Almost Linear Convergence under Minimal Score Assumptions: Quantized Transition Diffusion

Xunpeng Huang[*] HUANGXUNPENG746@GMAIL.COM
University of California San Diego

Yingyu Lin[*] YIL208@UCSD.EDU
University of California San Diego

Nikki Lijing Kuang NIKKI.KUANG@GMAIL.COM
University of California San Diego

Hanze Dong HENDRYDONG@GMAIL.COM
Microsoft Research

Difan Zou[†] DZOU@CS.HKU.HK
University of Hong Kong

Yian Ma[†] YIANMA@UCSD.EDU
University of California San Diego

Tong Zhang TONGZHANG@TONGZHANG-ML.ORG
University of Illinois Urbana-Champaign

Editors: Steve Hanneke and Tor Lattimore

Abstract

Continuous diffusion models have demonstrated remarkable generative performance across diverse domains but are often constrained by the computational cost of simulating reverse Ornstein–Uhlenbeck processes via SDE/ODE solvers. Existing theoretical results typically establish query complexities that scale polynomially with both the dimension d and the error tolerance ϵ (e.g., $\tilde{\mathcal{O}}(d/\epsilon)$). This mirrors the limitations of unadjusted Langevin algorithm, where standard first-order score solvers lack access to zeroth-order density information, precluding natural error-correction mechanisms and thus preventing the fast $\ln(1/\epsilon)$ convergence attainable by Metropolis-adjusted methods. In this paper, we develop an improved generative modeling method by introducing Quantized Transition Diffusion (QTD), a framework that reformulates continuous diffusion into a discrete generation problem through spatial quantization and the parameterization of zeroth-order information (e.g., density ratios). To sample from this discrete target, we propose a truncated uniformization algorithm that simulates the underlying continuous-time Markov chain of the discrete diffusion process without discretization error, while eliminating the restrictive bounded-score assumption required by prior uniformization-based approaches. Consequently, QTD attains ϵ-accuracy in total variation distance with a query complexity of $\mathcal{O}(d\ln^2(d/\epsilon))$, yielding a notable improvement in ϵ-dependence compared to existing continuous diffusion samplers. Crucially, our analysis capitalizes on a novel proof technique based on the infinitesimal chain rule of KL divergence, providing a fresh perspective on unifying continuous and discrete diffusion paradigms.

Keywords: Quantized diffusion, Truncated uniformization, Linear convergence

∗. Equal contribution.
†. Corresponding author.

1. Introduction

Diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020) have emerged as a powerful and widely adopted class of generative models, achieving state-of-the-art performance across diverse domains. Their core mechanism is a noising–denoising procedure: the forward process incrementally corrupts training data by adding noise, thereby mapping an unknown and potentially complex distribution to a tractable prior (e.g., a standard Gaussian), while the reverse process progressively denoises samples back toward the original data distribution by estimating the logarithmic gradient (*score*) of the noised distributions (Vincent, 2011; Song and Ermon, 2019). Despite their remarkable empirical success, advancing our theoretical understanding of these models—particularly regarding how to generate high-quality samples in high-dimensional settings with reduced computational cost—remains a significant challenge.

Current theoretical analyses of continuous diffusion samplers in $\mathbb{R}^d$ primarily address the simulation of reverse Ornstein–Uhlenbeck (OU) processes via SDE/ODE solvers (Chen et al., 2023b,a; Benton et al., 2024a; Li and Yan, 2024). These approaches yield complexities that are *polynomial* in the total variation error tolerance ϵ, ranging from $\tilde{\mathcal{O}}(d/\epsilon^2)$ to $\tilde{\mathcal{O}}(d^{1/3}/\epsilon^{2/3})$ (Chen et al., 2023a; Huang et al., 2024b; Li and Jiao, 2025). Due to the deep connections between diffusion-based inference and Monte Carlo methods (Huang et al., 2024a; He et al., 2024), the fundamental limitation of these approaches stems from the fact that standard diffusion samplers operate solely on first-order scores without access to zeroth-order density information; consequently, they cannot employ natural error-correction mechanisms, such as a Metropolis–Hastings filter, to eliminate the discretization bias introduced by Euler-type approximations. As a result, convergence is inherently constrained by the discretization step size, precluding the fast exponential rates attainable in high-precision MCMC. Although several recent works (Chen et al., 2026; Gatmiry et al., 2026) attempt to match the $\ln(1/\epsilon)$ TV-convergence rate of Metropolis-adjusted MCMC (see Table 2), they rely on strong assumptions—such as restricting the target distribution to the convolution of a bounded distribution and a Gaussian, or imposing smoothness conditions on the forward marginal scores—that limit direct comparison with existing methods requiring only minimal score assumptions. The challenge of reconciling fast convergence with mild score assumptions has motivated research into discrete diffusion on finite state spaces (e.g., $\{1, 2, \ldots, \mathrm{K}\}^d$), where density ratios (zeroth-order information) are explicitly parameterized. Chen and Ying (2024) demonstrate that the continuous-time Markov chain (CTMC) governing the inference process can be simulated via *uniformization*—a Poisson process simulation technique—using only $\mathcal{O}(d \ln(d/\epsilon))$ discrete score evaluations. Nonetheless, their analysis relies on the restrictive assumption that discrete score estimators remain bounded; allowing unbounded scores can lead to uncontrollable transition rates and redundant updates, thereby hindering convergence. Given these obstacles, two fundamental questions arise:

1. Can uniformization provide a *linear-convergence* alternative to continuous diffusion samplers without incurring significant computational overhead?

