_{j-1}} - \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\alpha_{t_j}\right)\alpha_{t_j}\mathbf{\Sigma}_{\mathbf{x}}\left(\alpha_{t_j}^2\mathbf{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right)^{-1}\right)\hat{\mathbf{x}}_{t_j} \\ &= \left(\frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\left(\alpha_{t_j}^2\mathbf{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right) + \left(\alpha_{t_{j-1}}\alpha_{t_j}\mathbf{\Sigma}_{\mathbf{x}} - \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}}\alpha_{t_j}^2\mathbf{\Sigma}_{\mathbf{x}}\right)\right)\left(\alpha_{t_j}^2\mathbf{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right)^{-1}\hat{\mathbf{x}}_{t_j} \\ &= \left(\sigma_{t_{j-1}}\sigma_{t_j}\mathbf{I} + \alpha_{t_{j-1}}\alpha_{t_j}\mathbf{\Sigma}_{\mathbf{x}}\right)\left(\alpha_{t_j}^2\mathbf{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I}\right)^{-1}\hat{\mathbf{x}}_{t_j}.\end{aligned} \tag{34}$$

Since $\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$ with $\mathbf{\Lambda} = \mathrm{diag}(\mu_1, \ldots, \mu_k)$, we have $(\alpha_{t_j}^2\mathbf{\Sigma}_{\mathbf{x}} + \sigma_{t_j}^2\mathbf{I})^{-1} = \mathbf{U}(\alpha_{t_j}^2\mathbf{\Lambda} + \sigma_{t_j}^2\mathbf{I})^{-1}\mathbf{U}^{-1}$. Then the updata formula (34) becomes

$$\begin{aligned}\hat{\mathbf{x}}_{t_{j-1}} &= \mathbf{U}\left(\sigma_{t_{j-1}}\sigma_{t_j}\mathbf{I} + \alpha_{t_{j-1}}\alpha_{t_j}\mathbf{\Lambda}\right)\mathbf{U}^{-1}\mathbf{U}(\alpha_{t_j}^2\mathbf{\Lambda} + \sigma_{t_j}^2\mathbf{I})^{-1}\mathbf{U}^{-1}\hat{\mathbf{x}}_{t_j} \\ &= \mathbf{U}\left(\sigma_{t_{j-1}}\sigma_{t_j}\mathbf{I} + \alpha_{t_{j-1}}\alpha_{t_j}\mathbf{\Lambda}\right)(\alpha_{t_j}^2\mathbf{\Lambda} + \sigma_{t_j}^2\mathbf{I})^{-1}\mathbf{U}^{-1}\hat{\mathbf{x}}_{t_j} \\ &\triangleq \mathbf{U}\mathbf{D}_j\mathbf{U}^{-1}\hat{\mathbf{x}}_{t_j},\end{aligned} \tag{35}$$

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

$$\mathbf{D}_j = \left(\alpha_{t_{j-1}}\alpha_{t_j}\mathbf{\Lambda} + \sigma_{t_{j-1}}\sigma_{t_j}\mathbf{I}\right)\left(\alpha_{t_j}^2\mathbf{\Lambda} + \sigma_{t_j}^2\mathbf{I}\right)^{-1} = \mathrm{diag}(d_{j,1}, \ldots, d_{j,k}), \tag{36}$$

with the diagonal entries

$$d_{j,\ell} = \frac{\alpha_{t_{j-1}}\alpha_{t_j}\mu_\ell + \sigma_{t_{j-1}}\sigma_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}, \qquad \ell = 1, \ldots, k. \tag{37}$$

Then, using the diagonalized linear update (35) repeatedly with the time discretization sequence $\{t_i\}_{i=0}^N$ yields

$$\hat{\mathbf{x}}_{t_0} = \mathbf{U}\left(\prod_{j=1}^N \mathbf{D}_j\right)\mathbf{U}^{-1}\hat{\mathbf{x}}_{t_N}, \tag{38}$$

with the diagonal matrix

$$\prod_{j=1}^N \mathbf{D}_j = \mathrm{diag}\left(\prod_{j=1}^N d_{j,1}, \ldots, \prod_{j=1}^N d_{j,k}\right). \tag{39}$$

By (6), in practical reverse sampling, the reverse process is initialized from the terminal prior:

$$\hat{\mathbf{x}}_{t_N} \sim \mathcal{N}(\mathbf{0}, \sigma_{t_N}^2\mathbf{I}), \tag{40}$$

where $\sigma_{t_N}=1$ in the VP setting, and in the VE setting, $\sigma_{t_N}=\sigma_{\max}$. Since the update (38) is affine transformation on a Gaussian distribution, $\hat{\mathbf{x}}_{t_0}$ remains Gaussian:

$$p(\hat{\mathbf{x}}_{t_0})=\mathcal{N}\big(\mathbf{0},\,\mathbf{U}\mathbf{M}\mathbf{U}^{-1}\big)\,, \tag{41}$$

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

$$\mathbf{M}\triangleq\sigma_{t_N}^2\left(\prod_{j=1}^{N}\mathbf{D}_j\right)\left(\prod_{j=1}^{N}\mathbf{D}_j\right)^{\top}=\mathrm{diag}(m_1,\ldots,m_k), \tag{42}$$

with its diagonal entries

$$m_\ell=\sigma_{t_N}^2\prod_{j=1}^{N}d_{j,\ell}^2,\qquad \ell=1,\ldots,k, \tag{43}$$

where $d_{j,\ell}$ is given in (37). This yields the result stated in Lemma 1.

