

DC-DPM: A DIVIDE-AND-CONQUER APPROACH FOR DIFFUSION REVERSE PROCESS

Anonymous authors

Paper under double-blind review

ABSTRACT

Diffusion models have achieved great success in generative tasks. However, previous approaches typically approximate the reversed transition kernel with a Gaussian distribution. This approximation can diverge from real scenarios, necessitating multiple iterative steps for high-quality sample generation and limiting the real-time inference performance of diffusion models. In this paper, we propose a **Divide-and-Conquer** strategy to improve the traditional single Gaussian transition kernel representation in each denoising step of **Diffusion Probabilistic Models** (DC-DPM), thus enhancing generation quality particularly over a limited number of timesteps. By dividing the data into clusters, our DC-DPM learns specific kernels for each partition. We design two merging strategies for these cluster-specific kernels along with corresponding training and sampling methods. We provide theoretical proof of DC-DPM’s convergence to the true data distribution from a novel perspective. Experimental results demonstrate the superior generation quality of our method compared to the traditional single Gaussian kernel. Furthermore, our DC-DPM can synergize with previous kernel optimization methods, enhancing their generation quality, especially with a small number of timesteps.

1 INTRODUCTION

Diffusion models have recently gained prominence in generating multi-modal content across various tasks, including image generation (Dhariwal & Nichol, 2021; Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022), image super-resolution (Li et al., 2022), video generation (Ho et al., 2022a;b), text-to-speech synthesis (Popov et al., 2021), 3D generation (Poole et al., 2022), and motion planning (Carvalho et al., 2023). Diffusion models generate data by iteratively predicting noise and solving diffusion SDEs to denoise (Song et al., 2020b). In each step, the transition kernel is approximated as a Gaussian, which can significantly differ from the true transition kernel (Guo et al., 2024). Therefore, hundreds or thousands of denoising steps are needed to produce high-quality data, limiting the real-time applicability of diffusion models.

Extensive research progress has been made to accelerate the diffusion reverse process. Besides the training-based methods including knowledge distillation (Sauer et al., 2023; Meng et al., 2023; Song et al., 2023; Luo et al., 2023) and training-free methods involving faster ODE/SDE solvers (Lu et al., 2022a;b; Zheng et al., 2023b; Xu et al., 2023; Sabour et al., 2024; Li et al., 2024), Analytic-DPM (Bao et al., 2022b) and Extended-Analytic-DPM (Bao et al., 2022a) estimate the optimal variance for the Gaussian transition kernel. GMS (Guo et al., 2024) represents the kernel as a Gaussian mixture by estimating high-order moments. However, due to computational constraints, GMS uses a mixture of only two Gaussians combined with hand-crafted weights. Furthermore, the sampling process in GMS involves multiple approximation operations, making it challenging to ensure convergence to the true data distribution for GMS.

In this paper, we propose DC-DPM, a novel approach to represent the transition kernel in the reverse process of diffusion probabilistic models (DPM) using a Divide-and-Conquer strategy. We first provide a proof for the convergence of the conventional single Gaussian kernel from a novel perspective without using Kolmogorov equations, and then generalize this convergence proof from the case of a single Gaussian kernel to our proposed divide-and-conquer transition kernel method. We provide theoretical proof that our proposed divide-and-conquer representation of the transition kernel converges to the true underlying data distribution, regardless of the data division patterns.

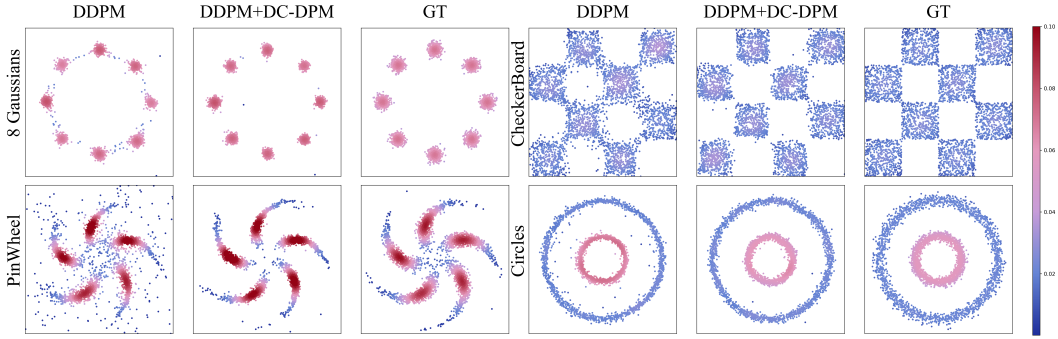


Figure 1: **DC-DPM improves generation quality on small timesteps.** Employing a divide-and-conquer approach to approximate transition kernels in diffusion reverse process, DC-DPM generates samples closer to ground truth distribution (GT) on 20 denoising steps. Colors represent data density.

Starting by clustering the data into different partitions, DC-DPM learns reversed transition kernels for each data cluster and models the overall transition kernel as a composition of these cluster-specific kernels. To determine how to combine the cluster-specific kernels, we design two strategies along with corresponding training and sampling methods. DC-DPM can collaborate with previous diffusion sampling optimization methods, particularly those focused on optimizing the Gaussian transition kernel design, such as Extended-Analytic-DPM and GMS, by utilizing the representations proposed in these works for each cluster-specific kernel in our DC-DPM.

Experimental results on 2D toy datasets and image datasets demonstrate that our method enhances the generation quality of diffusion models compared to the traditional single Gaussian transition kernel representation. Furthermore, our approach significantly improves the performance of previous transition kernel optimization methods, including Extended-Analytic-DPM and GMS, especially in scenarios with limited sampling steps.

Proofs for all Propositions are given in the Appendix.

2 BACKGROUND: DIFFUSION PROBABILISTIC MODELS AND ITS CONVERGENCE

2.1 DIFFUSION PROBABILISTIC MODELS AND TRANSITION KERNELS

Given a finite set of data samples $\{\mathbf{y}_i \in \mathbb{R}^d | i = 1, 2, \dots, N\}$, where d represents the data dimension and N is the number of samples. The distribution of these samples is characterized by:

$$p_{data}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{y}_i), \tag{1}$$

where $\delta(\mathbf{x})$ represents Dirac delta function. As real training processes are typically conducted on such finite datasets, we assume that the ground truth data distribution adheres to Eq. (1).

Diffusion probabilistic models define two Markov chains including forward process and reverse process. The forward process is typically hand-designed with Gaussian transition kernel to perturb data to noise and can be expressed as (Zhang et al., 2024):

$$p(\mathbf{x}_t, t | \mathbf{x}_s, s) = \mathcal{N}(\mathbf{x}_t; \alpha_{t|s} \mathbf{x}_s, \sigma_{t|s}^2 \mathbf{I}), \tag{2}$$

where t, s are two timesteps and $0 \leq s < t \leq 1$. $\alpha_{t|s} = \frac{\alpha_t}{\alpha_s}$ and $\sigma_{t|s} = \sqrt{1 - \alpha_{t|s}^2}$, where α_t is a hyperparameter which decreases monotonically from 1 to 0 over time t (Kingma et al., 2021).

Hence, the conditional distribution of \mathbf{x}_t given \mathbf{x}_0 can be derived as:

$$p(\mathbf{x}_t, t | \mathbf{x}_0, 0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}). \tag{3}$$

Taking the initial condition Eq. (1) into account, the single time marginal distribution of \mathbf{x}_t is:

$$p(\mathbf{x}_t, t) = \frac{1}{N} \sum_i (2\pi\sigma_t^2)^{-\frac{d}{2}} \exp\left(-\frac{\|\mathbf{x}_t - \alpha_t \mathbf{y}_i\|^2}{2\sigma_t^2}\right). \tag{4}$$

The reverse process reverses the forward one with a learned kernel. Based on Eq. (2) and (4), the ground truth reversed transition kernel can be derived with Bayes' rule:

$$\begin{aligned} p(\mathbf{x}_s, s | \mathbf{x}_t, t) &= p(\mathbf{x}_t, t | \mathbf{x}_s, s) \frac{p(\mathbf{x}_s, s)}{p(\mathbf{x}_t, t)} \\ &= (2\pi\sigma_{s|t})^{-\frac{d}{2}} \sum_i w_i(\mathbf{x}_t, t) \exp\left\{-\frac{1}{2\sigma_{s|t}^2} \left\| \mathbf{x}_s - \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2} \mathbf{x}_t - \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2} \mathbf{y}_i \right\|^2\right\}, \end{aligned} \quad (5)$$

where $\sigma_{s|t} = \sigma_{t|s} \frac{\sigma_s}{\sigma_t}$. $w_i(\mathbf{x}_t, t) = \frac{u_i(\mathbf{x}_t, t)}{\sum_j u_j(\mathbf{x}_t, t)}$ while $u_i(\mathbf{x}_t, t) = \exp(-\frac{\|\mathbf{x}_t - \alpha_t \mathbf{y}_i\|^2}{2\sigma_t^2})$.

Existing methods typically approximate the learnable reversed transition kernel as a single Gaussian distribution (Ho et al., 2020). The transition kernel can be expressed as (Zhang et al., 2024):

$$\tilde{p}(\mathbf{x}_s, s | \mathbf{x}_t, t) = (2\pi\sigma_{s|t})^{-\frac{d}{2}} \exp\left\{-\frac{1}{2\sigma_{s|t}^2} \left\| \mathbf{x}_s - \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2} \mathbf{x}_t - \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2} \bar{\mathbf{y}}(\mathbf{x}_t, t) \right\|^2\right\}, \quad (6)$$

The mean of this Gaussian distribution is related to $\bar{\mathbf{y}}(\mathbf{x}_t, t)$, which is estimated by a neural network $\mathbf{y}_\theta(\mathbf{x}_t, t)$ in \mathbf{x} -prediction methods, while the variance is isotropic and only depends on the timestep s and t . Another commonly used parameterization is ϵ -prediction, which employs a noise prediction network to estimate the noise $\epsilon(\mathbf{x}_t, t)$ (Salimans & Ho, 2022). Despite its difference from \mathbf{x} -prediction, these two parameterizations are equivalent, as demonstrated by the relationship

$$\mathbf{x}_t = \alpha_t \bar{\mathbf{y}}(\mathbf{x}_t, t) + \sigma_t \epsilon(\mathbf{x}_t, t). \quad (7)$$

In this paper, we utilize \mathbf{x} -prediction for simplicity in our proofs. For clarity, we will refer to $p(\mathbf{x}_t)$ and $p(\mathbf{x}_s | \mathbf{x}_t)$ instead of $p(\mathbf{x}_t, t)$ and $p(\mathbf{x}_s, s | \mathbf{x}_t, t)$ when there is no ambiguity.

2.2 CONVERGENCE WITH KOLMOGOROV EQUATIONS

Define $f_t = \frac{d \log \alpha_t}{dt}$ and $g_t = -2f_t$, and the stochastic differential equation (SDE)

$$d\mathbf{x}_t = f_t \mathbf{x}_t dt + g_t d\mathbf{B}_t, \quad (8)$$

where \mathbf{B}_t is the standard Brownian motion. According to Anderson (1982), its reverse process is

$$d\mathbf{x}_t = (f_t \mathbf{x}_t - g_t^2 \nabla_{\mathbf{x}_t} p(\mathbf{x}_t)) dt + g_t d\tilde{\mathbf{B}}_t. \quad (9)$$

Previous efforts to prove the convergence of DPM heavily depend on the Kolmogorov equations of Eq. (9) For instance, Lee et al. (2022) defines the discretization approximation

$$d\mathbf{x}_t = (f_t \mathbf{x}_t - g_t^2 \nabla_{\mathbf{x}_t} p(\mathbf{x}_t)) dt + g_t d\tilde{\mathbf{B}}_t, \quad (10)$$

and establish the corresponding Kolmogorov forward equation for the single time marginal distribution of Eq. (10), denoted as $q(\mathbf{x}_t)$. Ultimately, the Chi-square divergence $\chi^2(q(\mathbf{x}_t) || p(\mathbf{x}_t))$ is estimated using the Kolmogorov equations. Numerous subsequent studies have embraced this configuration (Lee et al., 2023; Chen et al., 2022; 2023b;a). However, this proof has its limitations as it's based on the Kolmogorov equations. This means it cannot be applied to other types of discretizations where constructing the Kolmogorov equations is challenging. Therefore, a proof that can be readily adapted to a wider range of discretizations would be beneficial.

3 METHOD

In this section, we first introduce a novel method to demonstrate that the distribution generated by conventional diffusion model closely matches the actual data distribution without using Kolmogorov equations. We then propose to approximate the reverse process transition kernel in a divide-and-conquer manner and prove its convergence using this novel method. We further propose merging strategies for these kernels and present the corresponding training and sampling methods.

To start with, we outline some assumptions regarding the initial distribution and the neural network approximation errors, which will be referenced throughout this paper:

Assumption 1 The initial distribution is a sum of Dirac deltas and $\max_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\| \leq M$ for some positive constant M .

Assumption 2 For all $t \in [0, 1]$, \mathbf{y}_θ and $\bar{\mathbf{y}}$ are close in $L^2(p)$:

$$\int_{\mathbb{R}^d} p(\mathbf{x}_t) \|\mathbf{y}_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t)\|^2 d\mathbf{x}_t < \varepsilon_y^2 < 1. \quad (11)$$

Assumption 3 α_t is a predefined function, which decreases monotonically from 1 to 0, with its derivatives bounded; specifically, $0 > \frac{d\alpha_t}{dt} \geq -C_\alpha$ for some positive constant C_α .

3.1 CONVERGENCE OF DPM FROM A NOVEL PERSPECTIVE

Considering that the sampling process occurs in discrete steps, we introduce the notation for time discretization as $\mathcal{D} = \{0 < t_{\min} = t_0 < t_1 < \dots < t_T = t_{\max} < 1\}$. Subsequently, the approximated single-time marginal distribution with the accurate $\bar{\mathbf{y}}(\mathbf{x}_t, t)$ is:

$$\tilde{p}(\mathbf{x}_{t_i}) = \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} \tilde{p}(\mathbf{x}_{t_i} | \mathbf{x}_{t_{i+1}}) \dots \tilde{p}(\mathbf{x}_{t_{T-1}} | \mathbf{x}_{t_T}) \tilde{p}(\mathbf{x}_{t_T}) d\mathbf{x}_{t_{i+1}} \dots d\mathbf{x}_{t_T}. \quad (12)$$

By substituting $\bar{\mathbf{y}}(\mathbf{x}_t, t)$ in Eq. (12) and (6) with the network prediction $\mathbf{y}_\theta(\mathbf{x}_t, t)$, we obtain $p_\theta(\mathbf{x}_{t_i})$ and $p_\theta(\mathbf{x}_{t_i} | \mathbf{x}_{t_{i+1}})$. We also define $\Delta t_i = t_{i+1} - t_i$ and denote the maximum Δt_i as $|\mathcal{D}|$.

As pointed out in previous study (Zhang et al., 2024), singularities arise near $t = 0$ and $t = 1$, necessitating specific treatment. To address this, we divide the time interval into three distinct segments: the left interval $[0, t_{\min}]$, the middle interval $[t_{\min}, t_{\max}]$, and the right interval $(t_{\max}, 1]$. Each section is handled independently. Previous work provides local error estimates for the middle and right intervals (Zhang et al., 2024). However, their assertions are not strong enough to achieve global convergence. We enhance these estimates to ensure global convergence. We refine the error bound for the middle interval from $(t - s)^{\frac{1}{4}}$ (Zhang et al., 2024) to $(t - s)^{\frac{3}{2} - \beta}$.

Proposition 1 For all $t_{\min} \leq s < t \leq t_{\max}$ and $0 < \beta < 1$, there exist $\delta > 0$ and $C_1, C_2 > 0$ depending on β, t_{\min} and t_{\max} , such that if $t - s < \delta$, the inequality $KL(p(\mathbf{x}_s) || p_\theta(\mathbf{x}_s)) \leq KL(p(\mathbf{x}_t) || p_\theta(\mathbf{x}_t)) + C_1(t - s)^{\frac{3-\beta}{2}} + C_2(t - s)\varepsilon_y$ holds.

For the right interval, we improve the error bound from $(1 - s)^{\frac{1}{2}}$ (Zhang et al., 2024), to $(1 - s)^2$.