2. Is it possible to remove the bounded-score assumption for discrete estimators without compromising efficiency?

In this paper, we introduce QTD, a novel **Quantized Transition Diffusion** framework that directly addresses these challenges, achieving the convergence of the total variation (TV) distance in $\mathcal{O}\big(d \ln^2(d/\epsilon)\big)$ expected score evaluations under minimal assumptions. To achieve linear convergence, we leverage zeroth-order information by adapting the density-ratio parameterization of

discrete diffusion to the continuous setting. Specifically, we partition high-probability regions of the target distribution into fine hypercubes and approximate the local density within each as uniform. By sufficiently refining these partitions, we ensure the induced histogram-like distribution approximates the true continuous distribution arbitrarily well in TV distance, effectively reducing the continuous sampling task to a discrete generation problem. Crucially, the availability of parameterized zeroth-order information allows for effective error correction; unlike Euler-type ODE/SDE solvers, where discretization error grows with step size, our method ensures that errors remain negligible even over long-distance state transitions. Furthermore, to eliminate the reliance on bounded discrete scores, we introduce *truncated uniformization*. This technique constrains the outgoing transition rates of each state, ensuring that the effective discrete scores remain naturally bounded. We prove that this modification preserves the *linear inference convergence* rate of standard uniformization. Our key contributions are summarized as follows:

- We propose the QTD framework and provably improve the inference rate from polynomial to logarithmic dependence on ϵ. Specifically, QTD generate d–dimensional samples to approximate the data distribution with $\Theta(\epsilon)$–TV error with only $\mathcal{O}(d \ln^2(d/\epsilon))$ expected score evaluations.
- We provide a constructive framework that models continuous distributions via efficient discrete diffusion. Under this framework, employing an efficient uniformization-based sampler ensures the output distribution remains arbitrarily close to the continuous target distribution.
- We introduce the *truncated uniformization* technique for an unbiased and tractable CTMC simulation. This method removes the restricted bounded-score assumption imposed in prior discrete diffusion analyses (Chen and Ying, 2024; Zhang et al., 2024).
- We develop a novel proof technique which is simpler and more general than Girsanov-based analyses used in (Chen and Ying, 2024; Zhang et al., 2024) for the discrete diffusion inference. In particular, we leverage the chain rule of KL divergence over infinitesimal time intervals to derive convergence guarantees.

2. Preliminaries

Given that our strategy rests on quantizing the continuous target distribution to obtain an approximation supported on a discrete space, we begin this section by formally defining the discrete forward and reverse Markov processes, parameterized by transition rate matrices. Subsequently, we detail the construction of noising and denoising dynamics on this discrete distribution for the purposes of training and inference. Furthermore, we introduce uniformization (van Dijk, 1992; van Dijk et al., 2018) as a technique for the efficient simulation of these processes. For reference, all notations introduced herein are summarized in Table 3 in Appendix A.

Problem setup. Without loss of generality, we focus on distributions that admit probability density functions in Euclidean space. These continuous density functions are represented by $p\colon \mathbb{R}^d \to \mathbb{R}^+$. Specifically, let the data distribution be $p_* \propto \exp(-f_*)$ for some potential function f_*. We consider the task of approximating p_* using some discrete distribution with probability mass function $q_*\colon \mathcal{Y} \to \mathbb{R}_0^+$, defined on a finite discrete space $\mathcal{Y}$. This discrete approximation is modeled via a forward Markov process $\{\mathbf{y}_t^{\rightarrow}\}_{t=0}^T$ and named as discrete diffusion model (Lou et al., 2024; Zhang et al., 2024; Chen and Ying, 2024), with initial distribution $q_0^{\rightarrow} = q_*$ that evolves toward

the uniform distribution. Then, the marginal distribution at time t is denoted by $q_t^{\rightarrow}$, the joint and conditional distributions over different time steps $t' > t$ are given by

$$(\mathbf{y}_{t'}^{\rightarrow}, \mathbf{y}_t^{\rightarrow}) \sim q_{t',t}^{\rightarrow} \quad \text{and} \quad q_{t'|t}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y}) = q_{t',t}^{\rightarrow}(\boldsymbol{y}', \boldsymbol{y})/q_t^{\rightarrow}(\boldsymbol{y}).$$

For simplicity, we set the forward process to be a time-homogeneous CTMC constructed via the transition rate function $R^{\rightarrow}\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, which implies that both conditional and marginal distributions satisfy

$$\frac{\mathrm{d}q_{t|s}^{\rightarrow}}{\mathrm{d}t}(\boldsymbol{y}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}} R^{\rightarrow}(\boldsymbol{y}, \boldsymbol{y}') \cdot q_{t|s}^{\rightarrow}(\boldsymbol{y}') \quad \text{and} \quad \frac{\mathrm{d}q_t^{\rightarrow}}{\mathrm{d}t}(\boldsymbol{y}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}} R^{\rightarrow}(\boldsymbol{y}, \boldsymbol{y}') \cdot q_t^{\rightarrow}(\boldsymbol{y}'). \tag{1}$$

The transition rate function $R^{\rightarrow}$ characterizes the instantaneous rate of transitioning from state $\boldsymbol{y}'$ to $\boldsymbol{y}$ and is formally defined as

$$R^{\rightarrow}(\boldsymbol{y}, \boldsymbol{y}') \coloneqq \lim_{\Delta t \to 0} \left[\Delta t^{-1} \cdot \left(q_{\Delta t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}') - \delta_{\boldsymbol{y}'}(\boldsymbol{y})\right)\right], \tag{2}$$

where $\delta_{\boldsymbol{y}'}(\boldsymbol{y}) = 1$ if $\boldsymbol{y} = \boldsymbol{y}'$ and 0 otherwise.

Reverse process. Additional properties of $R^{\rightarrow}$ are discussed in Appendix B.1. To sample from the target distribution $q_* = q_0^{\rightarrow}$ in practice, we simulate the reverse-time process $\mathbf{y}_t^{\leftarrow}$ that starts from $q_T^{\rightarrow}$ and moves backward.

$$\{\mathbf{y}_t^{\leftarrow}\}_{t=0}^{T} \quad \text{where} \quad \mathbf{y}_t^{\leftarrow} \sim q_t^{\leftarrow} = q_{T-t}^{\rightarrow}, \quad (\mathbf{y}_{t'}^{\leftarrow}, \mathbf{y}_t^{\leftarrow}) \sim q_{t',t}^{\leftarrow}, \quad \text{and} \quad q_{t',t}^{\leftarrow} = q_t^{\leftarrow} \cdot q_{t'|t}^{\leftarrow},$$

whose dynamic follows from

$$\frac{\mathrm{d}q_t^{\leftarrow}}{\mathrm{d}t}(\boldsymbol{y}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}} R_t^{\leftarrow}(\boldsymbol{y}, \boldsymbol{y}') \cdot q_t^{\leftarrow}(\boldsymbol{y}') \quad \text{where} \quad R_t^{\leftarrow}(\boldsymbol{y}, \boldsymbol{y}') \coloneqq R^{\rightarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \frac{q_t^{\leftarrow}(\boldsymbol{y})}{q_t^{\leftarrow}(\boldsymbol{y}')}, \tag{3}$$