The KL divergence between two zero-mean Gaussians is

$$D_{\mathrm{KL}}(\mathcal{N}(0,\Sigma_p)\|\mathcal{N}(0,\Sigma_q))=\frac{1}{2}\left(\mathrm{tr}(\mathbf{S})-\log\det(\mathbf{S})-k\right), \tag{44}$$

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

$$\begin{aligned}D_{\mathrm{KL}}(p(\hat{\mathbf{x}}_{t_0})\,\|\,q(\mathbf{x}_{t_0}))&=D_{\mathrm{KL}}\big(\mathcal{N}\big(\mathbf{0},\,\mathbf{U}\mathbf{M}\mathbf{U}^{-1}\big)\,\|\,\mathcal{N}\big(\mathbf{0},\,\mathbf{U}\mathbf{N}\mathbf{U}^{-1}\big)\big)\\&=\frac{1}{2}\big(\mathrm{tr}(\mathbf{N}^{-1}\mathbf{M})-\log\det(\mathbf{N}^{-1}\mathbf{M})-k\big)\\&=\frac{1}{2}\sum_{\ell=1}^{k}\left(\frac{m_\ell}{n_\ell}-\log\frac{m_\ell}{n_\ell}-1\right),\end{aligned} \tag{45}$$

where m_ℓ and n_ℓ are respectively given in (43) and (10).

APPENDIX B
PROOF OF LEMMA 2

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

$$S_\ell^N=\sum_{j=1}^{N}\log\left(\frac{\alpha_{t_{j-1}}\alpha_{t_j}\mu_\ell+\sigma_{t_{j-1}}\sigma_{t_j}}{\alpha_{t_j}^2\mu_\ell+\sigma_{t_j}^2}\right). \tag{46}$$

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 $h\to 0$:*

$$S_\ell^N=\sum_{j=1}^{N}\Big(F_\ell(t_j)\,h+G_\ell(t_j)\,h^2\Big)+\mathcal{O}(h^2), \tag{47}$$

where the coefficients are given by

$$F_\ell(t_j)=-\frac{\alpha_{t_j}\dot{\alpha}_{t_j}\mu_\ell+\sigma_{t_j}\dot{\sigma}_{t_j}}{\alpha_{t_j}^2\mu_\ell+\sigma_{t_j}^2}, \tag{48}$$

$$G_\ell(t_j)=\frac{\alpha_{t_j}\ddot{\alpha}_{t_j}\mu_\ell+\sigma_{t_j}\ddot{\sigma}_{t_j}}{2\big(\alpha_{t_j}^2\mu_\ell+\sigma_{t_j}^2\big)}-\frac{\big(\alpha_{t_j}\dot{\alpha}_{t_j}\mu_\ell+\sigma_{t_j}\dot{\sigma}_{t_j}\big)^2}{2\big(\alpha_{t_j}^2\mu_\ell+\sigma_{t_j}^2\big)^2}. \tag{49}$$

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

$$\alpha_{t_{j-1}} = \alpha(t_j - h) = \alpha_{t_j} - h\,\dot{\alpha}_{t_j} + \frac{h^2}{2}\,\ddot{\alpha}_{t_j} + \mathcal{O}(h^3),$$
$$\sigma_{t_{j-1}} = \sigma(t_j - h) = \sigma_{t_j} - h\,\dot{\sigma}_{t_j} + \frac{h^2}{2}\,\ddot{\sigma}_{t_j} + \mathcal{O}(h^3), \tag{50}$$

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 (50) into the numerator term of the summand in (46), we obtain

$$\begin{aligned}\alpha_{t_{j-1}}\alpha_{t_j}\mu_\ell + \sigma_{t_{j-1}}\sigma_{t_j} &= \left(\alpha_{t_j} - h\,\dot{\alpha}_{t_j} + \frac{h^2}{2}\,\ddot{\alpha}_{t_j}\right)\alpha_{t_j}\mu_\ell + \left(\sigma_{t_j} - h\,\dot{\sigma}_{t_j} + \frac{h^2}{2}\,\ddot{\sigma}_{t_j}\right)\sigma_{t_j} + \mathcal{O}(h^3) \\ &= \alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2 - h\big(\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}\big) + \frac{h^2}{2}\big(\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}\big) + \mathcal{O}(h^3). \end{aligned} \tag{51}$$

Then, using (51), the sum S_ℓ^N becomes

$$\begin{aligned} S_\ell^N &= \sum_{j=1}^N \log\left(\frac{\alpha_{t_{j-1}}\alpha_{t_j}\mu_\ell + \sigma_{t_{j-1}}\sigma_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\right) \\ &= \sum_{j=1}^N \log\left(1 + \frac{-h\big(\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}\big) + \frac{h^2}{2}\big(\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}\big)}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2} + \mathcal{O}(h^3)\right) \\ &\triangleq \sum_{j=1}^N \log\left(1 + \delta_j\right), \end{aligned} \tag{52}$$

where the term δ_j is expanded as

$$\begin{aligned} \delta_j &= \frac{-h\big(\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}\big) + \frac{h^2}{2}\big(\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}\big)}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2} + \mathcal{O}(h^3) \\ &= -\frac{\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\,h + \frac{\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}}{2(\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2)}\,h^2 + \mathcal{O}(h^3) \\ &\triangleq a_j h + b_j h^2 + \mathcal{O}(h^3). \end{aligned} \tag{53}$$

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

$$\delta_j^2 = a_j^2 h^2 + \mathcal{O}(h^3) = \left(\frac{\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\right)^2 h^2 + \mathcal{O}(h^3). \tag{54}$$