Proposition 2 For all $0 < s < 1$, there are constants $C_1, C_2 > 0$, such that $KL(p(\mathbf{x}_s) || p_\theta(\mathbf{x}_s)) \leq C_1(1 - s)^2 + C_2(1 - s)^2\varepsilon_y$.

As for the left interval, it's not feasible to compute the Kullback-Leibler divergence for Dirac deltas. We adapt the idea from Theorem 2.1 in prior work (Lee et al., 2023), which applies the Wasserstein distance to the left interval.

Proposition 3 Given $0 < t_{\min} < 1$, the 2-Wasserstein distance

$$W_2(p(\mathbf{x}_0), p(\mathbf{x}_{t_{\min}})) < \sqrt{2dC_\alpha t_{\min}}. \quad (13)$$

By combining the local error bounds above, we can establish global convergence as follows:

Proposition 4 For all $0 < \beta < 1$, there exist $\delta > 0$ and $C_1, C_2, C_3 > 0$, such that for all time discretizations \mathcal{D} with $|\mathcal{D}| < \delta$, the Kullback-Leibler divergence $KL(p(\mathbf{x}_{t_{\min}}) || p_\theta(\mathbf{x}_{t_{\min}})) \leq C_1|\mathcal{D}|^{\frac{1-\beta}{2}} + C_2\varepsilon_y$. Moreover, $W_2^2(p(\mathbf{x}_0), p(\mathbf{x}_{t_{\min}})) < C_3|\mathcal{D}|$.

3.2 DIVIDE-AND-CONQUER APPROXIMATION AND ITS CONVERGENCE

Based on the analysis above, besides the single Gaussian transition kernel, any distribution submitted to Proposition 4 could serve as the transition kernel.

As demonstrated in Eq. (5), the ground truth transition kernel of the diffusion reverse process is a mixture of standard Gaussian distributions. Previous work provides evidence that traditional approaches, which approximate this Gaussian mixture kernel with a single Gaussian distribution, can significantly diverge from the true reverse transition kernel (Guo et al., 2024). This motivates our *divide-and-conquer* (DC) transition kernel approximation.

Specifically, we propose to partition data and use cluster-specific kernels to represent data samples in each segment. The true kernel is then approximated by integrating these cluster-specific kernels. Consider a scenario where the training data is divided into L classes: $\{\mathbf{y}_i \in \mathbb{R}^d | i = 1, 2, \dots, N\} = \bigcup_{l=1}^L \{\mathbf{y}_i^l \in \mathbb{R}^d | i = 1, 2, \dots, N_l\}$. This partition can be arbitrary, and we will prove that any method of data division result in convergence. We define a new approximation

$$\hat{p}(\mathbf{x}_s | \mathbf{x}_t) = \sum_{l=1}^L a^l(\mathbf{x}_t, t) \hat{p}^l(\mathbf{x}_s | \mathbf{x}_t), \quad (14)$$

where

$$\hat{p}^l(\mathbf{x}_s | \mathbf{x}_t) = \sum_{i=1}^{N_l} (2\pi\sigma_{s|t})^{-\frac{d}{2}} u_i^l(\mathbf{x}_t, t) \exp\left\{-\frac{1}{2\sigma_{s|t}^2} \left\| \mathbf{x}_s - \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2} \mathbf{x}_t - \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2} \bar{\mathbf{y}}^l(\mathbf{x}_t, t) \right\|^2\right\}, \quad (15)$$

$$w_i^l(\mathbf{x}_t, t) = \frac{\exp\left\{-\frac{\|\mathbf{x}_t - \alpha_t \mathbf{y}_i^l\|^2}{2\sigma_t^2}\right\}}{\sum_{l,j} \exp\left\{-\frac{\|\mathbf{x}_t - \alpha_t \mathbf{y}_j^l\|^2}{2\sigma_t^2}\right\}}, \quad a^l(\mathbf{x}_t, t) = \sum_{i=1}^{N_l} w_i^l(\mathbf{x}_t, t), \quad u_i^l(\mathbf{x}_t, t) = \frac{w_i^l(\mathbf{x}_t, t)}{a^l(\mathbf{x}_t, t)} \quad \text{and} \quad \bar{\mathbf{y}}^l(\mathbf{x}_t, t) =$$

$\sum_{i=1}^{N_l} u_i^l(\mathbf{x}_t, t) \mathbf{y}_i^l$. The single time marginal distribution of this approximation $\hat{p}(\mathbf{x}_t)$ is defined in the same way as in Eq. (12). Each cluster-specific kernel $\hat{p}^l(\mathbf{x}_s | \mathbf{x}_t)$ can be approximated using any method suitable for a standard diffusion probabilistic model, including a single Gaussian approximation in DDPM, single Gaussian with optimized variances in Extended-Analytic-DPM (Bao et al., 2022a), as well as GMS (Guo et al., 2024), which computes high-order moments and estimates $\hat{p}^l(\mathbf{x}_s | \mathbf{x}_t)$ as a mixture of two Gaussians with hand-crafted weights.

In this scenario, we need L neural networks to approximate $\bar{\mathbf{y}}^l(\mathbf{x}_t, t)$. In practice, we use a conditional network $\mathbf{y}_\theta(\mathbf{x}_t, t, l)$ (also denoted as $\mathbf{y}_\theta^l(\mathbf{x}_t, t)$). Additionally, a neural network $\mathbf{a}_\phi(\mathbf{x}_t, t)$ is necessary to approximate $\mathbf{a}(\mathbf{x}_t, t) \stackrel{\text{def}}{=} (a^1(\mathbf{x}_t, t), \dots, a^L(\mathbf{x}_t, t))^T \in \mathbb{R}^L$. To derive the error bound, we also make the assumption that $\mathbf{y}_\theta^l(\mathbf{x}_t, t)$ and $\mathbf{a}_\phi(\mathbf{x}_t, t)$ approximate $\bar{\mathbf{y}}^l(\mathbf{x}_t, t)$ and $a^l(\mathbf{x}_t, t)$ in $L^2(p)$.

Assumption 4 For all $t \in [t_{\min}, 1]$ and $1 \leq l \leq L$, \mathbf{y}_θ^l and \mathbf{a}_ϕ^l are close to $\bar{\mathbf{y}}^l$ and a^l in $L^2(p)$ respectively:

$$\int_{\mathbb{R}^d} p(\mathbf{x}_t) \|\mathbf{y}_\theta^l(\mathbf{x}_t, t) - \bar{\mathbf{y}}^l(\mathbf{x}_t, t)\|^2 d\mathbf{x}_t < \varepsilon_{yl}^2 < 1, \quad (16)$$

and

$$\int_{\mathbb{R}^d} p(\mathbf{x}_t) (a_\phi^l(\mathbf{x}_t, t) - a^l(\mathbf{x}_t, t))^2 d\mathbf{x}_t < \varepsilon_{al}^2 < 1. \quad (17)$$

Moreover, $a_\phi^l(\mathbf{x}_t, t)$ and $a^l(\mathbf{x}_t, t)$ are uniformly lower bounded by a constant C_a .

Then, $\hat{p}(\mathbf{x}_s | \mathbf{x}_t)$ is defined by \mathbf{y}_θ^l s and \mathbf{a}_ϕ^l s. $\hat{p}_\theta(\mathbf{x}_t)$ is defined in a manner consistent with equation (12).

To estimate the error boundary of the Divide-and-Conquer Diffusion Probabilistic Models (DC-DPM), we employ a strategy that transforms it into a single Gaussian case. Taking into account

270 that

$$271 \quad p(\mathbf{x}_s|\mathbf{x}_t) = \sum_{l=1}^L a^l(\mathbf{x}_t, t) p^l(\mathbf{x}_s|\mathbf{x}_t), \quad (18)$$

274 where

$$275 \quad p^l(\mathbf{x}_s|\mathbf{x}_t) = \sum_{i=1}^{N_l} (2\pi\sigma_{s|t})^{-\frac{d}{2}} u_i^l(\mathbf{x}_t, t) \exp\left\{-\frac{1}{2\sigma_{s|t}^2} \left\| \mathbf{x}_s - \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2} \mathbf{x}_t - \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2} \mathbf{y}_i^l \right\|^2\right\}, \quad (19)$$

278 and given the convexity of the Kullback-Leibler divergence, we can deduce that

$$281 \quad KL(p(\mathbf{x}_s|\mathbf{x}_t)||\hat{p}(\mathbf{x}_s|\mathbf{x}_t)) \leq \sum_l a^l(\mathbf{x}_t, t) KL(p^l(\mathbf{x}_s|\mathbf{x}_t)||\hat{p}^l(\mathbf{x}_s|\mathbf{x}_t)). \quad (20)$$

282 Eq. (20) allows us to bound the error of the divide-and-conquer approximations by the sum of its individual components. Utilizing Propositions 1, 2, and 4, we can deduce the following corollaries:

286 **Corollary 1** For all $t_{min} \leq s < t \leq t_{max}$ and $0 < \beta < 1$, there exist $\delta > 0$ and $C_1, C_2, C_3 > 0$ depending on β, t_{min} and t_{max} , such that if $t - s < \delta$, the inequality $KL(p(\mathbf{x}_s)||\hat{p}_\theta(\mathbf{x}_s)) \leq KL(p(\mathbf{x}_t)||\hat{p}_\theta(\mathbf{x}_t)) + C_1(t-s)^{\frac{3-\beta}{2}} + C_2(t-s)\varepsilon_{yl} + C_3\varepsilon_{al}$ holds.

291 **Corollary 2** For all $0 < s < 1$, there are constants $C_1, C_2, C_3 > 0$, such that $KL(p(\mathbf{x}_s)||p_\theta(\mathbf{x}_s)) \leq C_1(1-s)^2 + C_2\varepsilon_{yl} + C_3\varepsilon_{al}$.

294 **Corollary 3** For all $0 < \beta < 1$, there exist $\delta > 0$ and $C_1, C_2, C_3, C_4 > 0$, such that for all time discretizations \mathcal{D} with $|\mathcal{D}| < \delta$, the Kullback-Leibler divergence $KL(p(\mathbf{x}_{t_{min}})||\hat{p}_\theta(\mathbf{x}_{t_{min}})) \leq C_1|\mathcal{D}|^{\frac{1-\beta}{2}} + C_2\varepsilon_{yl} + C_3T\varepsilon_{al}$. Moreover, $W_2(p(\mathbf{x}_0), p(\mathbf{x}_{t_{min}})) < C_4|\mathcal{D}|$.

300 It is worthy to note that Corollaries above can not be easily proved with the methods based on Kolmogorov equations as in the previous works, because it is not trivial to construct a Ito diffusion with equation (14) being the solution to its corresponding Kolmogorov equations.

303 The term ε_{al} in Corollary 3 includes a coefficient T , representing the inference time step. This factor inhibits the error from converging to zero as $|\mathcal{D}|$ approaches zero. However, due to the simplistic structure of $\mathbf{a}(\mathbf{x}_t, t)$, the network $\mathbf{a}_\phi(\mathbf{x}_t, t)$ is relatively easy to train. This results in ε_{al} being significantly smaller than ε_{yl} . Consequently, this maintains the error of DC-DPM at a reasonably low value.

309 3.3 MERGING CLUSTER-SPECIFIC KERNELS

310 The divide-and-conquer representation of the reversed transition in Eq. (14) consists of combination coefficients $a^l(\mathbf{x}_t, t)$, referred to as the *class part*, and cluster-specific kernels $p^l(\mathbf{x}_s|\mathbf{x}_t)$, referred to as the *diffusion part*. For the diffusion part, to learn L cluster-specific kernels, we propose training a single conditional network $\mathbf{y}_\theta(\mathbf{x}_t, t, l)$ to represent them, rather than training L independent networks, in order to save computational overhead. For the class part, we propose two approaches to estimate it: *label diffusion approximation (LD)* and *fixed class approximation (FC)*.

317 Label diffusion approximation (LD) learns the class part in a manner similar to the diffusion part. Define L_i as the one-hot vector representing the class to which data point y_i belongs. Then we construct $a(\mathbf{x}_t, t)$ as:

$$320 \quad a(\mathbf{x}_t, t) = \sum_i w_i(\mathbf{x}_t, t) L_i, \quad (21)$$

322 where $w_i(\mathbf{x}_t, t)$ represents the coefficients in the ground truth transition kernel in Eq. (5). Substituting Eq. (21) into Eq. (18) aligns with the ground truth transition kernel in Eq. (5). Given that the structure of $\mathbf{a}(\mathbf{x}_t, t)$ closely mirrors that of $\bar{\mathbf{y}}(\mathbf{x}_t, t)$, a neural network $\mathbf{a}_\phi(\mathbf{x}_t, t)$ can be trained in a

manner analogous to the \mathbf{x} -prediction networks in diffusion models. The training process to learn the class part can be formulated as:

Proposition 5 Let $L(\mathbf{x}_0)$ denote the one-hot class vector of \mathbf{x}_0 , the optimal $\mathbf{a}_\phi(\mathbf{x}_t, t)$ for the two objective functions

$$\mathcal{L}_2 = \mathbb{E}_{\mathbf{x}_0 \sim p_{data}, \mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{x}_0), t \sim \mathcal{U}(0,1)} \|L(\mathbf{x}_0) - \mathbf{a}_\phi(\mathbf{x}_t, t)\|^2, \quad (22)$$

and

$$\mathcal{L}_{CE} = \mathbb{E}_{\mathbf{x}_0 \sim p_{data}, \mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{x}_0), t \sim \mathcal{U}(0,1)} CE(L(\mathbf{x}_0), \mathbf{a}_\phi(\mathbf{x}_t, t)), \quad (23)$$

are the same and equal to $\mathbf{a}(\mathbf{x}_t, t)$, where CE represents the cross-entropy loss.

Based on the analysis above, we present the algorithms for training the diffusion model $\mathbf{y}_\theta(\mathbf{x}_t, t, l)$ in Algorithm 1 and the label model $\mathbf{a}_\phi(\mathbf{x}_t, t)$ in Algorithm 2.

Algorithm 1 Training of diffusion model \mathbf{y}_θ

- 1: **Repeat**
 - 2: $\mathbf{x}_0 \sim p_{data}$
 - 3: $t \sim \text{Uniform}(t_1, t_2, \dots, t_T)$
 - 4: $\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{x}_0)$
 - 5: Take gradient descent step on $\|\nabla_\theta \|\mathbf{y}_\theta(\mathbf{x}_t, t, l(\mathbf{x}_0)) - \mathbf{x}_0\|^2$
 - 6: **Until** converged
-

Algorithm 2 Training of label model \mathbf{a}_ϕ

- 1: **Repeat**
 - 2: $\mathbf{x}_0 \sim p_{data}$
 - 3: $t \sim \text{Uniform}(t_1, t_2, \dots, t_T)$
 - 4: $\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{x}_0)$
 - 5: Take gradient descent step on $\|\nabla_\phi \|\mathbf{a}_\phi(\mathbf{x}_t, t) - L(\mathbf{x}_0)\|^2$
 - 6: **Until** converged
-

The second approach, *fixed class approximation (FC)*, first samples a label l from Eq. (21) at $t = 1$. In this scenario, $\mathbf{a}(x_1, 1) = (b^1, b^2, \dots, b^L)^T$, where $b^l = \frac{N_l}{N}$ represents the proportion of samples in cluster l relative to the total number of samples in the dataset. Then the label adheres to this value along time t :

$$\mathbf{a}(\mathbf{x}_t, t) = \mathbf{e}^{(l)}. \quad (24)$$

Since the sampled label remains consistent over time t in the FC approximation, the class part $\mathbf{a}_\phi(\mathbf{x}_t, t)$ is necessitated solely at $t = 1$.