proven in Appendix B.2. Similar to the $R^{\rightarrow}$ in the forward process, $R_t^{\leftarrow}$ characterizes the transition rates for the time-inhomogeneous reverse process $\{\mathbf{y}_t^{\leftarrow}\}_{t=0}^{T}$, i.e.,

$$R_t^{\leftarrow}(\boldsymbol{y}, \boldsymbol{y}') \coloneqq \lim_{\Delta t \to 0} \left[\frac{q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}|\boldsymbol{y}') - \delta_{\boldsymbol{y}'}(\boldsymbol{y})}{\Delta t^{-1}}\right], \tag{4}$$

as shown in Appendix B.3. In practice, the probability density ratio $q_t^{\leftarrow}(\boldsymbol{y})/q_t^{\leftarrow}(\boldsymbol{y}')$ will usually be approximated with neural networks due to its unknown closed form, which is presented as

$$\tilde{v}_{t,\boldsymbol{y}'}(\cdot) \approx v_{t,\boldsymbol{y}'}(\cdot) = q_t^{\leftarrow}(\cdot)/q_t(\boldsymbol{y}').$$

To simulate the reverse process in Eq. (3), we must estimate the time-varying rate matrix $R_t^{\leftarrow}$, which depends on the intractable ratio $q_t^{\leftarrow}(\boldsymbol{y})/q_t^{\leftarrow}(\boldsymbol{y}')$. We approximate this ratio using neural networks trained via score entropy minimization (Lou et al., 2024; Benton et al., 2024b),

$$L_{\mathrm{SE}}(\hat{v}) = \int_0^T \mathbb{E}_{\mathbf{y}_t \sim q_t^{\rightarrow}} \left[\sum_{\boldsymbol{y} \neq \mathbf{y}_t} R^{\rightarrow}(\mathbf{y}_t, \boldsymbol{y}) \cdot D_\phi\left(v_{T-t,\mathbf{y}_t}(\boldsymbol{y}) \middle\| \tilde{v}_{T-t,\mathbf{y}_t}(\boldsymbol{y})\right)\right] \mathrm{d}t, \tag{5}$$

where $D_\phi\left(\cdot\middle\|\cdot\right)$ denotes the Bregman divergence, and $\phi(c) = c \ln c$. Similar to the score estimation loss in continuous cases (Chen et al., 2023b), the loss L_{SE} is not directly estimable. Instead, implicit score entropy and denoising score entropy (Lou et al., 2024; Benton et al., 2024b) are introduced to enable an equivalent minimization.

Uniformization. Uniformization (van Dijk, 1992; van Dijk et al., 2018) is a sampling technique that enables the exact simulation of the discrete reverse process by a Poisson process (Chen and Ying, 2024). Consider the reverse process shown in Eq. (3), the conditional transition probability is

$$q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) = \begin{cases} \Delta t \cdot R^{\leftarrow}_t(\boldsymbol{y}', \boldsymbol{y}) & \boldsymbol{y}' \neq \boldsymbol{y} \\ 1 - \Delta t \sum_{\tilde{\boldsymbol{y}} \neq \boldsymbol{y}} R^{\leftarrow}_t(\tilde{\boldsymbol{y}}, \boldsymbol{y}) & \boldsymbol{y}' = \boldsymbol{y} \end{cases} \tag{6}$$

in an infinitesimal time Δt due to Eq. (4), where the $o(\Delta t)$ term is omitted. Suppose the probability of transitioning to a different state is upper bounded by $\Delta t \cdot \beta$:

$$\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R^{\leftarrow}_t(\boldsymbol{y}', \boldsymbol{y}) \coloneqq R^{\leftarrow}_t(\boldsymbol{y}) \leq \beta, \ \ \forall t. \tag{7}$$

We can then simulate Eq. (3) with the following uniformization procedure:

1. With probability $\Delta t \cdot \beta$, allow a state transition.

2. Conditioning on an allowed transition, move from $\boldsymbol{y}$ to $\boldsymbol{y}'$ with probability

$$\boldsymbol{M}_t(\boldsymbol{y}'|\boldsymbol{y}) = \begin{cases} \beta^{-1} R^{\leftarrow}_t(\boldsymbol{y}', \boldsymbol{y}) & \boldsymbol{y}' \neq \boldsymbol{y} \\ 1 - \beta^{-1} R^{\leftarrow}_t(\boldsymbol{y}) & \text{otherwise} \end{cases}.$$

Under these two steps, the practical conditional probability satisfies

$$\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) = \begin{cases} \Delta t \cdot \beta \cdot R^{\leftarrow}_t(\boldsymbol{y}', \boldsymbol{y}) \cdot \beta^{-1} = \Delta t \cdot R^{\leftarrow}_t(\boldsymbol{y}', \boldsymbol{y}) & \boldsymbol{y}' \neq \boldsymbol{y} \\ 1 - \Delta t \cdot \beta + \Delta t \cdot \beta \cdot (1 - \beta^{-1} \cdot R^{\leftarrow}_t(\boldsymbol{y})) = 1 - \Delta t \cdot R^{\leftarrow}_t(\boldsymbol{y}) & \boldsymbol{y}' = \boldsymbol{y}, \end{cases} \tag{8}$$

which exactly matches Eq. (6). Under this condition, the number of transition events within a time interval $[s, t]$ follows a Poisson distribution (van Dijk, 1992; van Dijk et al., 2018) whose expectation is $\beta(t - s)$, which coincides with the number of required evaluations of the transition rate function $R^{\leftarrow}_t$. This implies choosing a tighter upper bound β directly leads to better complexity.

3. Quantized Transition Diffusion

In this section, we introduce Quantized Transition Diffusion (QTD), a framework designed for efficient sampling from continuous data distributions. At a high level, QTD approximates the continuous target via a histogram-like distribution, thereby transforming the continuous generation task into a sampling problem over a discrete CTMC defined on a hypercube partition. Then, we construct the noising dynamics based on the adjacency structure induced by a binary encoding of this partition. This scheme effectively balances the trade-off between the out-degree of each state and the global connectivity of the transition graph, ensuring rapid forward mixing while maintaining low iteration complexity for the inference updates. Finally, we propose a *truncated uniformization* scheme to simulate the reverse-time CTMC. Crucially, this approach enables efficient simulation without relying on the restrictive boundedness assumption on discrete scores, a limitation prevalent in prior uniformization-based inference algorithms (Chen and Ying, 2024; Liang et al., 2025).