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

$$\begin{aligned} \log(1+\delta_j) = \delta_j - \frac{1}{2}\delta_j^2 + \mathcal{O}(\delta_j^3) &= \big(a_j h + b_j h^2\big) - \frac{1}{2}\left(a_j^2 h^2\right) + \mathcal{O}(h^3) \\ &= -\frac{\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}}{\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2}\,h + \left[\frac{\mu_\ell\alpha_{t_j}\ddot{\alpha}_{t_j} + \sigma_{t_j}\ddot{\sigma}_{t_j}}{2(\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2)} - \frac{\big(\mu_\ell\alpha_{t_j}\dot{\alpha}_{t_j} + \sigma_{t_j}\dot{\sigma}_{t_j}\big)^2}{2(\alpha_{t_j}^2\mu_\ell + \sigma_{t_j}^2)^2}\right] h^2 + \mathcal{O}(h^3) \\ &\triangleq F_\ell(t_j)h + G_\ell(t_j)h^2 + \mathcal{O}(h^3), \end{aligned} \tag{55}$$

where we used the fact that $\mathcal{O}(\delta_j^3) = \mathcal{O}(h^3)$. Finally, substituting (55) 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

$$S_\ell^N = \sum_{j=1}^N \Big(F_\ell(t_j)\,h + G_\ell(t_j)\,h^2\Big) + \mathcal{O}(h^2), \tag{56}$$

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]:

$$\sum_{j=1}^{N} f(jh)\,h = \int_0^1 f(s)\,\mathrm{d}s + \frac{h}{2}\big(f(1) - f(0)\big) + \frac{h^2}{12}\big(f'(1) - f'(0)\big) + R(f,h), \tag{57}$$

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

Applying (57) to the Riemann sums in Proposition 1 yields, as $h \to 0$,

$$\begin{aligned} \sum_{j=1}^{N} F_\ell(t_j)\,h &= \int_0^1 F_\ell(t)\,\mathrm{d}t + \frac{h}{2}\big(F_\ell(1) - F_\ell(0)\big) + \mathcal{O}(h^2), \\ \sum_{j=1}^{N} G_\ell(t_j)\,h &= \int_0^1 G_\ell(t)\,\mathrm{d}t + \mathcal{O}(h), \end{aligned} \tag{58}$$

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

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

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

Recalling the expansion in (47) and substituting the approximations from (58), we obtain

$$\begin{aligned} S_\ell^N &= \sum_{j=1}^{N}\Big(F_\ell(t_j)\,h + G_\ell(t_j)\,h^2\Big) + \mathcal{O}(h^2) \\ &= \sum_{j=1}^{N} F_\ell(t_j)\,h + h\sum_{j=1}^{N} G_\ell(t_j)\,h + \mathcal{O}(h^2) \\ &= \int_0^1 F_\ell(t)\,\mathrm{d}t + \frac{h}{2}\big(F_\ell(1) - F_\ell(0)\big) + h\left(\int_0^1 G_\ell(t)\,\mathrm{d}t + \mathcal{O}(h)\right) + \mathcal{O}(h^2) \\ &= \int_0^1 F_\ell(t)\,\mathrm{d}t + E_\ell^1\,h + \mathcal{O}(h^2), \end{aligned} \tag{61}$$

where the leading order coefficient is given by

$$E_\ell^1 = \frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) + \int_0^1 G_\ell(t)\,\mathrm{d}t. \tag{62}$$

Define the continuous counterpart

$$I_\ell \triangleq \int_0^1 F_\ell(t)\,\mathrm{d}t. \tag{63}$$

Then by (61) and (63), the discretization error $r_\ell \triangleq S_\ell^N - I_\ell$ satisfies

$$r_\ell = \frac{E_\ell^1}{N} + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{64}$$

Next, we derive closed-form expressions for I_ℓ and E_ℓ^1.

A. Closed Form of Integral I_ℓ

Substituting the definition of $F_\ell(t)$ (59) 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:

$$
\begin{aligned}
I_\ell &= \int_0^1 F_\ell(t)\,\mathrm{d}t \\
&= \int_0^1 -\frac{\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma}}{\alpha^2\mu_\ell + \sigma^2}\,\mathrm{d}t \\
&= -\frac{1}{2}\int_0^1 \frac{\mathrm{d}(\alpha^2\mu_\ell + \sigma^2)}{\alpha^2\mu_\ell + \sigma^2} \\
&= -\frac{1}{2}\log(\alpha^2\mu_\ell + \sigma^2)\Big|_{t=0}^{t=1} \\
&= \frac{1}{2}\log\left(\frac{\alpha^2(0)\mu_\ell + \sigma^2(0)}{\alpha^2(1)\mu_\ell + \sigma^2(1)}\right).
\end{aligned} \tag{65}
$$

To interpret this result, we first recall that $n_\ell \triangleq \alpha^2(0)\mu_\ell + \sigma^2(0)$, consistent with the notation in (10). We now consider the following two boundary conditions, as we mentioned before in Section II-A

- Variance Preserving (VP): The boundary conditions are specified as $\alpha(0) = 1$, $\sigma(0) = 0$, and typically $\alpha(1) = 0$, $\sigma(1) = 1$. Substituting these into (65) yields

$$
I_\ell^{(\mathrm{VP})} = \frac{1}{2}\log\left(\frac{n_\ell}{0\cdot\mu_\ell + 1}\right) = \frac{1}{2}\log n_\ell. \tag{66}
$$