3.4 SAMPLING METHOD FOR DC-DPM

Conventionally, DDPM reverse approximation in Eq. (6) can be realized by the trajectory:

$$\mathbf{x}_{t_{i-1}} = \frac{\alpha_{t_i|t_{i-1}} \sigma_{t_{i-1}}^2}{\sigma_{t_i}^2} \mathbf{x}_{t_i} + \frac{\alpha_{t_{i-1}} \sigma_{t_i|t_{i-1}}^2}{\sigma_{t_i}^2} \mathbf{y}_\theta(\mathbf{x}_{t_i}, t_i) + \sigma_{t_{i-1}|t_i} \mathbf{z}_{t_i}, \quad (25)$$

where $\mathbf{z}_{t_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{x}_{t_T} \sim p(\mathbf{x}_{t_T}, t_T)$. Our reverse process, using a mixture of cluster-specific kernels, requires an additional random variable $\mathbf{y}(\mathbf{x}_{t_i}, t_i)$ to represent the trajectory:

$$\mathbf{x}_{t_{i-1}} = \frac{\alpha_{t_i|t_{i-1}} \sigma_{t_{i-1}}^2}{\sigma_{t_i}^2} \mathbf{x}_{t_i} + \frac{\alpha_{t_{i-1}} \sigma_{t_i|t_{i-1}}^2}{\sigma_{t_i}^2} \mathbf{y}(\mathbf{x}_{t_i}, t_i) + \sigma_{t_{i-1}|t_i} \mathbf{z}_{t_i}. \quad (26)$$

The density of $\mathbf{y}(\mathbf{x}_{t_i}, t_i)$ is

$$p(\mathbf{y}(\mathbf{x}_{t_i}, t_i) = \mathbf{y}) = \sum_l a^l(\mathbf{x}_{t_i}, t_i) \delta(\mathbf{y} - \mathbf{y}_\theta(\mathbf{x}_{t_i}, t_i, l)), \quad (27)$$

and $\mathbf{y}(\mathbf{x}_{t_i}, t_i)$ is independent of \mathbf{z}_{t_i} .

For label diffusion approximation (LD), our method samples two random variables in each step: weight sampling in line 3 and diffusion sampling in line 5 as shown in Algorithm 3. Fixed class approximation (FC) samples the weight from the discrete distribution $(b^1, b^2, \dots, b^L)^T$. Since the weight term remains consistent over time t , weight sampling is executed only once, as shown in line 2 of Algorithm 4. After weight sampling, the model generates samples within one fixed class l_0 . The generated distribution is:

$$\hat{p}^{l_0}(\mathbf{x}_{t_i}) = \int_{\mathcal{R}^d} \dots \int_{\mathcal{R}^d} \hat{p}^{l_0}(\mathbf{x}_{t_i} | \mathbf{x}_{t_{i+1}}) \dots \hat{p}^{l_0}(\mathbf{x}_{t_{T-1}} | \mathbf{x}_{t_T}) p^{l_0}(\mathbf{x}_{t_T}, t_T) d\mathbf{x}_{t_{i+1}} \dots d\mathbf{x}_{t_T}. \quad (28)$$

According to Proposition 4, $\hat{p}^{l_0}(\mathbf{x}_0)$ approximates $p_{data}^{l_0}(\mathbf{x}) = \frac{1}{N_{l_0}} \sum_i \delta(\mathbf{x} - \mathbf{y}_i^{l_0})$. Thus the distribution $\sum_l b^l \hat{p}^l(\mathbf{x}_0)$ approximate $p(\mathbf{x}_0) = \sum_l b^l p_{data}^l$, where $b^l = \frac{N_l}{N}$.

Algorithm 3 Sampling process of label diffusion approximation (LD)

```

1:  $\mathbf{x}_{t_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $i = T, \dots, 1$  do
3:    $l \sim \mathbf{a}_\theta(\mathbf{x}_{t_i}, t_i)$ 
4:    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:    $\mathbf{x}_{t_{i-1}} = \frac{\alpha_{t_i|t_{i-1}} \sigma_{t_{i-1}}^2}{\sigma_{t_i}^2} \mathbf{x}_{t_i}$ 
      $+ \frac{\alpha_{t_{i-1}} \sigma_{t_i|t_{i-1}}^2}{\sigma_{t_i}^2} \mathbf{y}_\theta(\mathbf{x}_{t_i}, t_i, l) + \sigma_{t_{i-1}|t_i} \mathbf{z}$ 
6: end for
7: return  $\mathbf{x}_{t_0}$ 

```

Algorithm 4 Sampling process of fixed class approximation (FC)

```

1:  $\mathbf{x}_{t_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2:  $l \sim (b^1, b^2, \dots, b^C)$ 
3: for  $i = T, \dots, 1$  do
4:    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:    $\mathbf{x}_{t_{i-1}} = \frac{\alpha_{t_i|t_{i-1}} \sigma_{t_{i-1}}^2}{\sigma_{t_i}^2} \mathbf{x}_{t_i}$ 
      $+ \frac{\alpha_{t_{i-1}} \sigma_{t_i|t_{i-1}}^2}{\sigma_{t_i}^2} \mathbf{y}_\theta(\mathbf{x}_{t_i}, t_i, l) + \sigma_{t_{i-1}|t_i} \mathbf{z}$ 
6: end for
7: return  $\mathbf{x}_{t_0}$ 

```

Algorithm 5 Sampling process for LD approximation with ODE-based methods

```

1:  $\mathbf{x}_{t_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $i = T, \dots, 1$  do
3:    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
4:    $\mathbf{x}_{t_{i-1}} = \text{ODE}(\mathbf{x}_{t_i}, t_i, l)$ 
5: end for
6: return  $\mathbf{x}_{t_0}$ 

```

Algorithm 6 Sampling process for FC approximation with ODE-based methods

```

1:  $\mathbf{x}_{t_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2:  $l \sim (b^1, b^2, \dots, b^L)$ 
3: for  $i = T, \dots, 1$  do
4:    $\mathbf{x}_{t_{i-1}} = \text{ODE}(\mathbf{x}_{t_i}, t_i, l)$ 
5: end for
6: return  $\mathbf{x}_{t_0}$ 

```

As the probability flow ODE keeps the single-time marginals (Song et al., 2020b), we can replace the diffusion sampling method with probability flow ODE-based methods, such as DDIM (Song et al., 2020a), DPM Solver (Lu et al., 2022a), PNDM (Liu et al., 2022) etc. We summarize this in Algorithm 5, where $\text{ODE}(\mathbf{x}_t, t, l)$ represents the ODE-based sampling methods. Similarly, the fixed class approximation is also applicable to ODE-based methods, as presented in Algorithm 6

4 EXPERIMENTS

4.1 IMAGE-SPACE RESULTS

Table 1: **FID ↓ on CIFAR-10 Dataset.** Employing our Divide-and-Conquer (DC) kernel approximation strategy on previous DPM methods enhances their generation quality especially on small timesteps. LD represents merging kernels with label diffusion approximation while FC represents fixed class approximation. SN-DDPM is short for Extended-Analytic-DPM (Bao et al., 2022a).

# TIMESTEPS	CIFAR-10 (Linear Schedule)						CIFAR-10 (Cosine Schedule)					
	10	25	50	100	200	1000	10	25	50	100	200	1000
DDPM	43.14	21.63	15.21	10.94	8.23	5.11	34.76	16.18	11.11	8.38	6.66	4.92
+DC-LD (Ours)	39.40	21.95	15.54	10.78	7.91	4.98	27.78	15.52	10.12	7.29	5.61	4.11
+DC-FC (Ours)	34.48	21.05	15.12	10.67	7.82	4.50	25.80	14.58	9.66	6.72	5.03	3.46
SN-DDPM	21.87	6.91	4.58	3.74	3.34	3.71	16.33	6.05	4.19	3.83	3.72	4.08
+DC-LD (Ours)	16.77	6.39	4.29	3.40	2.97	3.30	12.85	6.54	4.56	3.63	3.35	3.51
+DC-FC (Ours)	11.90	4.98	3.62	2.98	2.55	2.93	9.92	4.95	3.35	2.67	2.53	2.74
GMS	17.43	5.96	4.16	3.26	3.01	2.76	13.80	5.48	4.00	3.46	3.34	4.23
+DC-LD (Ours)	14.54	5.89	4.22	3.41	3.58	5.19	10.80	6.22	4.53	3.64	3.34	4.35
+DC-FC (Ours)	10.40	4.84	3.61	3.00	3.00	2.86	8.76	4.91	3.43	2.76	2.60	3.35

We quantitatively compare the sample quality using the widely recognized Fréchet Inception Distance (FID) score (Heusel et al., 2017). Utilizing the semantic labels from the CIFAR-10 dataset, we

432 categorize the data into 10 classes. We then apply our proposed divide-and-conquer approximation
 433 to various transition kernel designs, including DDPM (Ho et al., 2020), Extended-Analytic-DPM
 434 (Bao et al., 2022a), and GMS (Guo et al., 2024). These kernels are merged using both the label
 435 diffusion (LD) and fixed class (FC) approximation strategies. As illustrated in Table 1, our DC-
 436 DPM approach significantly enhances the performance of existing methods, particularly at smaller
 437 denoising timesteps. Specifically, DC-DPM achieves improvements of 25.78% for DDPM, 45.58%
 438 for Extended-Analytic-DPM, and 40.33% for GMS in scenarios with 10 denoising steps.

439 Furthermore, as shown in Table 2, DC-DPM can also be applied to the ODE-based sampler DDIM,
 440 resulting in a 22.38% generation quality improvement with 10 steps on the CIFAR-10 dataset.
 441

442 **4.2 LATENT-SPACE RESULTS**

443 We also apply DC-DPM to latent diffusion models (Rombach et al., 2022). We perform comparative
 444 experiments for unconditional generation on the CelebA-HQ-256 image dataset. To classify the data,
 445 we first compute the VAE latent space of each image Kingma & Welling (2013), extract the primary
 446 dimension using principal component analysis (PCA) Abdi & Williams (2010), and then cluster the
 447 images into 10 classes using the K-Means algorithm. Both the quantitative results in Table 3 and the
 448 qualitative results in Fig. 3 demonstrate that DC-DPM improves the generation quality of diffusion
 449 models in latent space.
 450

451 Table 2: **FID ↓ on CIFAR-10 (Linear Schedule) with DDIM.** DC-DPM can be ap-
 452 plied to ODE-based samplers like DDIM.
 453

# STEPS	10	25	50
DDIM	21.31	10.70	7.74
+DC-LD (Ours)	20.43	11.39	8.38
+DC-FC (Ours)	16.54	9.15	6.60

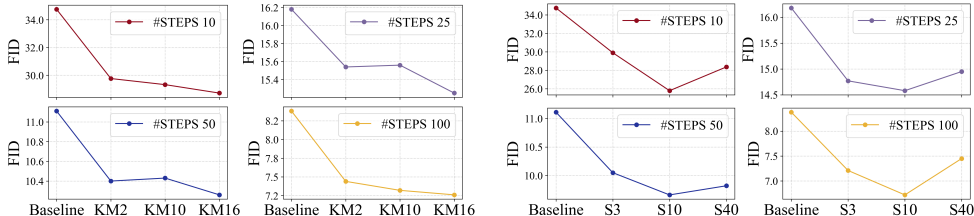
454 Table 3: **FID ↓ on CelebA-HQ-256.** DC-DPM is applicable to latent diffusion models to
 455 improve the generation quality.
 456

# STEPS	10	25	50
DDPM	35.21	18.60	14.16
+DC-LD (Ours)	30.58	15.76	12.25
+DC-FC (Ours)	30.47	15.37	12.16

457 **4.3 ABLATIONS**

458 We conduct experiments to examine the impact of different classification approaches and varying
 459 numbers of classes on the CIFAR-10 dataset. We flatten the input images and apply K-Means clus-
 460 tering to the raw image values, classifying the training data into different numbers of clusters. The
 461 FIDs for various denoising timesteps are presented in Fig. (2a). Our results indicate that the quality
 462 of generated images improves as the number of classes increases, although the rate of improvement
 463 diminishes with a higher number of classes. Notably, the generation quality for clusters created via
 464 K-Means remains inferior to that achieved using semantic labels, even when divided into 16 clusters.
 465

466 In Fig. (2b), we compare scenarios with different numbers of semantic labels. The label S3 denotes
 467 three semantic classes. Specifically, we consolidated the original 10 classes of the CIFAR-10 dataset
 468 into three broader categories: vehicles, animals, and others. For S40, we divide each of the original
 469 10 classes into four finer sub-classes using K-Means, resulting in a total of 40 classes. The gen-
 470 eration quality initially improves and then deteriorates as the number of classes increases. While
 471 having more classes makes each cluster-specific kernel easier to learn, it simultaneously raises the
 472 complexity of managing all these classes within a single conditional diffusion network $y_{\theta}(x_t, t, l)$.
 473



482 (a) FIDs Using K-Means with Varying Cluster Counts. (b) FIDs with Semantic Label Classification at
 483 Different Class Counts.
 484

485 **Figure 2: Ablations on Classification Approaches and Number of Classes on CIFAR-10.**



497 Figure 3: **Qualitative Results on CelebA-HQ-256 on 10 and 25 Denoising Steps.**

498 5 RELATED WORK

501 Significant research has focused on improving diffusion model performance on fewer timesteps,
502 broadly categorized into three approaches. Training-based methods includes trainable sampling
503 schedules (Watson et al., 2021), truncated diffusion (Lyu et al., 2022; Zheng et al., 2022), neural
504 operators (Zheng et al., 2023a), and distillation (Salimans & Ho, 2022; Sauer et al., 2023; Meng
505 et al., 2023; Song et al., 2023; Luo et al., 2023). The second category enhances the efficiency of
506 SDE and ODE solvers in the reverse process, including faster SDE and ODE solvers (Lu et al.,
507 2022a;b; Zheng et al., 2023b; Xu et al., 2023; Li et al., 2024), adaptive step size solvers (Jolicoeur-
508 Martineau et al., 2021), predictor-corrector methods (Song et al., 2020b; Zhao et al., 2023), and
509 stochastic-calculus-based optimization (Sabour et al., 2024).

510 The third category focuses on improving the design of the transition kernel in the diffusion reverse
511 process. Analytic-DPM (Bao et al., 2022b) and Extended-Analytic-DPM (Bao et al., 2022a) esti-
512 mate the optimal variance. Our work also falls within this category, with the most closely related
513 prior work being GMS (Guo et al., 2024). GMS represents the transition kernel as a mixture of
514 two Gaussians based on the estimation of higher-order moments. In contrast, the highlight of our
515 method is to divide data into clusters and construct the kernel function in a divide-and-conquer man-
516 ner. We construct a more general framework and previous Analytic-DPM, Extended-Analytic-DPM,
517 and GMS can serve as the cluster-specific kernel in our method.

518 6 CONCLUSION

519 In this paper, we propose DC-DPM, a novel divide-and-conquer approach for approximating the
520 transition kernel in the reverse process of diffusion probabilistic models. We provide convergence
521 proof for diffusion models from a new perspective, generalizing the transition kernel representation
522 from a conventional single Gaussian to a divide-and-conquer framework. This framework utilizes
523 cluster-specific kernels to represent segmented data, which are then merged to form an overall repre-
524 sentation. We propose two merging strategies along with their corresponding training and sampling
525 methods. Experimental results demonstrate the effectiveness of our approach, significantly enhanc-
526 ing generation quality, particularly over a limited number of timesteps.

527 REFERENCES

- 528
529
530
531 Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.
- 532
533 Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- 534
535 Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. *arXiv preprint arXiv:2206.07309*, 2022a.
- 536
537 Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022b.
- 538
539

- 540 Joao Carvalho, An T Le, Mark Baiert, Dorothea Koert, and Jan Peters. Motion planning diffusion:
541 Learning and planning of robot motions with diffusion models. In *2023 IEEE/RSJ International*
542 *Conference on Intelligent Robots and Systems (IROS)*, pp. 1916–1923. IEEE, 2023.
- 543 Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling:
544 User-friendly bounds under minimal smoothness assumptions. In *International Conference on*
545 *Machine Learning*, pp. 4735–4763. PMLR, 2023a.
- 546 Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and
547 distribution recovery of diffusion models on low-dimensional data. In *International Conference*
548 *on Machine Learning*, pp. 4672–4712. PMLR, 2023b.
- 549 Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy
550 as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint*
551 *arXiv:2209.11215*, 2022.
- 552 Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances*
553 *in neural information processing systems*, 34:8780–8794, 2021.
- 554 Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola.
555 A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.
- 556 Hanzhong Guo, Cheng Lu, Fan Bao, Tianyu Pang, Shuicheng Yan, Chao Du, and Chongxuan Li.
557 Gaussian mixture solvers for diffusion models. *Advances in Neural Information Processing Sys-*
558 *tems*, 36, 2024.
- 559 Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter.
560 Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in*
561 *neural information processing systems*, 30, 2017.
- 562 Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in*
563 *neural information processing systems*, 33:6840–6851, 2020.
- 564 Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P
565 Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition
566 video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.
- 567 Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J
568 Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–
569 8646, 2022b.
- 570 Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas.
571 Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*,
572 2021.
- 573 Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In
574 *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 21696–21707, 2021.
- 575 Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint*
576 *arXiv:1312.6114*, 2013.
- 577 Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with
578 polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882,
579 2022.
- 580 Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for
581 general data distributions. In *International Conference on Algorithmic Learning Theory*, pp.
582 946–985. PMLR, 2023.
- 583 Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating conver-
584 gence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*, 2024.
- 585 Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting
586 Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*,
587 479:47–59, 2022.