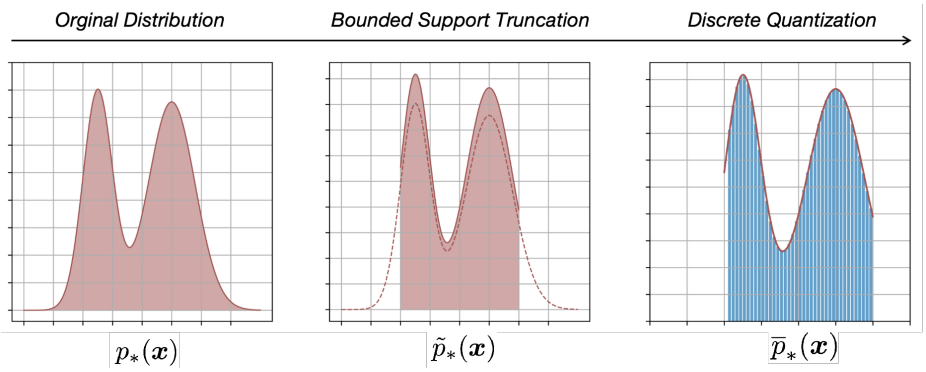

Figure 1: Visualization of the histogram approximation. The first step regularizes the original distribution in some bounded sets but controls the TV gap by Lemma 10. The second step quantizes the probability density to a histogram-like distribution but controls the TV gap by Lemma 11.

3.1. Histogram Approximation

To approximate the target distribution p_* defined in the Euclidean space $\mathbb{R}^d$ with a histogram-like distribution, we first restrict its support to a bounded region, which can be represented by a cube of side length L as follows:

$$\mathrm{Cube}\,(L) \coloneqq \{\boldsymbol{x} \,|\, -L \le \boldsymbol{x}_j \le L,\ \forall j \in \{1, 2, \ldots, d\}\}\,.$$

Given that $\mathrm{Cube}\,(L)$ covers most probability mass of p_*, we construct a probability density restricted to this region to approximate p_*:

$$\tilde{p}_*(\boldsymbol{x}) \coloneqq \frac{p_*(\boldsymbol{x})}{\int_{\boldsymbol{x}\in\mathrm{Cube}(L)} p_*(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}} \quad \forall \boldsymbol{x} \in \mathrm{Cube}\,(L)\,. \tag{9}$$

Standard concentration arguments allow us to control the TV distance between p_* and $\tilde{p}_*$. Next, we quantize $\tilde{p}_*$ over $\mathrm{Cube}\,(L)$ by discretizing each dimension into $K \coloneqq 2L/l$ intervals of width l, with partition points defined by:

$$l_i = -L + i \cdot l \quad i \in \{0, 1, \ldots, K\} \quad \text{and} \quad -L \le l_i \le L.$$

That means the high-dimensional cube $\mathrm{Cube}\,(L)$ will be decomposed into K^d cells (subsets), and each cell will cover a small region shown as follows

$$\mathrm{Cell}\,(i_0, i_1, \ldots, i_{d-1}) \coloneqq \left\{\boldsymbol{x} | l_{i_j} < \boldsymbol{x}_j \le l_{i_j+1}, \forall j \in \{0, 1, \ldots, d-1\}\right\}. \tag{10}$$

We construct the piecewise constant distribution $\overline{p}_*(\boldsymbol{x})$ by averaging the original density $\tilde{p}_*$ over each quantization cell. Specifically, for each cell $\mathrm{Cell}\,(i_0, i_1, \ldots, i_{d-1})$, we assign a constant density to all points $\boldsymbol{x}$ in the cell:

$$\overline{p}_*(\boldsymbol{x}) = l^{-d} \cdot \int_{\boldsymbol{u}\in\mathrm{Cell}(i_0,i_1,\ldots,i_{d-1})} \tilde{p}_*(\boldsymbol{u})\mathrm{d}\boldsymbol{u}, \quad \boldsymbol{x} \in \mathrm{Cell}\,(i_0, i_1, \ldots, i_{d-1})\,. \tag{11}$$

By construction, the quantized distribution satisfies the normalization condition $\int_{\mathrm{Cube}(L)} \overline{p}_*(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. The following lemma establishes a bound on the TV distance between the approximation $\overline{p}_*$ and

the ground truth p_*. This error analysis relies on two key properties: first, the sub-Gaussianity of the target distribution ensures that the bounded domain $\mathrm{Cube}\,(L)$ captures almost all the probability mass; second, the smoothness of p_* ensures that the histogram-like approximation incurs a controllable discretization error. Consequently, with an appropriate selection of the domain size L and grid resolution l, $\overline{p}_*$ converges to the continuous target distribution p_*.

Lemma 1 *Suppose the target distribution $p_* \propto \exp(-f_*)$ is σ sub-Gaussian and f_* is H–smooth, we can construct $\overline{p}_*$ defined on a finite cube $\mathrm{Cube}\,(L)$ with length*

$$L = \sigma \cdot \sqrt{2\ln(2d/\epsilon)} \quad \text{and} \quad l = \left[2H \cdot \left(\sigma\sqrt{2d\ln(2d/\epsilon)} + d + \sqrt{dm_0}\right)\right]^{-1} \cdot \epsilon,$$

to satisfy $\mathrm{TV}\,(p_*, \overline{p}_*) \le 3\epsilon$.

We defer the proof to Appendix C. Under this condition, we have constructed a histogram-like distribution to approximate p_*, which can be visualized by Fig. 1.

Under this transformation, although distribution $\overline{p}_*$ remains defined on $\mathbb{R}^d$, due to the histogram shape, it can be sampled by introducing a discrete distribution $\overline{q}_*$:

$$\overline{q}_*(\boldsymbol{y}) \propto \overline{p}_*(-L \cdot \mathbf{1} + l \cdot (\boldsymbol{y} - 0.5 \cdot \mathbf{1})), \quad \text{where} \quad \boldsymbol{y} \in \overline{\mathcal{Y}} \coloneqq \{0, 1, \ldots, K-1\}^d. \tag{12}$$

Here, each cell $\mathrm{Cell}\,(\mathbf{y})$ in the partition is treated as a discrete state whose probability mass is proportional to the probability density of $\overline{p}_*$ at the midpoint of that cell. Consequently, sampling from $\overline{p}$ reduces to the following two-stage procedure:

1. Sample from the discrete distribution $\overline{q}_*$ defined on $\overline{\mathcal{Y}}$;
2. Uniformly draw a sample from the cell, i.e., $\mathrm{Cell}\,(\boldsymbol{y})$.