- Variance Exploding (VE): The boundary conditions are $\alpha(t) \equiv 1$, $\sigma(0) = 0$, and $\sigma(1) = \sigma_{\max}$. The integral (65) becomes

$$
I_\ell^{(\mathrm{VE})} = \frac{1}{2}\log\left(\frac{n_\ell}{\mu_\ell + \sigma_{\max}^2}\right). \tag{67}
$$

  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

$$
I_\ell^{(\mathrm{VE})} + \log\sigma_{\max} = \frac{1}{2}\log n_\ell + \mathcal{O}\left(\frac{\mu_\ell}{\sigma_{\max}^2}\right). \tag{68}
$$

We thus establish the unified relationship

$$
\frac{1}{2}\log n_\ell \approx I_\ell + \log\sigma_N, \tag{69}
$$

where σ_N is defined in (40). Note that the relation in (69) 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 (43) and (46), we have the exact identity

$$
\frac{1}{2}\log m_\ell = S_\ell^N + \log\sigma_N. \tag{70}
$$

Subtracting (69) from (70), we obtain

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

Consequently, to minimize the KL divergence given in (45), 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 (60) into two parts, denoted as $A_\ell(t)$ and $B_\ell(t)$:

$$G_\ell(t) = \underbrace{\frac{\alpha\ddot{\alpha}\mu_\ell + \sigma\ddot{\sigma}}{2(\alpha^2\mu_\ell + \sigma^2)}}_{A_\ell(t)} - \underbrace{\frac{(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{2(\alpha^2\mu_\ell + \sigma^2)^2}}_{B_\ell(t)}. \tag{71}$$

Substituting this decomposition into the expression for E_ℓ^1 (62), we obtain

$$\begin{aligned} E_\ell^1 &= \frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) + \int_0^1 G_\ell(t)\,\mathrm{d}t \\ &= \frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) + \int_0^1 A_\ell(t)\,\mathrm{d}t - \int_0^1 B_\ell(t)\,\mathrm{d}t. \end{aligned} \tag{72}$$

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

$$\begin{aligned} \int_0^1 A_\ell(t)\,\mathrm{d}t &= \int_0^1 \frac{1}{2}\left(\frac{\alpha\mu_\ell}{\alpha^2\mu_\ell + \sigma^2}\ddot{\alpha} + \frac{\sigma}{\alpha^2\mu_\ell + \sigma^2}\ddot{\sigma}\right)\mathrm{d}t \\ &= \frac{1}{2}\int_0^1 \frac{\alpha\mu_\ell}{\alpha^2\mu_\ell + \sigma^2}\,\mathrm{d}(\dot{\alpha}) + \frac{1}{2}\int_0^1 \frac{\sigma}{\alpha^2\mu_\ell + \sigma^2}\,\mathrm{d}(\dot{\sigma}) \\ &= \frac{1}{2}\left[\frac{\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma}}{\alpha^2\mu_\ell + \sigma^2}\right]_0^1 - \frac{1}{2}\int_0^1 \left(\dot{\alpha}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\alpha\mu_\ell}{\alpha^2\mu_\ell + \sigma^2} + \dot{\sigma}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\sigma}{\alpha^2\mu_\ell + \sigma^2}\right)\mathrm{d}t \\ &= -\frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) - \frac{1}{2}\int_0^1 \frac{(\dot{\alpha}^2\mu_\ell + \dot{\sigma}^2)(\alpha^2\mu_\ell + \sigma^2) - 2(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})}{(\alpha^2\mu_\ell + \sigma^2)^2}\mathrm{d}t \\ &= -\frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) - \frac{1}{2}\int_0^1 \frac{(\dot{\alpha}^2\mu_\ell + \dot{\sigma}^2)(\alpha^2\mu_\ell + \sigma^2) - 2(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t, \end{aligned} \tag{73}$$

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 (59).

Substituting the result from (73) back into the expression of E_ℓ^1 given in (72), we obtain

$$\begin{aligned} E_\ell^1 &= \frac{1}{2}\big(F_\ell(1) - F_\ell(0)\big) + \int_0^1 A_\ell(t)\,\mathrm{d}t - \int_0^1 B_\ell(t)\,\mathrm{d}t \\ &= -\frac{1}{2}\int_0^1 \frac{(\dot{\alpha}^2\mu_\ell + \dot{\sigma}^2)(\alpha^2\mu_\ell + \sigma^2) - 2(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t - \int_0^1 \frac{(\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{2(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t \\ &= -\frac{1}{2}\int_0^1 \frac{(\dot{\alpha}^2\mu_\ell + \dot{\sigma}^2)(\alpha^2\mu_\ell + \sigma^2) - (\alpha\dot{\alpha}\mu_\ell + \sigma\dot{\sigma})^2}{(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t \\ &= -\frac{\mu_\ell}{2}\int_0^1 \frac{(\alpha\dot{\sigma} - \sigma\dot{\alpha})^2}{(\alpha^2\mu_\ell + \sigma^2)^2}\,\mathrm{d}t. \end{aligned} \tag{74}$$

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 Proof of Theorem 1

In this appendix, we complement the result of Theorem 1 to the VE setting. The objective is identical to the VP setting shown in (21); 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:

$$\eta(0) = 0 \quad \text{and} \quad \lim_{t\to 1}\eta(1) = \sigma_{\max}. \tag{75}$$

Based on the definition $Q_\ell(t) \triangleq \frac{1}{\sqrt{\mu_\ell}} \arctan\left(\frac{\eta(t)}{\sqrt{\mu_\ell}}\right)$ from (22), 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 $L = \dot{Q}_\ell^2$ (25) 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:

$$Q_\ell(t) = \frac{t}{\sqrt{\mu_\ell}} \arctan\left(\frac{\sigma_{\max}}{\sqrt{\mu_\ell}}\right). \tag{76}$$

(*Remark:* As $\sigma_{\max} \to \infty$, this expression converges to the VP setting solution $Q_\ell(t) = \frac{\pi t}{2\sqrt{\mu_\ell}}$.)