- 594 Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on
595 manifolds. *arXiv preprint arXiv:2202.09778*, 2022.
596
- 597 Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast
598 ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural*
599 *Information Processing Systems*, 35:5775–5787, 2022a.
- 600 Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast
601 solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*,
602 2022b.
603
- 604 Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthe-
605 sizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.
- 606 Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. Accelerating diffusion models
607 via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*, 2022.
608
- 609 Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and
610 Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF*
611 *Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.
- 612 Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.
613
- 614 Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d
615 diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- 616 Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-
617 tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine*
618 *Learning*, pp. 8599–8608. PMLR, 2021.
619
- 620 Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-
621 conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- 622 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-
623 resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF confer-*
624 *ence on computer vision and pattern recognition*, pp. 10684–10695, 2022.
625
- 626 Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling
627 schedules in diffusion models. *arXiv preprint arXiv:2404.14507*, 2024.
- 628 Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar
629 Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic
630 text-to-image diffusion models with deep language understanding. *Advances in neural informa-*
631 *tion processing systems*, 35:36479–36494, 2022.
- 632 Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv*
633 *preprint arXiv:2202.00512*, 2022.
634
- 635 Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion dis-
636 tillation. *arXiv preprint arXiv:2311.17042*, 2023.
- 637 Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv*
638 *preprint arXiv:2010.02502*, 2020a.
639
- 640 Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben
641 Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint*
642 *arXiv:2011.13456*, 2020b.
- 643 Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint*
644 *arXiv:2303.01469*, 2023.
645
- 646 Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. Learning fast samplers
647 for diffusion models by differentiating through sample quality. In *International Conference on*
Learning Representations, 2021.

648 Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart
649 sampling for improving generative processes. *Advances in Neural Information Processing Sys-*
650 *tems*, 36:76806–76838, 2023.

651 Pengze Zhang, Hubery Yin, Chen Li, and Xiaohua Xie. Tackling the singularities at the endpoints
652 of time intervals in diffusion models. *arXiv preprint arXiv:2403.08381*, 2024.

653 Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-
654 corrector framework for fast sampling of diffusion models. *Advances in Neural Information*
655 *Processing Systems*, 36, 2023.

656 Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Aizzadenesheli, and Anima Anandkumar. Fast
657 sampling of diffusion models via operator learning. In *International Conference on Machine*
658 *Learning*, pp. 42390–42402. PMLR, 2023a.

659 Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion proba-
660 bilistic models and diffusion-based adversarial auto-encoders. *arXiv preprint arXiv:2202.09671*,
661 2022.

662 Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode
663 solver with empirical model statistics. *Advances in Neural Information Processing Systems*, 36,
664 2023b.

665 A PROOFS

666 A.1 PROOF OF PROPOSITION 1

667 **Lemma 1** For all positive integrable functions $f(\mathbf{x}), g(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}_+$, we have

$$668 \int_{\mathbb{R}^d} f(\mathbf{x}) \, d\mathbf{x} \log \frac{\int_{\mathbb{R}^d} f(\mathbf{x}) \, d\mathbf{x}}{\int_{\mathbb{R}^d} g(\mathbf{x}) \, d\mathbf{x}} \leq \int_{\mathbb{R}^d} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} \, d\mathbf{x} \quad (29)$$

669 *proof.* Let $F = \int_{\mathbb{R}^d} f(\mathbf{x}) \, d\mathbf{x}$, $G = \int_{\mathbb{R}^d} g(\mathbf{x}) \, d\mathbf{x}$ and $h(t) = t \log t$. $h(t)$ is convex because

$$670 \frac{d^2}{dt^2} h(t) = \frac{1}{t} > 0. \quad (30)$$

671 And then

$$672 \int_{\mathbb{R}^d} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} \, d\mathbf{x} = \int_{\mathbb{R}^d} g(\mathbf{x}) h\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) \, d\mathbf{x} \\ 673 = G \int_{\mathbb{R}^d} \frac{g(\mathbf{x})}{G} h\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) \, d\mathbf{x}. \quad (31)$$

674 According to the probabilistic form of Jensen’s inequality

$$675 G \int_{\mathbb{R}^d} \frac{g(\mathbf{x})}{G} h\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) \, d\mathbf{x} \geq Gh\left(\int_{\mathbb{R}^d} \frac{g(\mathbf{x})}{G} \frac{f(\mathbf{x})}{g(\mathbf{x})} \, d\mathbf{x}\right) = Gh\left(\frac{1}{G} \int_{\mathbb{R}^d} f(\mathbf{x}) \, d\mathbf{x}\right) \\ 676 = Gh\left(\frac{F}{G}\right) = F \log \frac{F}{G} \quad (32) \\ 677 = \int_{\mathbb{R}^d} f(\mathbf{x}) \, d\mathbf{x} \log \frac{\int_{\mathbb{R}^d} f(\mathbf{x}) \, d\mathbf{x}}{\int_{\mathbb{R}^d} g(\mathbf{x}) \, d\mathbf{x}}.$$

678 Note that the integrability of f ensures the validity of Jensen’s inequality. \square

679 **Lemma 2** Let $\sigma > 0$, $0 < \beta < 1$ and $B = \sigma \sqrt{(d+2) \log \frac{1}{\sigma^4} + \sigma^2 M}$, for all $\mathbf{v} \in \mathbb{R}^d$ with
680 $|\mathbf{v}| \leq M$, Let $\delta = \min(e^{-\frac{1}{4}}, e^{\frac{1}{4} W_{-1}(-\frac{1}{d+2})})$ and $C = M(1 + \sqrt{2\pi})$, then $\sigma < \delta$ indicates

$$681 (2\pi\sigma^2)^{-\frac{d}{2}} \int_{|\mathbf{x}| > B} \exp\left(-\frac{\|\mathbf{x} + \sigma^2 \mathbf{v}\|^2}{2\sigma^2}\right) (-\mathbf{x} + \sigma^2 \mathbf{v})^T \mathbf{v} \, d\mathbf{x} \leq C\sigma^4 \quad (33)$$

702 *proof.*

$$\begin{aligned}
703 & \\
704 & \\
705 & (2\pi\sigma^2)^{-\frac{d}{2}} \int_{\|\mathbf{x}\|>B} \exp\left(-\frac{\|\mathbf{x} + \sigma^2\mathbf{v}\|^2}{2\sigma^2}\right) (-\mathbf{x} + \sigma^2\mathbf{v})^T \mathbf{v} \, d\mathbf{x} \\
706 & \\
707 & \stackrel{(1)}{=} (2\pi\sigma^2)^{-\frac{d}{2}} \int_{\|\mathbf{z}-\sigma^2\mathbf{v}\|>B} \exp\left(-\frac{\|\mathbf{z}\|^2}{2\sigma^2}\right) (-\mathbf{z}^T \mathbf{v}) \, d\mathbf{z} \\
708 & \\
709 & \stackrel{(2)}{\leq} (2\pi\sigma^2)^{-\frac{d}{2}} \int_{\|\mathbf{z}\|>B-\sigma^2M} \exp\left(-\frac{\|\mathbf{z}\|^2}{2\sigma^2}\right) (M\|\mathbf{z}\|) \, d\mathbf{z} \tag{34} \\
710 & \\
711 & = (2\pi\sigma^2)^{-\frac{d}{2}} M \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} A(d) \sin^{d-2}(\varphi_1) \sin^{d-3}(\varphi_2) \cdots \sin(\varphi_{d-2}) \, d\varphi_{d-1} \cdots d\varphi_2 \, d\varphi_1 \\
712 & \\
713 & \stackrel{(3)}{=} (2\pi)^{-\frac{d}{2}} M \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} A(d), \\
714 & \\
715 & \\
716 & \\
717 &
\end{aligned}$$

718 where

$$719 \quad A(d) = \sigma^{-d} \int_{r>B-\sigma^2M} \exp\left(-\frac{r^2}{2\sigma^2}\right) M r^d \, dr, \tag{35}$$

720 and the derivation of equation (1) is attributed to the change in the integral variable of $Z = \mathbf{x} +$
721 $\sigma^2\mathbf{Y}$. The inequality in equation (2) arises from a broader integral domain and a Cauchy inequality.
722 Equation (3) is derived from the calculation of the $d - 1$ dimensional sphere S^{d-1} .

$$\begin{aligned}
723 & \\
724 & \\
725 & \\
726 & \\
727 & A(d) = \sigma \int_{r>B-\sigma^2M} \exp\left(-\frac{r^2}{2\sigma^2}\right) M \left(\frac{r}{\sigma}\right)^d \, d\frac{r}{\sigma} \tag{36} \\
728 & \\
729 & = \sigma M \int_{r'>\frac{B-\sigma^2M}{\sigma}} \exp\left(-\frac{r'^2}{2}\right) r'^d \, dr' \\
730 & \\
731 & = \sigma M E(d). \\
732 & \\
733 &
\end{aligned}$$

734 Let $\delta = e^{-\frac{1}{4}}$ and then $\log \frac{1}{\sigma^4} > 1$.

$$\begin{aligned}
735 & E(d) = \int_{r>\frac{B-\sigma^2M}{\sigma}} \exp\left(-\frac{r^2}{2}\right) r^d \, dr \\
736 & \\
737 & = -r^{d-1} \exp\left(-\frac{r^2}{2}\right) \Big|_{\frac{B-\sigma^2M}{\sigma}}^\infty + \int_{r>\frac{B-\sigma^2M}{\sigma}} \exp\left(-\frac{r^2}{2}\right) (d-1) r^{d-2} \, dr \\
738 & \\
739 & = \left(\frac{B-\sigma^2M}{\sigma}\right)^{d-1} \exp\left(-\frac{(B-\sigma^2M)^2}{2\sigma^2}\right) + (d-1)E(d-2) \\
740 & \\
741 & = \left((d+2) \log \frac{1}{\sigma^4}\right)^{\frac{d-1}{2}} \exp\left(-\frac{(d+2) \log \frac{1}{\sigma^4}}{2}\right) + (d-1)E(d-2) \\
742 & \\
743 & = \left((d+2) \log \frac{1}{\sigma^4}\right)^{\frac{d-1}{2}} (\sigma^4)^{\frac{d+2}{2}} + (d-1)E(d-2) \\
744 & \\
745 & = \left((d+2) \log \frac{1}{\sigma^4}\right)^{\frac{d-1}{2}} (\sigma^4)^{\frac{d+2}{2}} + (d-1) \left((d+2) \log \frac{1}{\sigma^4}\right)^{\frac{d-3}{2}} (\sigma^4)^{\frac{d+2}{2}} + \cdots \tag{37} \\
746 & \\
747 & + \begin{cases} (d-1)(d-3) \cdots 4 \left((d+2) \log \frac{1}{\sigma^4}\right) (\sigma^4)^{\frac{d+2}{2}} + 2E(1) & d \text{ is odd,} \\ (d-1)(d-3) \cdots 3 \left((d+2) \log \frac{1}{\sigma^4}\right)^{\frac{1}{2}} (\sigma^4)^{\frac{d+2}{2}} + E(0) & d \text{ is even,} \end{cases} \\
748 & \\
749 & \leq \frac{d+2}{2} \left((d+2) \log \frac{1}{\sigma^4}\right)^{\frac{d-1}{2}} (\sigma^4)^{\frac{d+2}{2}} + \begin{cases} 2E(1) & d \text{ is odd,} \\ E(0) & d \text{ is even,} \end{cases} \\
750 & \\
751 & \leq \left((d+2) \log \frac{1}{\sigma^4}\right)^{\frac{d+1}{2}} \sigma^4 + \begin{cases} 2E(1) & d \text{ is odd,} \\ E(0) & d \text{ is even.} \end{cases} \\
752 & \\
753 & \\
754 & \\
755 &
\end{aligned}$$

Let $\delta = e^{\frac{1}{4}W_{-1}(-\frac{1}{d+2})}$, we have $(d+2)\log\frac{1}{\sigma^4}\sigma^4 \leq 1$, where W_{-1} is the branch of Lambert W function labelled by -1.

$$\begin{aligned}
E(1) &= \int_{r > \frac{B-\sigma^2 M}{\sigma}} \exp(-\frac{r^2}{2}) r \, dr \\
&= - \int_{r > \frac{B-\sigma^2 M}{\sigma}} d \exp(-\frac{r^2}{2}) \\
&= - \exp(-\frac{r^2}{2}) \Big|_{\frac{B-\sigma^2 M}{\sigma}}^{\infty} \\
&= \exp(-\frac{(B-\sigma^2 M)^2}{2\sigma^2}) \\
&= (\sigma^4)^{\frac{d+2}{2}},
\end{aligned} \tag{38}$$

Since all $\lambda > 0$, $\exp(\lambda r - \lambda \sqrt{(d+2)\log\frac{1}{\sigma^4}}) \geq 1$, we have

$$\begin{aligned}
E(0) &= \int_{r > \frac{B-\sigma^2 M}{\sigma}} \exp(-\frac{r^2}{2}) \, dr \\
&\leq \int_{\mathbb{R}} \exp(\lambda r - \lambda \sqrt{(d+2)\log\frac{1}{\sigma^4}}) \exp(-\frac{r^2}{2}) \, dr \\
&= \sqrt{2\pi} \exp(\frac{\lambda^2}{2} - \lambda \sqrt{(d+2)\log\frac{1}{\sigma^4}}) \\
&\leq \sqrt{2\pi} \exp(-\frac{(d+2)\log\frac{1}{\sigma^4}}{2}) \\
&= \sqrt{2\pi} (\sigma^4)^{\frac{d+2}{2}}.
\end{aligned} \tag{39}$$

As a result, by setting $\delta = \min(e^{-\frac{1}{4}}, e^{\frac{1}{4}W_{-1}(-\frac{1}{d+2})})$ and $C = M(1 + \sqrt{2\pi})$, we can achieve the required inequality.

Lemma 3 Given the notations from Lemma 2, if $\delta = \min(e^{-\frac{1}{4(d+2)}}, e^{\frac{1}{4}W_{-1}(-\frac{1}{d+2})})$ we have

$$\frac{B + \sigma^2 M}{\sigma} \exp(-\frac{|B + \sigma^2 M|^2}{2\sigma^2}) < \sigma^4. \tag{40}$$

proof. When $\delta = \min(e^{-\frac{1}{4(d+2)}}, e^{\frac{1}{4}W_{-1}(-\frac{1}{d+2})})$, we have $\frac{B-\sigma^2 M}{\sigma} \geq 1$ and $(d+2)\log\frac{1}{\sigma^4}\sigma^4 \leq 1$. Since the function $t \exp(-\frac{t^2}{2})$ is decreasing when $t \geq 1$, we have

$$\begin{aligned}
\frac{B + \sigma^2 M}{\sigma} \exp(-\frac{|B + \sigma^2 M|^2}{2\sigma^2}) &\leq \frac{B - \sigma^2 M}{\sigma} \exp(-\frac{|B - \sigma^2 M|^2}{2\sigma^2}) \\
&= \sqrt{(d+2)\log\frac{1}{\sigma^4}} (\sigma^4)^{\frac{d+2}{2}} \\
&= \sqrt{(d+2)\log\frac{1}{\sigma^4}\sigma^4} (\sigma^4)^{\frac{d-1}{2}} \sigma^4 \\
&\leq \sigma^4.
\end{aligned} \tag{41}$$

Lemma 4 Given the notations from Lemma 2, consider a set of vectors $\{\mathbf{v}_i \in \mathbb{R}^d \mid 1 \leq i \leq N\}$ such that $\max_i \|\mathbf{v}_i\| \leq M$, along with corresponding weights w_i for each vector \mathbf{v}_i satisfying $\sum_i w_i = 1$. For all $1 \leq i \leq N$, let $C = \frac{1}{2} \text{Tr}(\sum_j w_j \|\mathbf{v}_j\|^2) + M^3 + \frac{3}{2}M^4$ and $\delta = e^{\frac{1}{2\beta} W_{-1}(\frac{-\beta}{2(d+2)})}$ such that $\sigma < \delta$ indicates

$$(2\pi\sigma^2)^{-\frac{d}{2}} \underbrace{\int_{\|\mathbf{x}\| \leq B} \exp\left(-\frac{\|\mathbf{x} + \sigma^2 \mathbf{y}_i\|^2}{2\sigma^2}\right) \sum_j w_j \frac{1}{2} \mathbf{v}_j^T \mathbf{x} \mathbf{x}^T \mathbf{v}_j \, d\mathbf{x}}_I - \sum_j w_j \frac{1}{2} \sigma^2 \|\mathbf{v}_j\|^2 < C \sigma^{3-\beta}. \quad (42)$$

proof.