Thus far, we have reformulated the continuous generation task as a sampling problem defined over a hypercube partition. To address this within the framework of discrete diffusion, the remaining objective is to construct appropriate noising and denoising dynamics—modeled as a CTMC—for the quantized distribution $\overline{q}_*$.

3.2. The Discrete Forward Process and Binary Encoding

As established in Section 2, the forward process on the discrete state space $\overline{\mathcal{Y}}$ is governed by the transition rate matrix $R^{\rightarrow}$. The design of $R^{\rightarrow}$ necessitates a fundamental trade-off between *reverse iteration complexity* and *forward mixing time*. On one hand, the computational cost of the reverse update scales with the number of non-zero entries in $R_t^{\leftarrow}$. By Eq. (3), $R_t^{\leftarrow}$ inherits the sparsity structure of $R^{\rightarrow}$; thus, a sparse $R^{\rightarrow}$ is essential for efficient inference. On the other hand, the mixing time quantifies the rate of state space traversal; generally, a denser $R^{\rightarrow}$ facilitates faster global exploration, standing in direct tension with the sparsity requirement.

Figure 2 illustrates this tension. Consider two extreme topologies for a d-dimensional space with K bins per dimension. Restricting transitions to immediate neighbors yields a favorable $\mathcal{O}(d)$ iteration complexity but a prohibitive $\mathcal{O}(Kd)$ mixing time, as the process must traverse the support incrementally. Conversely, a fully connected topology ensures rapid mixing but incurs an intractable $\mathcal{O}(K^d)$ computational cost per iteration. We therefore seek a middle ground that balances these two factors, targeting both iteration complexity and mixing time of $\mathcal{O}(d \log K)$.

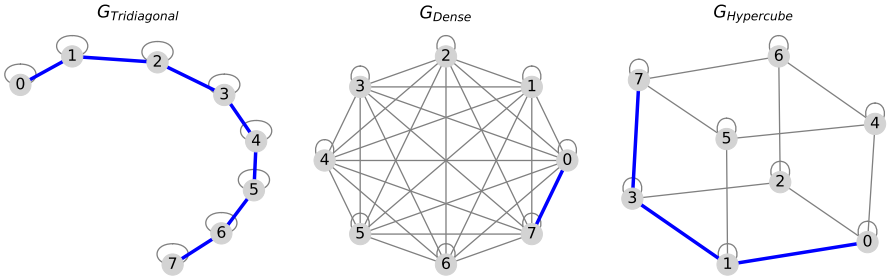

Figure 2: Comparison of transition graph topologies for discrete diffusion. We illustrate the connectivity of 8 discrete states ($d = 1, K = 8$) under different graph priors. Left ($G_{\text{Tridiagonal}}$): A simple neighbor-based chain yields low iteration complexity (degree 2) but suffers from a large diameter, leading to slow mixing. Center (G_{Dense}): A fully connected graph ensures one-step mixing but incurs a prohibitive $\mathcal{O}(K^d)$ cost. Right ($G_{\text{Hypercube}}$): Our proposed binary encoding (Section 3.2) maps states to a Boolean hypercube. This achieves an optimal balance, where both the out-degree (sparsity) and graph diameter (mixing) scale logarithmically with the number of bins, specifically $\mathcal{O}(d \log K)$. The bold blue edges highlight a diameter path—a shortest path between the two most distant vertices in each graph.

Specifically, we achieve this balance via Alg. 1 based on binary encoding. Assuming $\log_2 K$ is an integer, we define a bijection $\text{vBin}\colon \overline{\mathcal{Y}} \to \mathcal{Y} \coloneqq \{0,1\}^{d \log_2 K}$ that maps the original state $\overline{\boldsymbol{y}} = [\overline{\boldsymbol{y}}_0, \ldots, \overline{\boldsymbol{y}}_{d-1}] \in \overline{\mathcal{Y}}$ to a binary vector

$$\boldsymbol{y} = [\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{d \log_2 K - 1}], \quad \text{where} \quad \boldsymbol{y}_i = \left\lfloor \overline{\boldsymbol{y}}_{\lfloor i/\log_2 K \rfloor} \,/\, 2^{\left(i - \lfloor i/\log_2 K \rfloor\right)} \right\rfloor \bmod 2.$$

This embedding transforms the geometry into a Boolean hypercube. By restricting transitions to states with a Hamming distance of 1, we limit connectivity to immediate bit flips. Consequently, the neighbor count scales as $\mathcal{O}(d \log K)$, ensuring efficient reverse updates, while the graph diameter is similarly bounded by $\mathcal{O}(d \log K)$, facilitating rapid mixing. Formally, the forward transition rate $R^{\rightarrow}\colon \mathcal{Y} \times \mathcal{Y}$ on this binary space is defined as:

$$R^{\rightarrow}(\boldsymbol{y}, \boldsymbol{y}') = \begin{cases} 1 & \text{Ham}(\boldsymbol{y}, \boldsymbol{y}') = 1 \\ -\, d \log_2 K & \boldsymbol{y} = \boldsymbol{y}' \\ 0 & \text{otherwise} \end{cases}. \tag{13}$$

This definition yields a symmetric CTMC that behaves as a time-homogeneous random walk on the hypercube. As established in Lemma 2 (proof in Appendix D), the process exhibits exponential convergence, mirroring the behavior of continuous OU processes.

Lemma 2 *Suppose the transition rate function $R^{\rightarrow}$ of the CTMC $\{\mathbf{y}_t^{\rightarrow}\}_{t=0}^{T}$ is defined by Eq. (13). Then, the underlying distribution $q_t^{\rightarrow}$ of $\mathbf{y}_t^{\rightarrow}$ satisfies $\text{KL}\left(q_t^{\rightarrow} \middle\| q_{\infty}^{\rightarrow}\right) \le e^{-t} \cdot d \log_2 K$.*

3.3. Truncated Uniformization

Recall from Section 2 that the computational complexity of simulating a CTMC via uniformization is governed by the uniformization constant β (defined in Eq. (7)), which bounds the total transition rate out of any state. A tighter bound directly translates to improved simulation efficiency. Motivated by this, we analyze the time-dependent bound β_t for the time-inhomogeneous reverse CTMC