Subsequently, by inverting the transformation $Q_\ell(t)$, we recover the corresponding tangent law schedule for the VE setting $\eta_\ell(t)\big|_{\mathrm{VE}}$:

$$\eta_\ell(t)\big|_{\mathrm{VE}} = \sqrt{\mu_\ell} \tan\left(t \cdot \arctan\left(\frac{\sigma_{\max}}{\sqrt{\mu_\ell}}\right)\right). \tag{77}$$

Appendix D
Supplementary Proof of Theorem 2

In this appendix, we derive an explicit expression of the leading order coefficient E_ℓ^1 under the parameterized tangent law schedule.

Firstly, by the tangent law (20) and (77) derived in Theorem 1, we redefine the parameterized tangent law schedule for both VP and VE setting

$$\eta_\gamma(t) \triangleq \frac{\sigma_\gamma(t)}{\alpha_\gamma(t)} = \sqrt{\gamma}\, \tan\bigl(\Theta_\gamma t\bigr), \quad \gamma > 0, \tag{78}$$

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

$$\Theta_\gamma \triangleq \arctan\left(\frac{\eta_1}{\sqrt{\gamma}}\right). \tag{79}$$

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:
$$\Theta_\gamma\big|_{\mathrm{VP}} = \frac{\pi}{2}.$$
- Variance Exploding (VE): Since $\alpha(t) \equiv 1$ and $\sigma(1) = \sigma_{\max}$, we have $\eta_1 = \sigma_{\max}$,
$$\Theta_\gamma\big|_{\mathrm{VE}} = \arctan\left(\frac{\sigma_{\max}}{\sqrt{\gamma}}\right) = \frac{\pi}{2} - \arctan\left(\frac{\sqrt{\gamma}}{\sigma_{\max}}\right).$$

Same as we mentioned before in Appendix B-A, 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 (74), we proceed as follows:

$$
\begin{aligned}
E_\ell^1(\gamma) &= -\frac{\mu_\ell}{2}\int_0^1 \frac{(\alpha\dot{\sigma}-\sigma\dot{\alpha})^2}{(\alpha^2\mu_\ell+\sigma^2)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell}{2}\int_0^1 \frac{\dot{\eta}(t)^2}{(\mu_\ell+\eta(t)^2)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell}{2}\int_0^1 \frac{\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{\gamma}\tan\left(\Theta_\gamma t\right)\right)\right]^2}{\left(\mu_\ell+\gamma\tan^2\left(\Theta_\gamma t\right)\right)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell}{2}\int_0^1 \frac{\gamma\Theta_\gamma^2\sec^4(\Theta_\gamma t)}{\left(\frac{\mu_\ell\cos^2(\Theta_\gamma t)+\gamma\sin^2(\Theta_\gamma t)}{\cos^2(\Theta_\gamma t)}\right)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell\gamma\Theta_\gamma^2}{2}\int_0^1 \frac{\sec^4(\Theta_\gamma t)\cdot\cos^4(\Theta_\gamma t)}{\left(\mu_\ell\cos^2(\Theta_\gamma t)+\gamma\sin^2(\Theta_\gamma t)\right)^2}\,\mathrm{d}t \\
&= -\frac{\mu_\ell\gamma\Theta_\gamma}{2}\int_0^{\Theta_\gamma} \frac{1}{\left(\mu_\ell\cos^2 x+\gamma\sin^2 x\right)^2}\,\mathrm{d}x \\
&= -\frac{\mu_\ell\gamma\Theta_\gamma}{2}\int_0^{\Theta_\gamma} \frac{1}{\left(\mu_\ell+(\gamma-\mu_\ell)\sin^2 x\right)^2}\,\mathrm{d}x.
\end{aligned} \tag{80}
$$

The standard integral result from [26, formula 2.563.1 and 2.562.1 in pp.177] is

$$
\begin{aligned}
\int \frac{\mathrm{d}x}{(a+b\sin^2 x)^2} &= \frac{1}{2a(a+b)}\left[(2a+b)\int\frac{\mathrm{d}x}{a+b\sin^2 x}+\frac{b\sin x\cos x}{a+b\sin^2 x}\right] \\
&= \frac{1}{2a(a+b)}\left[\frac{(2a+b)\,\mathrm{sign}(a)}{\sqrt{a(a+b)}}\arctan\left(\sqrt{\frac{a+b}{a}}\tan x\right)+\frac{b\sin x\cos x}{a+b\sin^2 x}\right]\ (b>-a),
\end{aligned} \tag{81}
$$

where sign denotes the sign function. Applying this to $E_\ell^1(\gamma)$ (80) with $a=\mu_\ell$ and $b=\gamma-\mu_\ell$, we obtain