Let $\delta = e^{\frac{1}{2\beta} W_{-1}(\frac{-\beta}{2(d+2)})}$, we have $\sigma^{2\beta} \log \frac{1}{\sigma^4} \leq \frac{1}{d+2}$, which means $\sigma((d+2) \log \frac{1}{\sigma^4})^{-\frac{1}{2}} < \sigma^{1-\beta}$

Since the matrix $\mathbf{V} \stackrel{\text{def}}{=} \sum_j w_j \mathbf{v}_j \mathbf{v}_j^T$ is symmetric, it can be diagonalized: $\mathbf{V} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$ where $\mathbf{\Lambda}$ is an diagonal matrix, \mathbf{U} is an orthogonal matrix and

$$\text{Tr}(\mathbf{\Lambda}) = \text{Tr}\left(\sum_j w_j \mathbf{v}_j \mathbf{v}_j^T\right) = \sum_j w_j \text{Tr}(\mathbf{v}_j^T \mathbf{v}_j) = \sum_j w_j \|\mathbf{v}_j\|^2. \quad (43)$$

With the change of variable $\mathbf{z} = \mathbf{U}^T(\mathbf{x} + \sigma^2 \mathbf{v}_i)$, we have

$$\begin{aligned} \text{I} &= (2\pi\sigma^2)^{-\frac{d}{2}} \int_{\|\mathbf{x}\| \leq B} \exp\left(-\frac{\|\mathbf{x} + \sigma^2 \mathbf{y}_i\|^2}{2\sigma^2}\right) \sum_j w_j \frac{1}{2} \|(\mathbf{x} + \sigma^2 \mathbf{v}_i - \sigma^2 \mathbf{v}_i)^T \mathbf{v}_j\|^2 \, d\mathbf{x} \\ &= (2\pi\sigma^2)^{-\frac{d}{2}} \int_{\|\mathbf{U}\mathbf{z} - \sigma^2 \mathbf{v}_i\| \leq B} \exp\left(-\frac{\|\mathbf{z}\|^2}{2\sigma^2}\right) \sum_j w_j \frac{1}{2} (\|\mathbf{z}^T \mathbf{U}^T \mathbf{v}_j\|^2 - 2\sigma^2 \mathbf{z}^T \mathbf{U}^T \mathbf{v}_j \mathbf{v}_i^T \mathbf{v}_j \\ &\quad + \sigma^4 |\mathbf{v}_i^T \mathbf{v}_j|^2) \, d\mathbf{z} \\ &\leq (2\pi\sigma^2)^{-\frac{d}{2}} \int_{\|\mathbf{z}\| \leq B + \sigma^2 M} \exp\left(-\frac{\|\mathbf{z}\|^2}{2\sigma^2}\right) \frac{1}{2} \sum_j w_j \|\mathbf{z}^T \mathbf{U}^T \mathbf{v}_j\|^2 \, d\mathbf{z} + M^3 \sigma^2 (B + \sigma^2 M) + \frac{1}{2} M^4 \sigma^4 \\ &= (2\pi\sigma^2)^{-\frac{d}{2}} \int_{\|\mathbf{z}\| \leq B + \sigma^2 M} \exp\left(-\frac{\|\mathbf{z}\|^2}{2\sigma^2}\right) \frac{1}{2} \mathbf{z}^T \mathbf{\Lambda} \mathbf{z} \, d\mathbf{z} + M^3 \sigma^{3-\beta} + \frac{3}{2} M^4 \sigma^4 \\ &\leq (2\pi\sigma^2)^{-\frac{d}{2}} \frac{1}{2} \sum_l \int_{|z_l| \leq B + \sigma^2 M} \exp\left(-\frac{|z_l|^2}{2\sigma^2}\right) \mathbf{\Lambda}_{ll} z_l^2 \, dz_l \prod_{j=1, j \neq l}^d \int_{-\infty}^{\infty} \exp\left(-\frac{z_j^2}{2\sigma^2}\right) \, dz_j \\ &\quad + M^3 \sigma^{3-\beta} + \frac{3}{2} M^4 \sigma^4 \\ &= \text{Tr}(\mathbf{\Lambda}) (2\pi\sigma^2)^{-\frac{1}{2}} \frac{1}{2} \int_{|z| < B + \sigma^2 M} \exp\left(-\frac{z^2}{2\sigma^2}\right) z^2 \, dz + M^3 \sigma^{3-\beta} + \frac{3}{2} M^4 \sigma^4 \\ &= \text{Tr}(\mathbf{\Lambda}) \frac{\sigma^2}{\sqrt{2\pi}} \frac{1}{2} \left[2 \frac{B + \sigma^2 M}{\sigma} \exp\left(-\frac{|B + \sigma^2 M|^2}{2\sigma^2}\right) + \int_{|z| \leq B + \sigma^2 M} \exp\left(-\frac{z^2}{2\sigma^2}\right) \frac{z}{\sigma} \, dz \right] \\ &\quad + M^3 \sigma^{3-\beta} + \frac{3}{2} M^4 \sigma^4 \\ &\leq \text{Tr}(\mathbf{\Lambda}) \sigma^2 \frac{1}{2} (\sigma^4 + 1) + M^3 \sigma^{3-\beta} + \frac{3}{2} M^4 \sigma^4 \\ &= \sum_j \frac{1}{2} \sigma^2 w_j \|\mathbf{y}_j\|^2 + (\text{Tr}(\mathbf{\Lambda}) \frac{1}{2} \sigma (3 + \beta) + M^3 + \frac{3}{2} M^4 \sigma^{1+\beta}) \sigma^{3-\beta}. \end{aligned} \quad (44)$$

Let $C = \frac{1}{2} \text{Tr}(\mathbf{\Lambda}) + M^3 + \frac{3}{2} M^4$, we get the required equation (42). \square

Lemma 5 When $\|\mathbf{x}\| \leq B$ and $\sum_i w_i \mathbf{v}_i = 0$, let $\delta = \min(e^{\frac{1}{2\beta'}} W^{-1}(-\frac{\beta'}{2(d+2)}), (1+M)^{\beta'-1})$ and $C = \frac{1}{2}(1+M)M^3 + \frac{1}{8}M^4 + \frac{\epsilon}{6}(1+M)^3M^3 + \frac{1}{2}(1+M)^2M^4 + \frac{3}{2}(1+M)M^4 + \frac{1}{6}M^6$. If $\sigma \leq \delta$, we have

$$\begin{aligned} & \sum_j w_j \exp(-\mathbf{x}^T \mathbf{v}_j - \frac{1}{2}\sigma^2 \|\mathbf{v}_j\|^2) - 1 \\ & \leq \sum_j w_j (\frac{1}{2}\sigma^2 (\mathbf{v}_j^T \mathbf{x} \mathbf{x}^T \mathbf{v}_j - \|\mathbf{v}_j\|^2)) + C\sigma^{3-\beta} \end{aligned} \quad (45)$$

proof.

Let $\delta = \min(e^{\frac{1}{2\beta'}} W^{-1}(-\frac{\beta'}{2(d+2)}), (1+M)^{\beta'-1})$ and then $B = \sigma \sqrt{(d+2) \log \frac{1}{\sigma^4} + \sigma^2 M} < (1 + \sigma^{1+\beta'} M) \sigma^{1-\beta'} < (1+M) \sigma^{1-\beta'}$. Thus $\mathbf{x} \leq B \leq (1+M) \sigma^{1-\beta'} \leq 1$ and $|\mathbf{v}_j| \leq M$, where $\beta' = \frac{\beta}{3}$.

Expanding the function e^t at $t = 0$ with Lagrange's remainder, where $0 \leq \xi \leq B$

$$\begin{aligned} & \sum_j w_j \exp(-\mathbf{x}^T \mathbf{v}_j - \frac{1}{2}\sigma^2 \|\mathbf{v}_j\|^2) - 1 \\ & = \sum_j w_j \{ -\mathbf{x}^T \mathbf{v}_j - \frac{1}{2}\sigma^2 \|\mathbf{v}_j\|^2 + \frac{1}{2}(-\mathbf{x}^T \mathbf{v}_j - \frac{1}{2}\sigma^2 \|\mathbf{v}_j\|^2)^2 + \frac{1}{6}e^\xi (-\mathbf{x}^T \mathbf{v}_j - \frac{1}{2}\sigma^2 \|\mathbf{v}_j\|^2)^3 \} \\ & \leq \sum_j w_j \{ [-\frac{1}{2}\sigma^2 \|\mathbf{v}_j\|^2 + \frac{1}{2}\mathbf{v}_j^T \mathbf{x} \mathbf{x}^T \mathbf{v}_j] + \frac{1}{2}\sigma^2 \mathbf{x}^T \mathbf{v}_j \|\mathbf{v}_j\|^2 + \frac{1}{8}\sigma^4 \|\mathbf{v}_j\|^4 \\ & \quad + \frac{1}{6}e^B (\|\mathbf{x}^T \mathbf{v}_j\|^3 + 3\sigma^2 \|\mathbf{x}^T \mathbf{v}_j\|^2 \|\mathbf{v}_j\|^2 + 3\sigma^4 \|\mathbf{x}^T \mathbf{v}_j\| \|\mathbf{v}_j\|^4 + \sigma^6 \|\mathbf{v}_j\|^6) \} \\ & = \sum_j w_j (\frac{1}{2}\sigma^2 (\mathbf{v}_j^T \mathbf{x} \mathbf{x}^T \mathbf{v}_j - \|\mathbf{v}_j\|^2)) + \sum_j w_j (\frac{1}{2}(1+M)M^3 \sigma^{3-\beta'} + \frac{1}{8}M^4 \sigma^4 \\ & \quad + \frac{1}{6}e(1+M)^3 M^3 \sigma^{3-3\beta'} + \frac{1}{2}(1+M)^2 M^4 \sigma^{4-2\beta'} + \frac{1}{2}3(1+M)M^5 \sigma^{5-\beta'} + \frac{1}{6}M^6 \sigma^6) \\ & \leq \sum_j w_j (\frac{1}{2}\sigma^2 (\mathbf{v}_j^T \mathbf{x} \mathbf{x}^T \mathbf{v}_j - \|\mathbf{v}_j\|^2)) + C\sigma^{3-\beta}, \end{aligned} \quad (46)$$

where $C = \frac{1}{2}(1+M)M^3 + \frac{1}{8}M^4 + \frac{\epsilon}{6}(1+M)^3M^3 + \frac{1}{2}(1+M)^2M^4 + \frac{3}{2}(1+M)M^4 + \frac{1}{6}M^6$. \square

Lemma 6 For all $t_{min} \leq s < t \leq t_{max}$ and $0 < \beta < 1$, there exists a $\delta > 0$ and $C > 0$, depending on β, t_{min} and t_{max} , such that if $t - s < \delta$, the inequality $KL(p(\mathbf{x}_s) \|\tilde{p}(\mathbf{x}_s)) \leq KL(p(\mathbf{x}_t) \|\tilde{p}(\mathbf{x}_t)) + C(t-s)^{\frac{3-\beta}{2}}$ holds.

proof.

Noting that

$$\begin{aligned} KL(p(\mathbf{x}_s) \|\tilde{p}(\mathbf{x}_s)) & = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_t) p(\mathbf{x}_t) d\mathbf{x}_t \log \frac{\int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_t) p(\mathbf{x}_t) d\mathbf{x}_t}{\int_{\mathbb{R}^d} \tilde{p}(\mathbf{x}_s | \mathbf{x}_t) \tilde{p}(\mathbf{x}_t) d\mathbf{x}_t} d\mathbf{x}_s \\ & \leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_t) p(\mathbf{x}_t) \log \frac{p(\mathbf{x}_s | \mathbf{x}_t) p(\mathbf{x}_t)}{\tilde{p}(\mathbf{x}_s | \mathbf{x}_t) \tilde{p}(\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s \\ & = \underbrace{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_t) p(\mathbf{x}_t) \log \frac{p(\mathbf{x}_s | \mathbf{x}_t)}{\tilde{p}(\mathbf{x}_s | \mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s}_I \end{aligned} \quad (47)$$

$$+ \underbrace{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log \frac{p(\mathbf{x}_t)}{\tilde{p}(\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s}_{\text{II}}.$$

$$\text{II} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log p(\mathbf{x}_t) d\mathbf{x}_t d\mathbf{x}_s - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log \tilde{p}(\mathbf{x}_t) d\mathbf{x}_t d\mathbf{x}_s \quad (48)$$

Since $p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log p(\mathbf{x}_t) \geq 0$ and $p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log \tilde{p}(\mathbf{x}_t) \leq 0$ for all \mathbf{x}_t and \mathbf{x}_s , according to Fubini's theorem, we have

$$\begin{aligned} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log p(\mathbf{x}_t) d\mathbf{x}_t d\mathbf{x}_s &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log p(\mathbf{x}_t) d\mathbf{x}_s d\mathbf{x}_t \\ &= \int_{\mathbb{R}^d} p(\mathbf{x}_t) \log p(\mathbf{x}_t) d\mathbf{x}_t \end{aligned} \quad (49)$$

and

$$\begin{aligned} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log \tilde{p}(\mathbf{x}_t) d\mathbf{x}_t d\mathbf{x}_s &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log \tilde{p}(\mathbf{x}_t) d\mathbf{x}_s d\mathbf{x}_t \\ &= \int_{\mathbb{R}^d} p(\mathbf{x}_t) \log \tilde{p}(\mathbf{x}_t) d\mathbf{x}_t. \end{aligned} \quad (50)$$

Since the entropy of Gaussian mixtures and the cross entropy between Gaussians are all finite, we have $\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log p(\mathbf{x}_t) d\mathbf{x}_t d\mathbf{x}_s$ and $\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log \tilde{p}(\mathbf{x}_t) d\mathbf{x}_t d\mathbf{x}_s$ are both integrable. As a result,

$$\begin{aligned} \text{II} &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log p(\mathbf{x}_t) d\mathbf{x}_t d\mathbf{x}_s - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log \tilde{p}(\mathbf{x}_t) d\mathbf{x}_t d\mathbf{x}_s \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log p(\mathbf{x}_t) d\mathbf{x}_s d\mathbf{x}_t - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log \tilde{p}(\mathbf{x}_t) d\mathbf{x}_s d\mathbf{x}_t \quad (51) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t) \log \frac{p(\mathbf{x}_t)}{\tilde{p}(\mathbf{x}_t)} d\mathbf{x}_s d\mathbf{x}_t \\ &= \text{KL}(p(\mathbf{x}_t)||\tilde{p}(\mathbf{x}_t)). \end{aligned}$$

Now, let us delve into a detailed analysis of Part I.

$$\begin{aligned} \text{I} &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t) \log \frac{p(\mathbf{x}_s|\mathbf{x}_t)}{\tilde{p}(\mathbf{x}_s|\mathbf{x}_t)} d\mathbf{x}_s p(\mathbf{x}_t) d\mathbf{x}_t \\ &= \int_{\mathbb{R}^d} [(2\pi\sigma_s^2|t|)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \sum_i w_i(\mathbf{x}_t, t) \exp(-\frac{\|\mathbf{x}_s - \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{x}_t - \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\mathbf{y}_i\|^2}{2\sigma_s^2|t|}) \log(\sum_j w_j(\mathbf{x}_t, t) \exp(\\ &\quad - \underbrace{(\underbrace{(\mathbf{x}_s - \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{x}_t - \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\bar{\mathbf{y}}(\mathbf{x}_t, t))^T}_{\mathbf{x}} \underbrace{\frac{\alpha_s}{\sigma_s^2}(\bar{\mathbf{y}}(\mathbf{x}_t, t) - \mathbf{y}_j)}_{\mathbf{y}_j}} \\ &\quad + \frac{1}{2}\sigma_s^2|t| \frac{\alpha_s}{\sigma_s^2} \|\bar{\mathbf{y}}(\mathbf{x}_t, t) - \mathbf{y}_j\|^2))] d\mathbf{x}_s p(\mathbf{x}_t) d\mathbf{x}_t \\ &= \int_{\mathbb{R}^d} [(2\pi\sigma_s^2|t|)^{-\frac{d}{2}} \underbrace{\int_{A(\mathbf{x}_t, B, \bar{\mathbf{y}}(\mathbf{x}_t, t))}_{\text{III}}}_{\text{III}} + \underbrace{\int_{A^c(\mathbf{x}_t, B, \bar{\mathbf{y}}(\mathbf{x}_t, t))}_{\text{IV}}}_{\text{IV}}] \sum_i w_i(\mathbf{x}_t, t) \exp(\\ &\quad - \frac{\|\mathbf{x} + \sigma_s^2|t|\mathbf{y}_i\|^2}{2\sigma_s^2|t|}) \log \sum_j w_j(\mathbf{x}_t, t) \exp(-\mathbf{x}^T \mathbf{y}_i - \frac{1}{2}\sigma_s^2|t|\|\mathbf{y}_i\|^2)] d\mathbf{x}_s p(\mathbf{x}_t) d\mathbf{x}_t. \quad (52) \end{aligned}$$