Algorithm 1 TRAINING DATA QUANTIZATION

Input: The training set $\mathcal{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$.
Initialize output set $\mathcal{Y} = \{\}$ and the parameters, e.g., L and l as shown in Lemma 1.
for $n = 1$ **to** N **do**
  Quantize the training sample $\boldsymbol{x}^{(n)}$ to $\overline{\boldsymbol{y}}^{(n)}$ via

$$\overline{\boldsymbol{y}}^{(n)} = [\overline{\boldsymbol{y}}_0^{(n)}, \overline{\boldsymbol{y}}_1^{(n)}, \ldots, \overline{\boldsymbol{y}}_{d-1}^{(n)}] \quad \text{where} \quad \overline{\boldsymbol{y}}_i^{(n)} = \lfloor (\boldsymbol{x}_i^{(n)} + L)/l \rfloor.$$

  Append the set $\mathcal{Y}$ with binary encoded $\boldsymbol{y}^{(n)} = \mathrm{vBin}(\overline{\boldsymbol{y}}^{(n)})$ where $\boldsymbol{y}^{(n)} \in \{0,1\}^{d \log_2 K}$.
end for
return $\mathcal{Y}$.

defined in Eq. (3). Specifically, when the forward transition rates are set according to Eq. (13), the resulting β_t satisfies the following lemma. The detailed proof is provided in Appendix E.1.

Lemma 3 *Consider a CTMC whose transition rate function $R^{\rightarrow}$ is defined as Eq. (13). Then, for any $\boldsymbol{y}$, the reverse transition rate function satisfies*

$$\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y}) \coloneqq R_t^{\leftarrow}(\boldsymbol{y}) \leq \beta_t \coloneqq (2d \log_2 K) \cdot \max\{1, (T-t)^{-1}\}. \tag{14}$$

Therefore, it is important for us to divide the entire reverse process into W segments. With a proper segmentation, we can assign a tight upper bound β_{t_w} for $R_t^{\leftarrow}(\boldsymbol{y})$ when $t \in [t_{w-1}, t_w)$ and minimize the expectation of transition events, given by $\sum_w \beta_{t_w} \cdot (t_w - t_{w-1})$. In practice, since the exact form $R_t^{\leftarrow}$ is intractable, we approximate it by minimizing Eq. (5):

$$R_t^{\leftarrow}(\boldsymbol{y}, \boldsymbol{y}') \approx \tilde{R}_t(\boldsymbol{y}, \boldsymbol{y}') = R^{\rightarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \tilde{v}_{t,\boldsymbol{y}'}(\boldsymbol{y}).$$

Here, $\tilde{v}_{t,\boldsymbol{y}'}(\boldsymbol{y})$ can approximate the ideal density ratio $q_t^{\leftarrow}(\boldsymbol{y})/q_t^{\leftarrow}(\boldsymbol{y}')$ with high accuracy. However, this approximation may violate the desired global rate bound in Lemma 3. To address this, prior work (Chen and Ying, 2024; Liang et al., 2025) impose an estimated score boundedness assumption for discrete diffusion inference:

$$\sum_{\boldsymbol{y} \neq \boldsymbol{y}'} \tilde{R}_t(\boldsymbol{y}, \boldsymbol{y}') \leq C d \log_2 K \cdot \max\{1, (T-t)^{-1}\} \tag{15}$$

We argue that this assumption can be safely removed by truncating the approximate transition rate function as follows:

$$\hat{R}_t(\boldsymbol{y}, \boldsymbol{y}') = \begin{cases} \tilde{R}_t(\boldsymbol{y}, \boldsymbol{y}') \cdot \beta_t / \tilde{R}_t(\boldsymbol{y}') & \tilde{R}_t(\boldsymbol{y}') > \beta_t \\ \tilde{R}_t(\boldsymbol{y}, \boldsymbol{y}') & \text{otherwise.} \end{cases}, \quad \forall \boldsymbol{y}' \neq \boldsymbol{y}, \tag{16}$$

and

$$\hat{R}_t(\boldsymbol{y}', \boldsymbol{y}') = -\sum_{\boldsymbol{y} \neq \boldsymbol{y}'} \hat{R}_t(\boldsymbol{y}, \boldsymbol{y}'). \tag{17}$$

It ensures that the total outgoing rate from any state does not exceed β_t, hence eliminating the need for explicit score bounds. Combining $\hat{R}_t$ with the two-step uniformization mentioned in Section 2, we obtain a practical and efficient inference algorithm, summarized in Alg. 2. Here, $\boldsymbol{e}_i$ denotes the one-hot vector with a 1 at position i and 0 elsewhere, and mod is an element-wise operator.

Algorithm 2 INFERENCE PROCESS WITH TRUNCATED UNIFORMIZATION

Input: Total time T, a time partition $0 = t_0 < \ldots < t_W = T - \delta$, parameters $\beta_{t_1}, \ldots, \beta_{t_W}$ set as Eq. (14), a reverse transition rate function $\hat{R}_t^{\leftarrow}$ obtained by the learnt score function $\tilde{v}_{t,\boldsymbol{y}'}(\cdot)$.
Draw an initial sample $\hat{\mathbf{y}}_{t_0} \sim \mathrm{Uniform}(\{0,1\}^{d \log_2 K})$.
for $w = 1$ **to** W **do**
  Draw $N \sim \mathrm{Poisson}(\beta_{t_w}(t_w - t_{w-1}))$;
  Sample N points i.i.d. uniformly from $[t_{w-1}, t_w]$ and sort them as $\tau_1 < \tau_2 < \ldots < \tau_N$;
  Set $\mathbf{z}_0 = \hat{\mathbf{y}}_{t_{w-1}}$;
  for $n = 1$ **to** N **do**
    Set

$$\mathbf{z}_n = \begin{cases} (\mathbf{z}_{n-1} + \boldsymbol{e}_i) \bmod 2, & w.p.\ \beta_{t_w}^{-1} \cdot \hat{R}_{\tau_n}(\boldsymbol{z}_{n-1} + \boldsymbol{e}_i, \boldsymbol{z}_{n-1}), \quad 0 \le i \le d \log_2 K - 1 \\ \mathbf{z}_{n-1}, & w.p.\ 1 - \beta_{t_w}^{-1} \cdot \hat{R}_{\tau_n}(\boldsymbol{z}_{n-1}). \end{cases}$$

  end for
  Set $\hat{\mathbf{y}}_{t_w} = \mathbf{z}_N$.
end for
Recover the cell index with $\overline{\mathbf{y}} = \mathrm{vBin}^{-1}(\hat{\mathbf{y}}_{t_W})$ and uniformly draw a sample $\hat{\mathbf{x}}$ from $\mathrm{Cell}(\overline{\mathbf{y}})$.
return $\hat{\mathbf{x}}$.