$$
\begin{aligned}
E_\ell^1(\gamma) &= -\frac{\mu_\ell\gamma\Theta_\gamma}{2}\int_0^{\Theta_\gamma}\frac{1}{\left(\mu_\ell+(\gamma-\mu_\ell)\sin^2 x\right)^2}\,\mathrm{d}x \\
&= -\frac{\mu_\ell\gamma\Theta_\gamma}{2}\cdot\frac{1}{2\mu_\ell\gamma}\left[\frac{\mu_\ell+\gamma}{\sqrt{\mu_\ell\gamma}}\arctan\left(\sqrt{\frac{\gamma}{\mu_\ell}}\tan\Theta_\gamma\right)+\frac{(\gamma-\mu_\ell)\sin\Theta_\gamma\cos\Theta_\gamma}{\mu_\ell\cos^2\Theta_\gamma+\gamma\sin^2\Theta_\gamma}\right] \\
&= -\frac{\Theta_\gamma}{4}\left[\frac{\mu_\ell+\gamma}{\sqrt{\mu_\ell\gamma}}\arctan\left(\sqrt{\frac{\gamma}{\mu_\ell}}\tan\Theta_\gamma\right)+\frac{(\gamma-\mu_\ell)\tan\Theta_\gamma}{\mu_\ell+\gamma\tan^2\Theta_\gamma}\right].
\end{aligned} \tag{82}
$$

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

$$
E_\ell^1(\gamma)\big|_{\mathrm{VP}} = -\frac{\pi/2}{4}\left(\frac{\mu_\ell+\gamma}{\sqrt{\mu_\ell\gamma}}\cdot\frac{\pi}{2}\right) = -\frac{\pi^2}{16}\left(\sqrt{\frac{\mu_\ell}{\gamma}}+\sqrt{\frac{\gamma}{\mu_\ell}}\right), \tag{83}
$$

thereby recovering the expression presented in (28).

Appendix E
Implementation Details

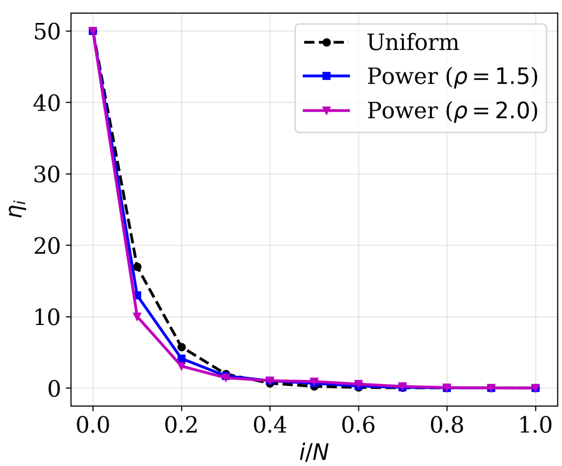
In this appendix, we provide detailed implementation settings and results for the applications discussed in Section III-C and Section IV.

a) Time discretization Strategy: Recall that the *power-uniform* discretization schedule is defined as:

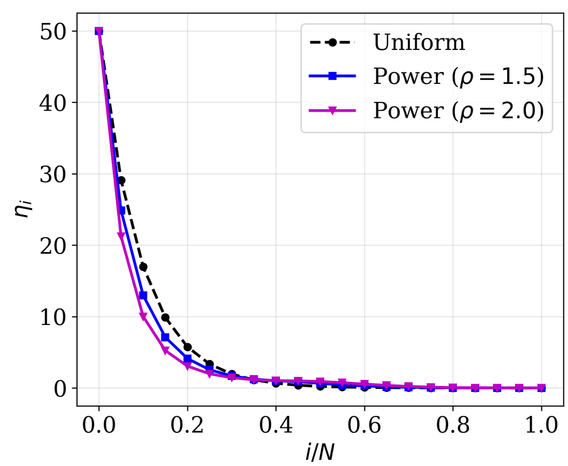
$$\lambda_i^{(\rho)} = \left(\lambda_{t_N}^{1/\rho} + \frac{i}{N}\left(\lambda_{t_0}^{1/\rho} - \lambda_{t_N}^{1/\rho}\right)\right)^{\rho}, \quad i = 0, \dots, N, \tag{84}$$

where $\lambda = \log(\alpha(t)/\sigma(t))$ represents the half-logSNR. The boundary values λ_{t_0} and λ_{t_N} are determined by the discrete values of noise schedule $\alpha(t)$ and $\sigma(t)$ used in the training stage (4). The hyperparameter ρ controls the distribution of the step size, with $\rho = 1$ corresponding to the widely used *uniform-λ* strategy.

To accommodate negative values in the base (i.e., when $\lambda < 0$), we employ 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 2, the *power-uniform* discretization with $\rho > 1$ induces a more rapid decay of η in the high-noise regime (where $\lambda < 0$), corresponding to the early stage of the reverse sampling process.



(a) Steps $N = 10$.



(b) Steps $N = 20$.

Fig. 2: Visualization of the $\eta_i = e^{-\lambda_i}$ under different time discretization settings.