According to Lemma 2, let $\delta_{\text{III}} = \min(e^{-\frac{1}{4}}, e^{\frac{1}{4}W^{-1}(-\frac{1}{4+2})})$ and $C_{\text{III}} = M_\sigma(1 + \sqrt{2\pi})$, where $M_\sigma = \frac{M}{\sigma_{t\min}^2}$, when $\sigma_s|t| \leq \delta_{\text{III}}$,

$$\text{III} = \int_{\mathbb{R}^d} (2\pi\sigma_s^2|t|)^{-\frac{d}{2}} \int_{A(\mathbf{x}_t, B, \bar{\mathbf{y}}(\mathbf{x}_t, t))} \sum_i w_i(\mathbf{x}_t, t) \exp(-\frac{\|\mathbf{x} + \sigma_s^2|t|\Delta\mathbf{y}_i\|^2}{2\sigma_s^2|t|}) \log(\sum_j w_j(\mathbf{x}_t, t) \exp(-\mathbf{x}^T \mathbf{y}_j - \frac{1}{2}\sigma_s^2|t|\|\mathbf{y}_j\|^2)) d\mathbf{x}_t$$

$$\begin{aligned}
& \frac{\sum_j w_j(\mathbf{x}_t, t) \exp(-\frac{\|\mathbf{x} + \sigma_{s|t}^2 \Delta \mathbf{y}_j\|^2}{2\sigma_{s|t}^2})}{\sum_j w_j(\mathbf{x}_t, t) \exp(-\frac{\|\mathbf{x}\|^2}{2\sigma_{s|t}^2})} d\mathbf{x}_s p(\mathbf{x}_t) d\mathbf{x}_t \\
& \leq \int_{\mathbb{R}^d} (2\pi\sigma_{s|t}^2)^{-\frac{d}{2}} \int_{A(\mathbf{x}_t, B, \bar{\mathbf{y}}(\mathbf{x}_t, t))} \sum_i w_i(\mathbf{x}_t, t) \exp(\\
& \quad - \frac{\|\mathbf{x} + \sigma_{s|t}^2 \Delta \mathbf{y}_i\|^2}{2\sigma_{s|t}^2}) \log\left(\frac{\exp(-\frac{\|\mathbf{x} + \sigma_{s|t}^2 \Delta \mathbf{y}_j\|^2}{2\sigma_{s|t}^2})}{\exp(-\frac{\|\mathbf{x}\|^2}{2\sigma_{s|t}^2})}\right) d\mathbf{x}_s p(\mathbf{x}_t) d\mathbf{x}_t \quad (53) \\
& = \int_{\mathbb{R}^d} (2\pi\sigma_{s|t}^2)^{-\frac{d}{2}} \int_{A(\mathbf{x}_t, B, \bar{\mathbf{y}}(\mathbf{x}_t, t))} \sum_i w_i(\mathbf{x}_t, t) \exp(\\
& \quad - \frac{\|\mathbf{x} + \sigma_{s|t}^2 \Delta \mathbf{y}_i\|^2}{2\sigma_{s|t}^2}) (-\mathbf{x}^T \Delta \mathbf{y}_i - \frac{1}{2} \sigma_{s|t}^2 \|\mathbf{y}_i\|^2) d\mathbf{x}_s p(\mathbf{x}_t) d\mathbf{x}_t \\
& \leq \int_{\mathbb{R}^d} C_{\text{III}} \sigma_{s|t}^4 p(\mathbf{x}_t) d\mathbf{x}_t = C_{\text{III}} \sigma_{s|t}^4.
\end{aligned}$$

According to Lemma 4 and Lemma 5, let $C_{\text{IV}} = \max(\frac{1}{2}M_\sigma + M_\sigma^3 + \frac{3}{2}M_\sigma^4, \frac{1}{2}(1 + M_\sigma)M_\sigma^3 + \frac{1}{8}M_\sigma^4 + \frac{e}{6}(1 + M_\sigma)^3 M_\sigma^3 + \frac{1}{2}(1 + M_\sigma)^2 M_\sigma^4 + \frac{3}{2}(1 + M_\sigma)M_\sigma^4 + \frac{1}{6}M_\sigma^6)$ and $\delta_{\text{IV}} = \min(e^{\frac{1}{3}\beta} W^{-1}(-\frac{\beta}{6(d+2)}), (1 + M_\sigma)^{\frac{1}{3}\beta-1})$, where $M_\sigma = \frac{M}{\sigma_{t_{\min}}^2}$, when $\sigma_{s|t} \leq \delta_{\text{IV}}$

$$\begin{aligned}
\text{IV} &= \sum_i \int_{\mathbb{R}^d} w_i(\mathbf{x}_t, t) (2\pi\sigma_{s|t}^2)^{-\frac{d}{2}} \int_{\|\mathbf{x}\| < B} \exp(-\frac{\|\mathbf{x} + \sigma_{s|t}^2 \mathbf{y}_i\|^2}{2\sigma_{s|t}^2}) \log[\\
& \quad \sum_j w_j(\mathbf{x}_t, t) \exp(-\mathbf{x}^T \mathbf{y}_j - \frac{1}{2} \sigma_{s|t}^2 \|\mathbf{y}_j\|^2)] d\mathbf{x} p(\mathbf{x}_t) d\mathbf{x}_t \\
&= \sum_i \int_{\mathbb{R}^d} w_i(\mathbf{x}_t, t) (2\pi\sigma_{s|t}^2)^{-\frac{d}{2}} \int_{\|\mathbf{x}\| < B} \exp(-\frac{\|\mathbf{x} + \sigma_{s|t}^2 \mathbf{y}_i\|^2}{2\sigma_{s|t}^2}) [\\
& \quad \sum_j w_j(\mathbf{x}_t, t) \exp(-\mathbf{x}^T \mathbf{y}_j - \frac{1}{2} \sigma_{s|t}^2 \|\mathbf{y}_j\|^2) - 1] d\mathbf{x}_s p(\mathbf{x}_t) d\mathbf{x}_t \\
&\leq \sum_i \int_{\mathbb{R}^d} w_i(\mathbf{x}_t, t) C_{\text{IV}} \sigma_{s|t}^{3-\beta} p(\mathbf{x}_t) d\mathbf{x}_t \\
&= C_{\text{IV}} \sigma_{s|t}^{3-\beta}. \quad (54)
\end{aligned}$$

Because $\sigma_{s|t}^2 = (1 - \frac{\alpha_s^2}{\alpha_t^2}) \frac{1 - \alpha_s^2}{1 - \alpha_t^2} \leq \frac{\alpha_s + \alpha_t}{\alpha_s^2} (\alpha_s - \alpha_t) \leq \frac{2C_\alpha}{\alpha_{t_{\min}}^2} (t - s)$. Let $\delta = \min(\delta_{\text{III}}, \delta_{\text{IV}})$ and $C = (C_{\text{III}} + C_{\text{IV}}) (\frac{2C_\alpha}{\alpha_{t_{\min}}^2})^{\frac{3}{2}}$, we get the result. \square

Proposition 1 For all $t_{\min} \leq s < t \leq t_{\max}$ and $0 < \beta < 1$, there exist $\delta > 0$ and $C_1, C_2 > 0$ depending on β, t_{\min} and t_{\max} , such that if $t - s < \delta$, the inequality $\text{KL}(p(\mathbf{x}_s) \| p_\theta(\mathbf{x}_s)) \leq \text{KL}(p(\mathbf{x}_t) \| p_\theta(\mathbf{x}_t)) + C_1 (t - s)^{\frac{3-\beta}{2}} + C_2 (t - s) \varepsilon_y$ holds.

proof.

Since $|\bar{\mathbf{y}}(\mathbf{x}_t, t)| = |\sum_i w_i(\mathbf{x}_t, t) \mathbf{y}_i| \leq \sum_i w_i(\mathbf{x}_t, t) |\mathbf{y}_i| \leq \sum_i w_i(\mathbf{x}_t, t) M = M$, we have

$$\int_{\mathbb{R}^d} p(\mathbf{x}_t) |\bar{\mathbf{y}}(\mathbf{x}_t, t)|^2 < M^2. \quad (55)$$

$$\text{KL}(p(\mathbf{x}_s) \| p_\theta(\mathbf{x}_s)) \leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_t) p(\mathbf{x}_t) \log \frac{p(\mathbf{x}_s | \mathbf{x}_t) p(\mathbf{x}_t)}{p_\theta(\mathbf{x}_s | \mathbf{x}_t) p_\theta(\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s$$

$$\begin{aligned}
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \log \frac{p(\mathbf{x}_s|\mathbf{x}_t)}{p_\theta(\mathbf{x}_s|\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s \\
&\quad + \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \log \frac{p(\mathbf{x}_t, t)}{p_\theta(\mathbf{x}_t, t)} d\mathbf{x}_t d\mathbf{x}_s. \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) [\log \frac{p(\mathbf{x}_s|\mathbf{x}_t)}{\tilde{p}(\mathbf{x}_s|\mathbf{x}_t)} + \log \frac{\tilde{p}(\mathbf{x}_s|\mathbf{x}_t)}{p_\theta(\mathbf{x}_s|\mathbf{x}_t)}] d\mathbf{x}_t d\mathbf{x}_s \quad (56) \\
&\quad + \text{KL}(p(\mathbf{x}_t)||p_\theta(\mathbf{x}_t)) \\
&\stackrel{(1)}{\leq} \underbrace{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \log \frac{\tilde{p}(\mathbf{x}_s|\mathbf{x}_t)}{p_\theta(\mathbf{x}_s|\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s}_I \\
&\quad + C_1(t-s)^{\frac{3-\beta}{2}} + \text{KL}(p(\mathbf{x}_t)||p_\theta(\mathbf{x}_t))
\end{aligned}$$

The inequality (1) use the conclusion in Proposition 6.

Since $\tilde{p}(\mathbf{x}_s|\mathbf{x}_t)$ and $p_\theta(\mathbf{x}_s|\mathbf{x}_t)$ are Gaussians with the same covariance matrix,

$$\begin{aligned}
I &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_t, t)p(\mathbf{x}_s|\mathbf{x}_t) \frac{1}{2\sigma_s^2|t} [(2\mathbf{x}_s - 2\frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{x}_t \\
&\quad - \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}(\bar{\mathbf{y}}(\mathbf{x}_t, t) + y_\theta(\mathbf{x}_t, t)))^T \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}(y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t))] d\mathbf{x}_s d\mathbf{x}_t \quad (57) \\
&= - \underbrace{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_t, t)p(\mathbf{x}_s|\mathbf{x}_t) \frac{1}{2\sigma_s^2|t} (\frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2})^2 (y_\theta(\mathbf{x}_t, t) + \bar{\mathbf{y}}(\mathbf{x}_t, t))^T (y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t)) d\mathbf{x}_s d\mathbf{x}_t}_{I_1} \\
&\quad + \underbrace{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_t, t)p(\mathbf{x}_s|\mathbf{x}_t) \frac{1}{\sigma_s^2|t} (\mathbf{x}_s - \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{x}_t)^T \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2} (y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t)) d\mathbf{x}_s d\mathbf{x}_t}_{I_2}
\end{aligned}$$

$$\begin{aligned}
I_1 &= - \int_{\mathbb{R}^d} p(\mathbf{x}_t) \frac{\alpha_s^2\sigma_{s|t}^2}{2\sigma_s^4} (y_\theta(\mathbf{x}_t, t) + \bar{\mathbf{y}}(\mathbf{x}_t, t))^T (y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t)) d\mathbf{x}_t \\
&\leq \frac{\alpha_s^2\sigma_{s|t}^2}{2\sigma_s^4} (\int_{\mathbb{R}^d} p(\mathbf{x}_t) \|(y_\theta(\mathbf{x}_t, t) + \bar{\mathbf{y}}(\mathbf{x}_t, t))\|^2 d\mathbf{x}_t \int_{\mathbb{R}^d} p(\mathbf{x}_t) \|(y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t))\|^2 d\mathbf{x}_t)^{\frac{1}{2}} \\
&\leq \frac{\alpha_s^2\sigma_{s|t}^2}{2\sigma_s^4} \varepsilon_y (\int_{\mathbb{R}^d} p(\mathbf{x}_t) \|(y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t) + 2\bar{\mathbf{y}}(\mathbf{x}_t, t))\|^2 d\mathbf{x}_t)^{\frac{1}{2}} \quad (58) \\
&\leq \frac{\alpha_s^2\sigma_{s|t}^2}{2\sigma_s^4} \varepsilon_y (\int_{\mathbb{R}^d} p(\mathbf{x}_t) 3(\|y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t)\|^2 + 4\|\bar{\mathbf{y}}(\mathbf{x}_t, t)\|^2) d\mathbf{x}_t)^{\frac{1}{2}} \\
&\leq 3 \frac{\alpha_s^2\sigma_{s|t}^2}{2\sigma_s^4} \varepsilon_y (\varepsilon_y^2 + 4M^2)^{\frac{1}{2}} \\
&\leq \frac{3}{2\sigma_{t_{min}}^4} (1 + 4M^2)^{\frac{1}{2}} \sigma_{s|t}^2 \varepsilon_y.
\end{aligned}$$

$$\begin{aligned}
I_2 &= \frac{\alpha_s}{\sigma_s^2} \int_{\mathbb{R}^d} p(\mathbf{x}_t) \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t) (\mathbf{x}_s - \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{x}_t)^T d\mathbf{x}_s (y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t)) d\mathbf{x}_t \\
&= \frac{\alpha_s}{\sigma_s^2} \int_{\mathbb{R}^d} p(\mathbf{x}_t) \sum_i w_i(\mathbf{x}_t, t) \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2} \mathbf{y}_i^T (y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t)) d\mathbf{x}_t \quad (59)
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{\alpha_s^2 \sigma_{s|t}^2}{\sigma_s^4} M \left(\int_{\mathbb{R}^d} p(\mathbf{x}_t) \|y_\theta(\mathbf{x}_t, t) - \bar{\mathbf{y}}(\mathbf{x}_t, t)\|^2 d\mathbf{x}_t \right)^{\frac{1}{2}} \\
&\leq \frac{\alpha_s^2}{\sigma_s^4} M \sigma_{s|t}^2 \varepsilon_y \leq \frac{M}{\sigma_{t_{\min}}^4} \sigma_{s|t}^2 \varepsilon_y
\end{aligned}$$

As as result, let $C_2 = \left(\frac{3}{2\sigma_{t_{\min}}^4} (1 + 4M^2)\right)^{\frac{1}{2}} + \frac{M}{\sigma_{t_{\min}}^4} \left(\frac{2C_\alpha}{\alpha_{t_{\min}}^2}\right)^{\frac{3}{2}}$ and δ use the value in Lemma 6, we get the required result. \square

A.2 PROOF OF PROPOSITION 2

Proposition 2 For all $0 < s < 1$, there are constants $C_1, C_2 > 0$, such that $KL(p(\mathbf{x}_s) || p_\theta(\mathbf{x}_s)) \leq C_1(1-s)^2 + C_2(1-s)^2 \varepsilon_y$.

proof.