Empirical Results. We empirically validate that QTD achieves acceleration over standard DDPM baselines based on reversing the OU process, using both synthetic and toy real-world datasets. Due to space constraints, detailed experimental results are deferred to Appendix G.

4. Theoretical Results

In this section, we first analyze the convergence of the discrete generated distribution $\hat{q}_{T-\delta}$ in Alg. 2, demonstrating that it converges to the discrete target distribution $\overline{q}_*$ in KL divergence, contingent on accurate score estimation. A central technical innovation is the application of the infinite-time chain rule to bound the KL divergence between the marginal distributions of the ideal and approximated reverse processes. This technique serves as a robust alternative to the discrete Girsanov theorem, effectively circumventing the restrictive bounded score assumption. Subsequently, we extend these results to the continuous domain. Leveraging standard regularity assumptions from the diffusion model literature, we establish that the QTD-based method (Alg. 2) achieves ϵ-accuracy in TV distance to the continuous target p_* only using $\tilde{\mathcal{O}}(d)$ discrete score evaluations. Finally, we provide a theoretical comparison between the proposed scheme and existing continuous inference algorithms, highlighting the superior scaling properties of our approach.

Truncated Uniformization Results. To analyze the convergence and the query complexity required to achieve a target TV distance in the discrete setting, we introduce the following assumption regarding the accuracy of the discrete score estimation:

[A1] Score Estimation Error. Let $\mathcal{Y}$ be the discrete training set obtained by quantizing the continuous set $\mathcal{X}$ via Alg. 1. We assume the discrete score function $\tilde{v}_t$, trained according to Eq. (5), satisfies a bounded estimation error condition, i.e., $L_{\mathrm{SE}}(\hat{v}) \le \epsilon_{\mathrm{score}}^2$.

Lemma 4 *Under Assumption [A1], if the reverse process is implemented via algorithm 2, then:*

$$\mathrm{KL}\left(q^{\leftarrow}_{T-\delta}\middle\|\hat{q}_{T-\delta}\right) \leq \mathrm{KL}\left(q^{\leftarrow}_{0}\middle\|\hat{q}_{0}\right) + (T-\delta)\epsilon^2_{score}$$

This lemma establishes that the introduction of truncation mechanisms does not impede the convergence of the uniformization-type sampler toward the target distribution $q^{\leftarrow}_{T-\delta} \approx q_*$. A key challenge is that the standard discrete Girsanov Theorem is inapplicable to $\hat{R}_t$ (Eq. (16)) due to the truncation operator. To address this, the proof of Lemma 4 follows the sketch outlined below:

1. First, we analyze the time derivative of the KL divergence using the chain rule, and then exchange the limit and expectation operations:

$$\frac{\mathrm{dKL}\left(q^{\leftarrow}_t\middle\|\hat{q}_t\right)}{\mathrm{d}t} \leq \sum_{\boldsymbol{y}\in\mathcal{Y}} q^{\leftarrow}_t(\boldsymbol{y}) \cdot \underbrace{\lim_{\Delta t\to 0}\left[\frac{\mathrm{KL}\left(q^{\leftarrow}_{t+\Delta t|t}(\cdot|\boldsymbol{y})\middle\|\hat{q}_{t+\Delta t|t}(\cdot|\boldsymbol{y})\right)}{\Delta t}\right]}_{\text{Term 1}}.$$

2. The limit of the conditional KL divergence gap, i.e., Term 1, can be decomposed into two components, each relating to the discrepancy between the ideal and practical reverse transition rate matrices ($R^{\leftarrow}_t$ and $\hat{R}_t$):

$$\lim_{\Delta t\to 0}\left[\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} \frac{q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}{\Delta t} \cdot \ln\frac{q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}{\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}\right] = \sum_{\boldsymbol{y}'\neq\boldsymbol{y}} R^{\leftarrow}_t(\boldsymbol{y}',\boldsymbol{y}) \cdot \ln\frac{R^{\leftarrow}_t(\boldsymbol{y}',\boldsymbol{y})}{\hat{R}_t(\boldsymbol{y}',\boldsymbol{y})},$$

and

$$\lim_{\Delta t\to 0}\left[\Delta t^{-1} \cdot \left(1 - \sum_{\boldsymbol{y}'\neq\boldsymbol{y}} q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})\right) \cdot \ln\frac{1-\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}{1-\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}\right] = \hat{R}_t(\boldsymbol{y}) - R^{\leftarrow}_t(\boldsymbol{y}).$$

3. This formulation allows the truncation effects to be seamlessly incorporated into our analysis while coupling naturally with the learned transition rate matrix $\tilde{R}_t$ employed during training:

$$\begin{aligned}
&\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} R^{\leftarrow}_t(\boldsymbol{y}',\boldsymbol{y}) \cdot \ln\frac{R^{\leftarrow}_t(\boldsymbol{y}',\boldsymbol{y})}{\hat{R}_t(\boldsymbol{y}',\boldsymbol{y})} + \hat{R}_t(\boldsymbol{y}) - R^{\leftarrow}_t(\boldsymbol{y}) \\
&= \underbrace{\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} R^{\leftarrow}_t(\boldsymbol{y}',\boldsymbol{y}) \ln\frac{R^{\leftarrow}_t(\boldsymbol{y}',\boldsymbol{y})}{\tilde{R}_t(\boldsymbol{y}',\boldsymbol{y})} + \tilde{R}_t(\boldsymbol{y}) - R^{\leftarrow}_t(\boldsymbol{y})}_{\text{related to } \epsilon_{\text{score}}} + \underbrace{\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} R^{\leftarrow}_t(\boldsymbol{y}',\boldsymbol{y}) \ln\frac{\tilde{R}_t(\boldsymbol{y}',\boldsymbol{y})}{\hat{R}_t(\boldsymbol{y}',\boldsymbol{y})} + \hat{R}_t(\boldsymbol{y}) - \tilde{R}_t(\boldsymbol{y})}_{\text{Term 2}}.
\end{aligned}$$

4. Term 2 depends solely on the truncation operator. By exploiting the connection between $\tilde{R}_t$ and $\hat{R}_t$, we show that this term is provably non-positive, thereby obviating the need for a bounded score assumption and completing the analysis.