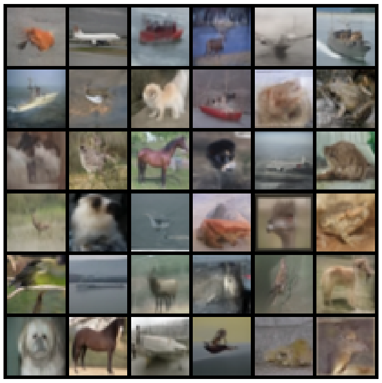
Based on the *power-uniform* discretization of λ, time steps t_i used in the reverse sampling process (3) are recovered via $t_i = \lambda^{-1}(\lambda_i^{(\rho)})$, which depends on the training choices of $\alpha(t)$ and $\sigma(t)$. For the specific case of the cosine schedule, defined as $\alpha(t) = \cos(\frac{\pi}{2}t)$ and $\sigma(t) = \sin(\frac{\pi}{2}t)$, this relationship yields the solution $t_i = \frac{2}{\pi}\arctan(e^{-\lambda_i^{(\rho)}})$.

b) Synthetic Distribution: We model the data 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:

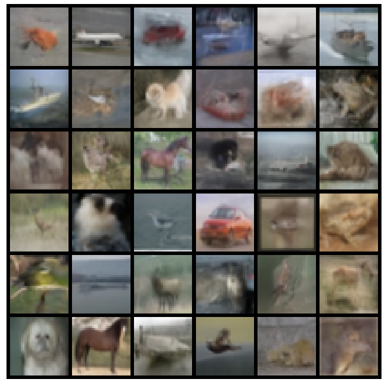
$$\mu_\ell = C \cdot (\ell + i_0)^{-p} + \varepsilon, \qquad \ell = 1, \dots, k, \tag{85}$$

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 $\mu_k \geq \mu_{\max}$.

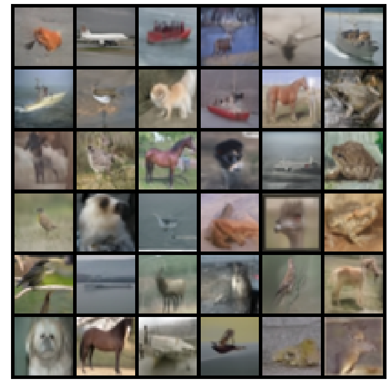
c) Generated Images Comparison: In this part, we provide further qualitative comparisons of samples generated by pretrained models using general ODE samplers under different time discretization strategies.



(a) $\rho = 1.5$.

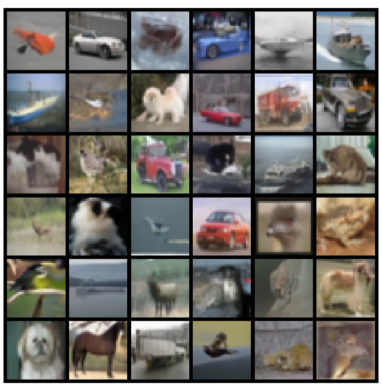


(b) Uniform-λ.

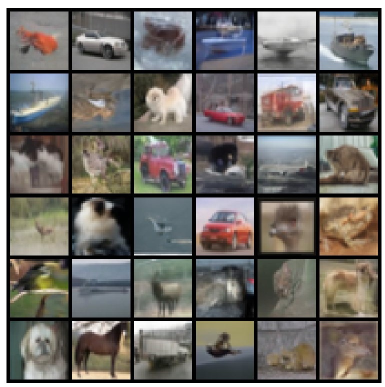


(c) Xue'24.

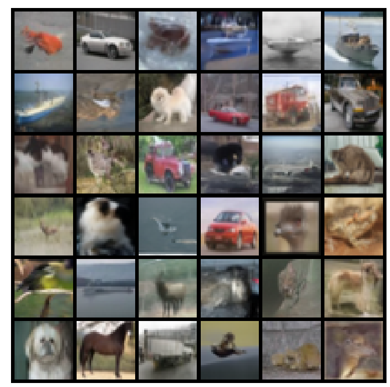
Fig. 3: EDM checkpoint with DPM-Solver++ on CIFAR-10 (NFEs=6).



(a) $\rho = 1.5$.

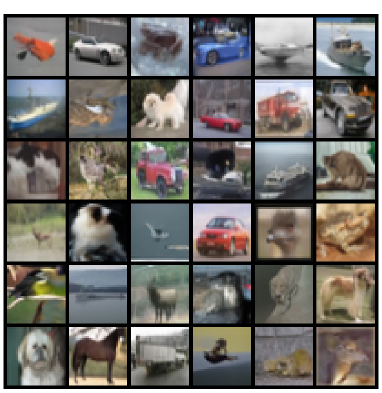


(b) Uniform-λ.

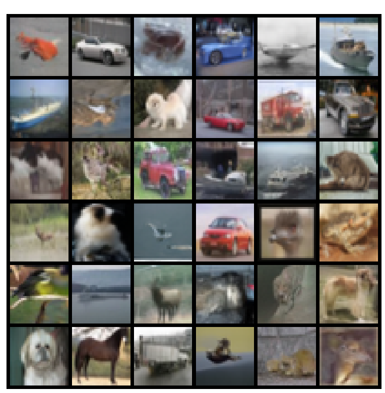


(c) Xue'24.

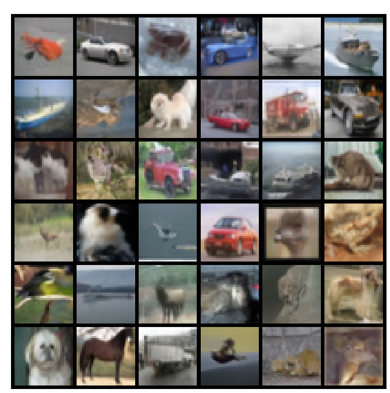
Fig. 4: EDM checkpoint with DPM-Solver++ on CIFAR-10 (NFEs=12).



(a) $\rho = 1.5$.

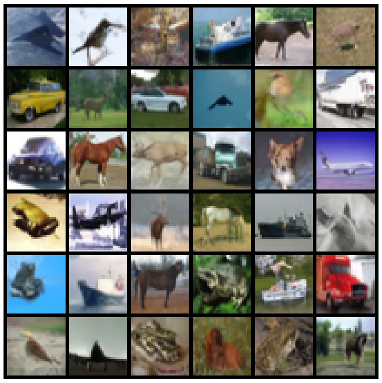


(b) Uniform-λ.