First, we consider the difference between $p(\mathbf{x}_s)$ and $\tilde{p}(\mathbf{x}_s)$. Since $p(\mathbf{x}_1) = \tilde{p}(\mathbf{x}_1) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$,

$$\begin{aligned}
&KL(p(\mathbf{x}_s) || \tilde{p}(\mathbf{x}_s)) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_1) p(\mathbf{x}_1) d\mathbf{x}_1 \log \frac{\int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_1) p(\mathbf{x}_1) d\mathbf{x}_1}{\int_{\mathbb{R}^d} \tilde{p}(\mathbf{x}_s | \mathbf{x}_1) \tilde{p}(\mathbf{x}_1) d\mathbf{x}_1} d\mathbf{x}_s \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_1) p(\mathbf{x}_1) \log \frac{p(\mathbf{x}_s | \mathbf{x}_1) p(\mathbf{x}_1)}{\tilde{p}(\mathbf{x}_s | \mathbf{x}_1) \tilde{p}(\mathbf{x}_1)} d\mathbf{x}_1 d\mathbf{x}_s \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_1) p(\mathbf{x}_1) \log \frac{p(\mathbf{x}_s | \mathbf{x}_1)}{\tilde{p}(\mathbf{x}_s | \mathbf{x}_1)} d\mathbf{x}_1 d\mathbf{x}_s \\
&= \int_{\mathbb{R}^d} (2\pi\sigma_s^2)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \sum_i w_i(\mathbf{x}_1, 1) \exp\left(-\frac{\|\mathbf{x}_s - \alpha_s \mathbf{y}_i\|^2}{2\sigma_s^2}\right) \log\left[\frac{\sum_i w_i(\mathbf{x}_1, 1) \exp\left(-\frac{\|\mathbf{x}_s - \alpha_s \mathbf{y}_i\|^2}{2\sigma_s^2}\right)}{\exp\left(-\frac{\|\mathbf{x}_s - \alpha_s \bar{\mathbf{y}}(\mathbf{x}_1, 1)\|^2}{2\sigma_s^2}\right)} \right] d\mathbf{x}_s p(\mathbf{x}_1) d\mathbf{x}_1 \\
&\leq \int_{\mathbb{R}^d} (2\pi\sigma_s^2)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \sum_i w_i(\mathbf{x}_1, 1) \exp\left(-\frac{\|\mathbf{x}_s - \alpha_s \mathbf{y}_i\|^2}{2\sigma_s^2}\right) \log\left[\frac{\exp\left(-\frac{\|\mathbf{x}_s - \alpha_s \mathbf{y}_i\|^2}{2\sigma_s^2}\right)}{\exp\left(-\frac{\|\mathbf{x}_s - \alpha_s \bar{\mathbf{y}}(\mathbf{x}_1, 1)\|^2}{2\sigma_s^2}\right)} \right] d\mathbf{x}_s p(\mathbf{x}_1) d\mathbf{x}_1 \\
&\leq \int_{\mathbb{R}^d} (2\pi\sigma_s^2)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \sum_i w_i(\mathbf{x}_1, 1) \exp\left(-\frac{\|\mathbf{x}_s - \alpha_s \mathbf{y}_i\|^2}{2\sigma_s^2}\right) \\
&\quad \cdot \frac{\alpha_s}{\sigma_s^2} (\mathbf{x}_s - \alpha_s \mathbf{y}_i)^T (\mathbf{y}_i - \bar{\mathbf{y}}(\mathbf{x}_1, 1)) + \frac{\alpha_s^2}{2\sigma_s^2} \|\mathbf{y}_i - \bar{\mathbf{y}}(\mathbf{x}_1, 1)\|^2 d\mathbf{x}_s p(\mathbf{x}_1) d\mathbf{x}_1 \\
&\leq \int_{\mathbb{R}^d} \frac{\alpha_s^2}{2\sigma_s^2} M^2 p(\mathbf{x}_1) d\mathbf{x}_1 \leq \frac{M^2}{2\sigma_{t_{\min}}^2} \alpha_s^2 \leq \frac{C_\alpha^2 M^2}{2\sigma_{t_{\min}}^2} (1-s)^2.
\end{aligned} \tag{60}$$

Upon considering equations (56), (58), and (59), and designating $C_2 = C_\alpha^2 \left(\frac{3}{2\sigma_{t_{\min}}^4} (1 + 4M^2)\right)^{\frac{1}{2}} + \frac{M}{\sigma_{t_{\min}}^4}$, we are able to derive the desired conclusion. \square

The method used to prove the previous Proposition cannot be applied to prove Proposition 1 because there is a $\sigma_{s|t}^2$ in the denominator. This results in an error bound of $\sigma_{s|t}^2$, which does not allow for global convergence.

A.3 PROOF OF PROPOSITION 3

Proposition 3 Given $0 < t_{min} < 1$, the 2-Wasserstein distance

$$W_2(p(\mathbf{x}_0), p(\mathbf{x}_{t_{min}})) < \sqrt{2dC_\alpha t_{min}}. \quad (61)$$

proof.

$$\begin{aligned} W_2(p(\mathbf{x}_0), p(\mathbf{x}_{t_{min}})) &\leq \frac{1}{N} \sum_{i=1}^N W_2(\delta(x - \mathbf{y}_i), \mathcal{N}(\mathbf{y}_i, \sigma_{t_{min}}^2 \mathbf{I})) \\ &= \frac{1}{N} \sum_{i=1}^N \sqrt{d} \sigma_{t_{min}} \\ &\leq \frac{1}{N} \sum_{i=1}^N \sqrt{2dC_\alpha t_{min}} = \sqrt{2dC_\alpha t_{min}}. \end{aligned} \quad (62)$$

□

A.4 PROOF OF PROPOSITION 4

Proposition 4 For all $0 < \beta < 1$, there exist $\delta > 0$ and $C_1, C_2, C_3 > 0$, such that for all time discretizations \mathcal{D} with $|\mathcal{D}| < \delta$, the Kullback-Leibler divergence $KL(p(\mathbf{x}_{t_{min}}) || p_\theta(\mathbf{x}_{t_{min}})) \leq C_1 |\mathcal{D}|^{\frac{1-\beta}{2}} + C_2 \varepsilon_y$. Moreover, $W_2^2(p(\mathbf{x}_0), p(\mathbf{x}_{t_{min}})) < C_3 |\mathcal{D}|$.

proof.

According to Proposition 1 and 2, for all $i \in \{1, 2, \dots, T\}$

$$\begin{aligned} KL(p(x_{t_i}) || p_\theta(x_{t_i})) &\leq KL(p(x_{t_{i+1}}) || p_\theta(x_{t_{i+1}})) + C_1(t_{i+1} - t_i)^{\frac{3-\beta}{2}} + C_2(t_{i+1} - t_i)\varepsilon_y \\ &\leq KL(p(x_1) || p_\theta(x_1)) + C_1 |\mathcal{D}|^{\frac{1-\beta}{2}} + C_2 \varepsilon_y. \end{aligned} \quad (63)$$

The final estimation using the 2-Wasserstein is simply the Proposition 3. □

A.5 PROOF OF COROLLARY 1

Corollary 1 For all $t_{min} \leq s < t \leq t_{max}$ and $0 < \beta < 1$, there exists a $\delta > 0$ and $C_1, C_2, C_3 > 0$, depending on β, t_{min} and t_{max} , such that if $t - s < \delta$, the inequality $KL(p(\mathbf{x}_s) || \hat{p}_\theta(\mathbf{x}_s)) \leq KL(p(\mathbf{x}_t) || \hat{p}_\theta(\mathbf{x}_t)) + C_1(t - s)^{\frac{3-\beta}{2}} + C_2(t - s)\varepsilon_{yl} + C_3 \varepsilon_{al}$ holds.

proof. In accordance with the convexity of the Kullback-Leibler divergence, we have

$$\begin{aligned} KL(p(\mathbf{x}_s | \mathbf{x}_t) || \hat{p}(\mathbf{x}_s | \mathbf{x}_t)) &\leq \sum_l a^l(\mathbf{x}_t, t) KL(p^l(\mathbf{x}_s | \mathbf{x}_t), \hat{p}^l(\mathbf{x}_s | \mathbf{x}_t)) \\ &\stackrel{(1)}{=} \sum_l a^l(\mathbf{x}_t, t) C_1(t - s)^{\frac{3-\beta}{2}} \\ &= C_1(t - s)^{\frac{3-\beta}{2}}. \end{aligned} \quad (64)$$

The equality in step (1) is a direct consequence of Proposition 1.

$$KL(p(\mathbf{x}_s) || \hat{p}_\theta(\mathbf{x}_s)) \leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s | \mathbf{x}_t) p(\mathbf{x}_t) \log \frac{p(\mathbf{x}_s | \mathbf{x}_t) p(\mathbf{x}_t)}{\hat{p}_\theta(\mathbf{x}_s | \mathbf{x}_t) \hat{p}_\theta(\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s$$

$$\begin{aligned}
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \log \frac{p(\mathbf{x}_s|\mathbf{x}_t)}{\hat{p}_\theta(\mathbf{x}_s|\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s \\
&\quad + \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \log \frac{p(\mathbf{x}_t, t)}{\hat{p}_\theta(\mathbf{x}_t, t)} d\mathbf{x}_t d\mathbf{x}_s. \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) [\log \frac{p(\mathbf{x}_s|\mathbf{x}_t)}{\hat{p}_\theta(\mathbf{x}_s|\mathbf{x}_t)} + \log \frac{\hat{p}_\theta(\mathbf{x}_s|\mathbf{x}_t)}{\hat{p}_\theta(\mathbf{x}_s|\mathbf{x}_t)}] d\mathbf{x}_t d\mathbf{x}_s \quad (65) \\
&\quad + \text{KL}(p(\mathbf{x}_t) \|\hat{p}_\theta(\mathbf{x}_t)) \\
&\stackrel{(1)}{\leq} \underbrace{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \log \frac{\hat{p}_\theta(\mathbf{x}_s|\mathbf{x}_t)}{\hat{p}_\theta(\mathbf{x}_s|\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s}_I \\
&\quad + C_1(t-s)^{\frac{3-\beta}{2}} + \text{KL}(p(\mathbf{x}_t) \|\hat{p}_\theta(\mathbf{x}_t))
\end{aligned}$$

The inequality (1) results from equation (64).

$$\begin{aligned}
I &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \log \frac{\sum_{l=1}^L a^l(\mathbf{x}_t, t)\hat{p}^l(\mathbf{x}_s|\mathbf{x}_t)}{\sum_{l=1}^L a_\phi^l(\mathbf{x}_t, t)\hat{p}_\theta^l(\mathbf{x}_s|\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \frac{1}{\sum_{k=1}^L a^k(\mathbf{x}_t, t)\hat{p}^k(\mathbf{x}_s|\mathbf{x}_t)} \\
&\quad \cdot \sum_{l=1}^L a^l(\mathbf{x}_t, t)\hat{p}^l(\mathbf{x}_s|\mathbf{x}_t) \log \frac{a^l(\mathbf{x}_t, t)\hat{p}^l(\mathbf{x}_s|\mathbf{x}_t)}{a_\phi^l(\mathbf{x}_t, t)\hat{p}_\theta^l(\mathbf{x}_s|\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \sum_l \log \frac{\hat{p}^l(\mathbf{x}_s|\mathbf{x}_t)}{\hat{p}_\theta^l(\mathbf{x}_s|\mathbf{x}_t)} d\mathbf{x}_t d\mathbf{x}_s \\
&\quad + \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_t)p(\mathbf{x}_t, t) \sum_l \log \frac{a^l(\mathbf{x}_t, t)}{a_\phi^l(\mathbf{x}_t, t)} d\mathbf{x}_t d\mathbf{x}_s \quad (66) \\
&\stackrel{(1)}{\leq} C_2\sigma_{s|t}^2\varepsilon_{yl} + \int_{\mathbb{R}^d} p(\mathbf{x}_t, t) \sum_l \log \frac{a^l(\mathbf{x}_t, t)}{a_\phi^l(\mathbf{x}_t, t)} d\mathbf{x}_t \\
&\leq C_2\sigma_{s|t}^2\varepsilon_{yl} + \int_{\mathbb{R}^d} p(\mathbf{x}_t, t) \sum_l \frac{1}{C_a} |a^l(\mathbf{x}_t, t) - a_\phi^l(\mathbf{x}_t, t)| d\mathbf{x}_t \\
&\leq C_2\sigma_{s|t}^2\varepsilon_{yl} + \left(\int_{\mathbb{R}^d} p(\mathbf{x}_t, t) \sum_l \frac{1}{C_a} |a^l(\mathbf{x}_t, t) - a_\phi^l(\mathbf{x}_t, t)|^2 d\mathbf{x}_t \right)^{\frac{1}{2}} \\
&= C_2\sigma_{s|t}^2\varepsilon_{yl} + C_3\varepsilon_{al}.
\end{aligned}$$

Given the established relationship between $\sigma_{s|t}$ and $t-s$, we are able to derive the necessary conclusion.

A.6 PROOF OF COROLLARY 2

Corollary 2 For all $0 < s < 1$, there are constants $C_1, C_2, C_3 > 0$, such that $\text{KL}(p(\mathbf{x}_s) \|\hat{p}_\theta(\mathbf{x}_s)) \leq C_1(1-s)^2 + C_2\varepsilon_{yl} + C_3\varepsilon_{al}$.

proof.

$$\begin{aligned}
&\text{KL}(p(\mathbf{x}_s) \|\hat{p}^\theta(\mathbf{x}_s)) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_1)p(\mathbf{x}_1) d\mathbf{x}_1 \log \frac{\int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_1)p(\mathbf{x}_1) d\mathbf{x}_1}{\int_{\mathbb{R}^d} \hat{p}^\theta(\mathbf{x}_s|\mathbf{x}_1)\hat{p}^\theta(\mathbf{x}_1) d\mathbf{x}_1} d\mathbf{x}_s
\end{aligned}$$

$$\begin{aligned}
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_1)p(\mathbf{x}_1) \log \frac{p(\mathbf{x}_s|\mathbf{x}_1)p(\mathbf{x}_1)}{\hat{p}^\theta(\mathbf{x}_s|\mathbf{x}_1)\hat{p}^\theta(\mathbf{x}_1)} d\mathbf{x}_1 d\mathbf{x}_s \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_1)p(\mathbf{x}_1) \log \frac{p(\mathbf{x}_s|\mathbf{x}_1)}{\hat{p}^\theta(\mathbf{x}_s|\mathbf{x}_1)} d\mathbf{x}_1 d\mathbf{x}_s \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_1)p(\mathbf{x}_1) \log \frac{p(\mathbf{x}_s|\mathbf{x}_1)}{\hat{p}(\mathbf{x}_s|\mathbf{x}_1)} d\mathbf{x}_1 d\mathbf{x}_s \\
&\quad + \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_1)p(\mathbf{x}_1) \log \frac{\hat{p}(\mathbf{x}_s|\mathbf{x}_1)}{\hat{p}^\theta(\mathbf{x}_s|\mathbf{x}_1)} d\mathbf{x}_1 d\mathbf{x}_s \tag{67} \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \sum_l a^l(\mathbf{x}_t, t) p^l(\mathbf{x}_s|\mathbf{x}_1)p(\mathbf{x}_1) \log \frac{p^l(\mathbf{x}_s|\mathbf{x}_1)}{\hat{p}^l(\mathbf{x}_s|\mathbf{x}_1)} d\mathbf{x}_1 d\mathbf{x}_s \\
&\quad + \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(\mathbf{x}_s|\mathbf{x}_1)p(\mathbf{x}_1) \frac{1}{\sum_k a^k(\mathbf{x}_t, 1)\hat{p}^k(\mathbf{x}_s|\mathbf{x}_1)} \\
&\quad \sum_l a^l(\mathbf{x}_s|\mathbf{x}_1)\hat{p}^l(\mathbf{x}_s|\mathbf{x}_1) \log \frac{a^l(\mathbf{x}_1, 1)\hat{p}^l(\mathbf{x}_s|\mathbf{x}_1)}{a_\phi^l(\mathbf{x}_1, 1)\hat{p}^\theta(\mathbf{x}_s|\mathbf{x}_1)} d\mathbf{x}_1 d\mathbf{x}_s \\
&\stackrel{(1)}{\leq} C_1(1-s)^2 + C_2\varepsilon_{yl} + C_3\varepsilon_{al}
\end{aligned}$$

The inequality (1) utilizes Proposition 2 in conjunction with the method employed in the proof of Corollary 1. \square

A.7 PROOF OF COROLLARY 3

Corollary 3 For all $0 < \beta < 1$, there is a $\delta > 0$ and $C_1, C_2, C_3, C_4 > 0$, such that for all time discretizations \mathcal{D} with $|\mathcal{D}| < \delta$, the Kullback-Leibler divergence $KL(p(\mathbf{x}_{t_{min}}|\hat{p}_\theta(\mathbf{x}_{t_{min}})) \leq C_1|\mathcal{D}|^{\frac{1-\beta}{2}} + C_2\varepsilon_{yl} + C_3T\varepsilon_{al}$. Moreover, $W_2(p(\mathbf{x}_0), p(\mathbf{x}_{t_{min}})) < C_4|\mathcal{D}|$.

proof.