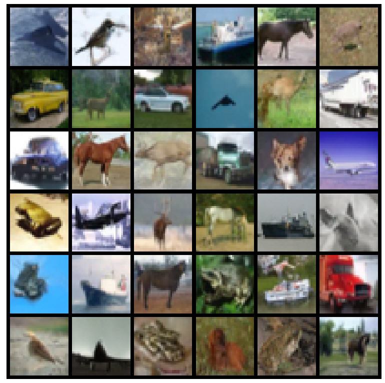


(c) Xue'24.

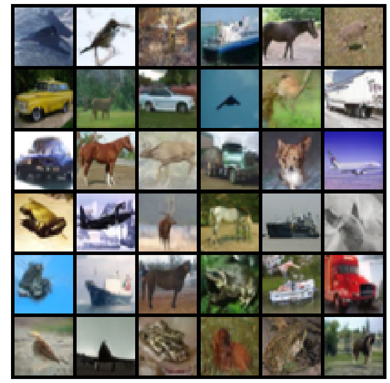
Fig. 5: EDM checkpoint with DPM-Solver++ on CIFAR-10 (NFEs=15).



(a) $\rho = 2.0$.

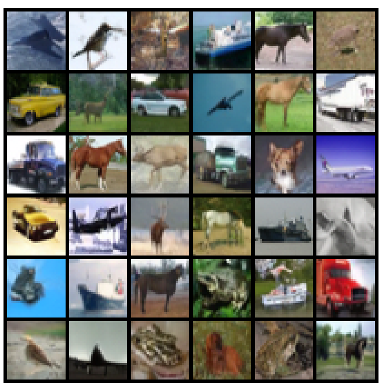


(b) Uniform-λ.

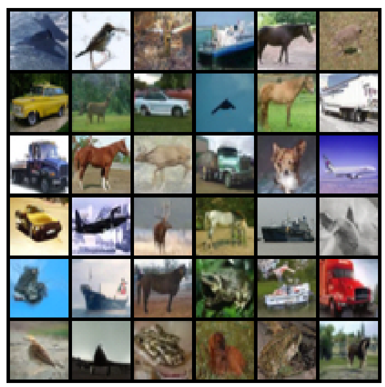


(c) Xue'24.

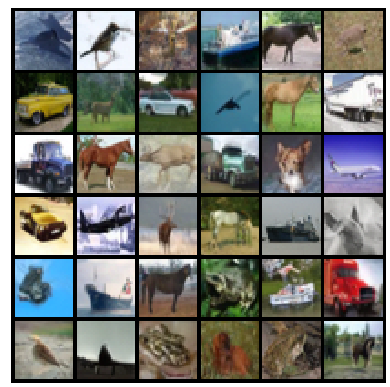
Fig. 6: VP-SDE checkpoint with UniPC on CIFAR-10 (NFEs=20).



(a) $\rho = 2.0$.

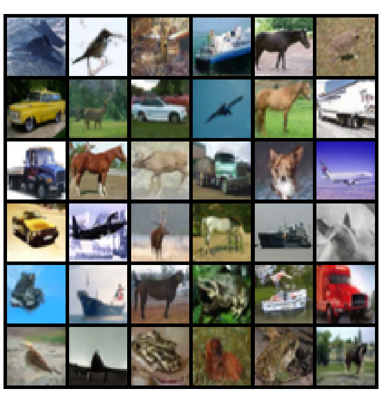


(b) Uniform-λ.

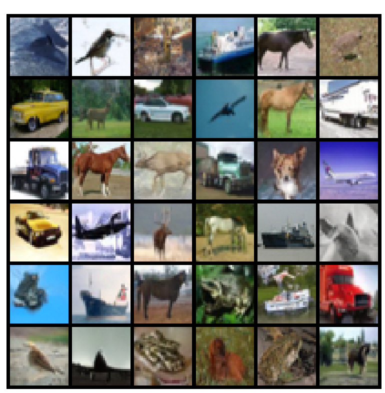


(c) Xue'24.

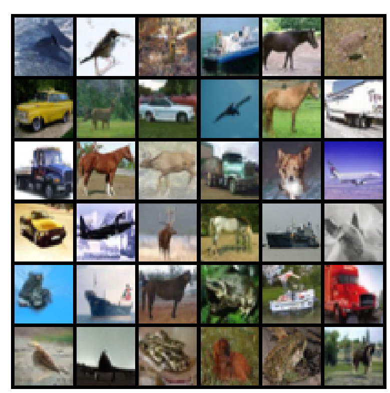
Fig. 7: VP-SDE checkpoint with UniPC on CIFAR-10 (NFEs=25).



(a) $\rho = 2.0$.



(b) Uniform-λ.



(c) Xue'24.

Fig. 8: VP-SDE checkpoint with UniPC on CIFAR-10 (NFEs=50).

(a) $\rho = 1.5$.

(b) Uniform-λ.

(c) Xue'24.

Fig. 9: EDM checkpoint with DPM-Solver++ on FFHQ-64 (NFEs=8).

(a) $\rho = 1.5$.

(b) Uniform-λ.

(c) Xue'24.

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

(a) $\rho = 1.5$.

(b) Uniform-λ.

(c) Xue'24.

Fig. 11: EDM checkpoint with DPM-Solver++ on FFHQ-64 (NFEs=12).