The methodology employed to prove this corollary mirrors that used in the proof of Proposition 4. The sole divergence lies in the inclusion of an additional term with ε_{al} . \square

A.8 PROOF OF PROPOSITION 5

Proposition 5 Let $L(\mathbf{x}_0)$ denote the one-hot class vector of \mathbf{x}_0 , the optimal $\mathbf{a}_\phi(\mathbf{x}_t, t)$ for the two objective functions

$$\mathcal{L}_2 = \mathbb{E}_{\mathbf{x}_0 \sim p_{data}, \mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_0), t \sim \mathcal{U}(0,1)} \|L(\mathbf{x}_0) - \mathbf{a}_\phi(\mathbf{x}_t, t)\|^2, \tag{68}$$

and

$$\mathcal{L}_{CE} = \mathbb{E}_{\mathbf{x}_0 \sim p_{data}, \mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_0), t \sim \mathcal{U}(0,1)} CE(L(\mathbf{x}_0), \mathbf{a}_\phi(\mathbf{x}_t, t)), \tag{69}$$

are the same and equal to $\mathbf{a}(\mathbf{x}_t, t)$, where CE represents the cross-entropy loss.

proof.

Given that the subscript is utilized for data indices, we opt to use superscripts for vector components within the context of this proof. Let L_i denotes the one-hot class vector of the data \mathbf{y}_i .

(1) Loss \mathcal{L}_2 . It is a constrained optimization problem:

$$\begin{cases} \arg \min & \mathcal{L}_2, \\ & \mathbf{a}_\phi \\ \text{s.t.} & \mathbb{1}^T \mathbf{a}_\phi = 1, a_\theta^l \geq 0, \end{cases} \tag{70}$$

where $\mathbb{1}$ is a column vector, all of whose elements are 1s. Using the KKT condition Nocedal & Wright (1999)

$$0 = \nabla_{\mathbf{a}_\phi(\mathbf{x}_t, t)} \mathcal{L}_2 + \nu(\mathbb{1}^T \mathbf{a}_\phi(\mathbf{x}_t, t) - 1) - \mu^T \mathbf{a}_\phi(\mathbf{x}_t, t)$$

$$\begin{aligned}
1296 &= \nabla_{\mathbf{a}_\phi(\mathbf{x}_t, t)} \sum_i \underbrace{\frac{1}{N} (2\pi\sigma_t^2)^{-\frac{d}{2}} v_i(\mathbf{x}_t, t)}_{A_t} |L_i - \mathbf{a}_\phi(\mathbf{x}_t, t)|^2 + \nu(\mathbb{1}^T \mathbf{a}_\phi(\mathbf{x}_t, t) - 1) - \mu^T \mathbf{a}_\phi(\mathbf{x}_t, t) \\
1297 &= \sum_i A_t v_i(\mathbf{x}_t, t) (\mathbf{a}_\phi(\mathbf{x}_t, t) - L_i) + \nu \mathbb{1} - \mu \\
1298 &= A_t \sum_i v_i(\mathbf{x}_t, t) \mathbf{a}_\phi(\mathbf{x}_t, t) - A_t \sum_i v_i(\mathbf{x}_t, t) L_i + \nu \mathbb{1} - \mu,
\end{aligned} \tag{71}$$

which leads to

$$\mathbf{a}_\phi^*(\mathbf{x}_t, t) = \frac{\sum_i v_i(\mathbf{x}_t, t) L_i - \nu \mathbb{1} / A_t + \mu / A_t}{\sum_j v_j(\mathbf{x}_t, t)}, \tag{72}$$

where $\mu^l \geq 0, \forall l$. Because $\mathbb{1}^T \mathbf{a}_\phi^*(\mathbf{x}_t, t) = 1$, we have

$$\mathbb{1}^T \mathbf{a}_\phi^*(\mathbf{x}_t, t) = \frac{\sum_i v_i(\mathbf{x}_t, t) \mathbb{1} L_i - \nu L / A_t + \mathbb{1}^T \mu / A_t}{\sum_j v_j(\mathbf{x}_t, t)} = 1 - \nu L / A_t + \mathbb{1}^T \mu / A_t, \tag{73}$$

which indicates $\nu L = \mathbb{1}^T \mu \geq 0$. Since $(\mathbf{a}_\phi^*(\mathbf{x}_t, t))^l \mu^l = 0, \forall l$, we have

$$\left(\frac{\sum_i v_i(\mathbf{x}_t, t) L_i - \mathbb{1} \nu / A_t + \mu / A_t}{\sum_j v_j(\mathbf{x}_t, t)} \right)^l \mu^l = 0. \tag{74}$$

If

$$\sum_i v_i(\mathbf{x}_t, t) L_i^l - \nu / A_t + \mu^l / A_t = 0, \tag{75}$$

then

$$\nu > \mu^l \geq 0, \tag{76}$$

which lead to the contradiction

$$\nu L > \mathbb{1}^T \mu. \tag{77}$$

As a result, $\mu = 0$ and $\nu = 0$, and

$$\mathbf{a}_\phi^*(\mathbf{x}_t, t) = \frac{\sum_i v_i(\mathbf{x}_t, t) L_i - \mathbb{1} \nu / A_t + \mu / A_t}{\sum_j v_j(\mathbf{x}_t, t)} = \sum_i w_i(\mathbf{x}_t, t) L_i = \mathbf{a}(\mathbf{x}_t, t). \tag{78}$$

The Lagrange multiplier ν and μ are zero, which means we can omit the constrain $\mathbb{1}^T \mathbf{a}_\phi(\mathbf{x}_t, t) = 1$ and $a_\phi^l \geq 0$. As the object function and the feasible set are all convex, the KKT condition is also sufficient.

(2) The loss \mathcal{L}_{CE} . Using the KKT condition Nocedal & Wright (1999)

$$\begin{aligned}
1331 &0 = \nabla_{\mathbf{a}_\phi(\mathbf{x}_t, t)} \mathcal{L}_{CE} + \nu(\mathbb{1}^T \mathbf{a}_\phi(\mathbf{x}_t, t) - 1) - \mu^T \mathbf{a}_\phi(\mathbf{x}_t, t) \\
1332 &= \nabla_{\mathbf{a}_\phi(\mathbf{x}_t, t)} \sum_i \underbrace{\frac{1}{N} (2\pi\sigma_t^2)^{-\frac{d}{2}} v_i(\mathbf{x}_t, t)}_{A_t} \sum_l -L_i^l \log(a_\theta^l(\mathbf{x}_t, t)) + \nu(\mathbb{1}^T \mathbf{a}_\phi(\mathbf{x}_t, t) - 1) - \mu \\
1333 &= - \sum_i A_t v_i(\mathbf{x}_t, t) \begin{bmatrix} L_i^1 / a_\theta^1(\mathbf{x}_t, t) \\ \vdots \\ L_i^L / a_\theta^L(\mathbf{x}_t, t) \end{bmatrix} + \nu \mathbb{1} - \mu,
\end{aligned} \tag{79}$$

which leads to

$$(\mathbf{a}_\phi^*(\mathbf{x}_t, t))^l = \sum_i \frac{A_t v_i(\mathbf{x}_t, t)}{\nu - \mu^l} L_i^l, \tag{80}$$

where $\mu^l \geq 0$. Since $(\mathbf{a}_\phi^*(\mathbf{x}_t, t))^l \mu^l = 0$, we must have $\mu^l = 0, \forall l$.

Because $\mathbb{1}^T \mathbf{a}_\phi^*(\mathbf{x}_t, t) = 1$, we have $\nu = A_t \sum_i v_i(\mathbf{x}_t, t)$. Thus

$$\mathbf{a}_\phi^*(\mathbf{x}_t, t) = \sum_i \frac{A_t v_i(\mathbf{x}_t, t)}{A_t \sum_j v_j(\mathbf{x}_t, t)} L_i = \sum_i w_i(\mathbf{x}_t, t) L_i = \mathbf{a}(\mathbf{x}_t, t). \tag{81}$$

In this case, the Lagrange multiplier ν is not zero, thus the constrain $\mathbb{1}^T \mathbf{a}_\phi(\mathbf{x}_t, t) = 1$ is essential. As the object function and the feasible set are all convex, the KKT condition is also sufficient. \square

#STEPS	Patterns (Number of Classes)														
	Circles (2)			Moons (2)			Pinwheel (5)			CheckerBoard (8)			Gaussians (8)		
	1 GS	FC	LD	1 GS	FC	LD	1 GS	FC	LD	1 GS	FC	LD	1 GS	FC	LD
500	2.788	8.017	0.8717	3.935	7.051	2.590	3.672	6.658	3.056	-1.301	6.561	-2.912	6.339	7.945	2.567
100	7.685	5.066	2.088	5.724	4.111	0.7585	3.891	9.002	3.789	0.1639	7.913	-4.804	6.364	11.80	5.768
50	12.42	4.764	2.594	13.21	5.457	0.08196	14.05	14.79	3.265	9.516	8.759	0.8225	18.93	13.66	7.536
30	20.82	10.86	15.55	21.38	8.021	3.508	24.37	17.64	8.619	33.57	12.18	7.559	45.70	8.005	1.572
20	50.96	18.90	40.57	40.36	10.44	10.12	52.51	20.48	9.817	115.5	19.88	10.22	90.04	6.098	3.801
10	121.0	33.44	71.94	149.1	43.29	44.30	169.3	21.07	12.32	494.9	54.94	90.71	211.1	22.03	16.62

Table 4: **Comparison on Synthetic Datasets.** 1 GS indicates the baseline which approximates each step as a single Gaussian. FC and LD represent our methods. Generation quality is assessed by Maximum Mean Discrepancy (MMD) \downarrow . Values in the table have been rescaled by a factor of 10^{-5} .

B EXPERIMENTAL DETAILS

B.1 TRAINING DETAILS

We use aligned training setting to that of the noise prediction network in Extended-Analytic-DPM (Bao et al., 2022a) and GMS (Guo et al., 2024) for image-space experiments on CIFAR-10. We use an exponential moving average (EMA) with a rate of 0.9999 and set the batch size as 128, learning rate as $2e-4$. We train 600K iterations and save a checkpoint every 10K iterations. For the latent-space experiments on CelebA-HQ-256, we align the setting with LDM (Rombach et al., 2022) and set the batch size as 48, learning rate as $9.6e-5$. We train 500K iterations and choose the best checkpoint to evaluate. We use the same training setting for the label model in label diffusion merging approximation. Training on CIFAR-10 and CelebA-HQ-256 both take about 48 hours on 8 Tesla V100 GPUs.

To apply our method to Extended-Analytic-DPM and GMS, two higher-order noise prediction networks need to be trained. We align the settings with Extended-Analytic-DPM and GMS and train two additional light-weight prediction heads with the backbone model frozen. Please refer to these two original papers for more details.

B.2 EVALUATION DETAILS

Following Extended-Analytic-DPM and GMS, we calculate the FID score on 50K generated samples, using the official implementation of FID for pytorch (<https://github.com/mseitzer/pytorch-fid>). The reference distribution statistics of FID are computed on the full training set. The parameters in sampling are kept aligned with those in Extended-Analytic-DPM, please refer to Appendix F.5 in the original paper of Extended-Analytic-DPM (Bao et al., 2022a) for more details.

B.3 RESULTS ON 2D SYNTHETIC DATASET

We validate our approach on five synthetic 2D datasets with varying distributions. Each dataset consists of continuous 2D points $(x, y) \in \mathbb{R}^2$, assigned class labels based on natural clustering. For each experiment, we generated 4K samples and assessed generation quality using Maximum Mean Discrepancy (MMD) with a Laplace kernel (bandwidth 0.1) (Gretton et al., 2012). Each computation was repeated 8 times, and we report the average MMD value, with lower values indicating better generation quality. As shown in Table 4, our LD and FC methods outperform the single Gaussian baseline, achieving lower MMD across different timesteps.

C QUALITATIVE RESULTS

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419



Figure 4: DDPM + DC-DPM on 10 Denoising Steps.

1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437

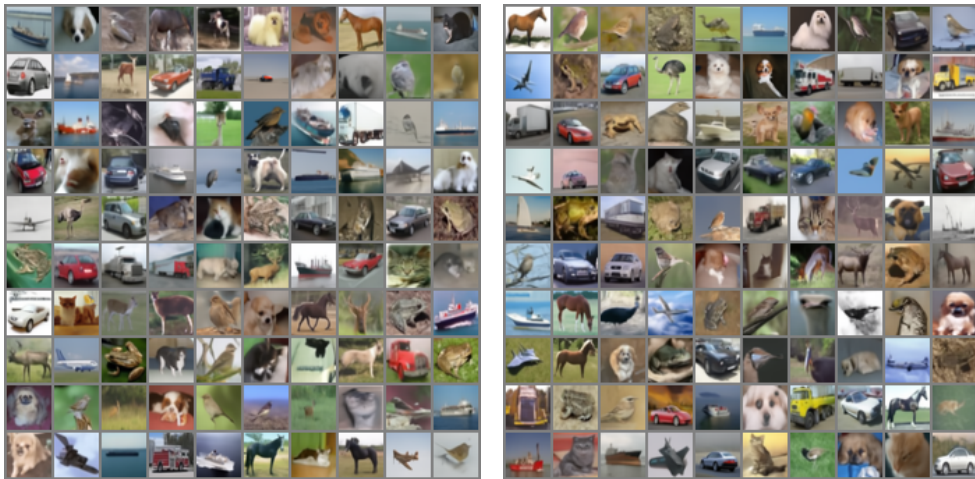


Figure 5: DDPM + DC-DPM on 25 Denoising Steps.

1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457



Figure 6: DDPM + DC-DPM on 50 Denoising Steps.

1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473



Figure 7: DDPM + DC-DPM on 100 Denoising Steps.

1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491



Figure 8: SN-DDPM (Bao et al., 2022a) + DC-DPM on 10 Denoising Steps.

1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

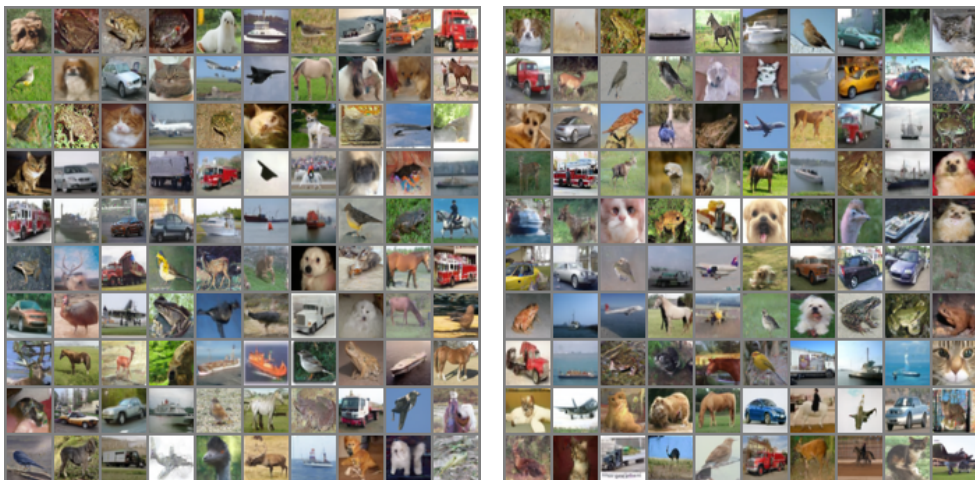


Figure 9: SN-DDPM + DC-DPM on 25 Denoising Steps.

1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527



Figure 10: SN-DDPM + DC-DPM on 50 Denoising Steps.

1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545



Figure 11: SN-DDPM + DC-DPM on 100 Denoising Steps.

1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565



Figure 12: GMS (Guo et al., 2024) + DC-DPM on 10 Denoising Steps.

1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581



Figure 13: **GMS + DC-DPM on 25 Denoising Steps.**

1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599



Figure 14: **GMS + DC-DPM on 50 Denoising Steps.**

1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619



Figure 15: **GMS + DC-DPM on 100 Denoising Steps.**