Since Alg. 2 serves as a self-contained sampling algorithm for discrete diffusion models, we provide the following convergence guarantee for the truncated uniformization sampler. We omit its proof, as it is nearly identical to that of Theorem 6.

Reference	Algorithm	Assumptions	Complexity (for TV)
Chen and Ying (2024)	Uniformization	**[A1]**, (15)	$\tilde{\mathcal{O}}(d)$
Zhang et al. (2024)	Euler-Method	**[A1]**	$\tilde{\mathcal{O}}(d^{4/3}\epsilon^{-4/3})$
Ren et al. (2025)	τ–leaping	**[A1]**	$\tilde{\mathcal{O}}(d\epsilon^{-1})$
This paper	Alg. 2	**[A1]**	$\tilde{\mathcal{O}}(d)$

Table 1: Comparison among discrete inference algorithms. Stopping time will be $T-\epsilon/d$ to guarantee the TV convergence with ϵ error tolerance.

Theorem 5 *Suppose Assumption **[A1]** holds and we aim to draw samples from the discrete set $\{0,1\}^{d\log_2 K}$ according to the distribution q_0. If we train a discrete diffusion model satisfying*

$$\epsilon_{score} \leq \frac{\epsilon}{\ln(d/\epsilon)+\ln\log_2 K} = \tilde{\mathcal{O}}(\epsilon),$$

and run Alg. 2, returning $\hat{\mathbf{y}}_{t_W}$, with

$$t_0=0, \quad t_{w+1}-t_w = 0.5\cdot(T-t_{w+1}), \quad t_W = T-\delta, \quad \textit{and} \quad \beta_{t_w} = \frac{2d\log_2 K}{\min\{1,\, T-t_w\}},$$

where

$$T=\ln(d/\epsilon)+\ln\log_2 K \quad \textit{and} \quad \delta \leq \frac{\epsilon}{d\log_2 K},$$

then the expected score estimation complexity of Alg. 2 is $\mathcal{O}(d\ln^2(d/\epsilon))$ to achieve $\mathrm{TV}\left(q_0, \hat{q}_{t_W}\right) = o(\epsilon)$, where $\hat{q}_{t_W}$ denotes the distribution of $\hat{\mathbf{y}}_{t_W}$.

In comparison to biased discrete inference, such as the Euler method (Zhang et al., 2024) and τ-leaping (Ren et al., 2025), which respectively require $\tilde{\mathcal{O}}(d^{4/3}\epsilon^{-4/3})$ and $\tilde{\mathcal{O}}(d\epsilon^{-1})$ complexity to ensure total variation convergence, truncated uniformization method only requires $\tilde{\mathcal{O}}(d)$ discrete score evaluations, significantly improving efficiency. Further distinguishing itself from standard uniformization methods, truncated uniformization removes the widely adopted assumption in Eq. (15), thereby significantly enhancing its practical applicability. Detailed comparison can be found in Table 1.

QTD-based Inference Results. To analyze the end-to-end convergence and gradient complexity required to achieve TV distance convergence, we make the following assumptions on p_*:

[A2] The second moment of p_* is bounded, i.e., $\mathbb{E}_{\mathbf{x}\sim p_*}[\|\mathbf{x}\|^2] \leq m_0$.

[A3] The energy function of p_* has bounded Hessian, i.e., $\|\nabla^2 \ln p_*\| \leq H$.

[A4] For any $\boldsymbol{u} \in \mathbb{R}^d$, there is a scalar sub-Gaussian tail, i.e.,

$$\mathbb{E}_{\mathbf{x}\sim p_*}\left[\exp\left(t\cdot \mathbf{x}^\top \boldsymbol{u}\right)\right] \leq \exp\left(\sigma^2 t^2 \|\boldsymbol{u}\|^2/2\right).$$

Assumptions **[A2]** and **[A3]** encapsulate the minimal smoothness conditions in Chen et al. (2023a), and Assumption **[A3]** can be circumvented via early stopping (Benton et al., 2024a; Chen et al., 2023a; Li and Yan, 2024). Though our analysis also assumes light tails, Assumption **[A4]** (the σ-sub-Gaussian property) is invoked mainly to ensure clear convergence guarantees; exponential tails would suffice. We impose no isoperimetric constraints, and p_* need not be log-concave or unimodal. Meanwhile, Assumption **[A1]** aligns with standard practice for discrete score estimation (Zhang et al., 2024; Chen and Ying, 2024). Crucially, no smoothness or boundedness is required for the intermediate scores $\tilde{v}_t$, enforcing only minimal conditions on the score. Under these assumptions, we establish the following theorem, with the proof deferred to Appendix E.2.

Theorem 6 *Suppose Assumptions **[A1]–[A4]** hold. If we apply Alg. 1 with*

$$L = \sigma \cdot \sqrt{2\ln(2d/\epsilon)}, \quad l = \left[2H \cdot \left(\sigma\sqrt{2d\ln(2d/\epsilon)} + d + \sqrt{dm_0}\right)\right]^{-1} \cdot \epsilon, \quad \text{and} \quad K = 2L/l$$

to quantize p_*, and run Alg. 2 with the same hyperparameter settings as in Theorem 5, then the expected score estimation complexity of Alg. 2 is $\mathcal{O}(d\ln^2(d/\epsilon))$ to achieve $\mathrm{TV}\left(p_*, \hat{p}\right) \leq 5\epsilon$, where $\hat{p}$ denotes the distribution of the generated samples.

Unlike conventional diffusion models that directly apply the noising–denoising procedure in Euclidean space, QTD achieves a state-of-the-art linear convergence rate with respect to the error tolerance ϵ, only requiring the additional mild sub-Gaussian assumption **[A4]**. Besides, we also provide Theorem 7, i.e., the early-stopping version of Theorem 6, in the following

Theorem 7 *Suppose Assumption **[A1]**, **[A2]** and **[A4]** hold. If we apply Alg. 1 with*

$$L \geq 2\sqrt{2d}\sigma \cdot \sqrt{\ln(2d\sigma/\epsilon)}, \quad l = \epsilon/\sqrt{d} \quad \text{and} \quad K = 2L/l,$$

to quantize p_* and run Alg. 2 with the same hyperparameter settings as in Theorem 5. Then the expectation of score estimation complexity of Alg. 2 will be $\mathcal{O}(d\ln^2(d/\epsilon))$ to achieve $\mathrm{TV}\left(\overline{p}_*, \hat{p}\right) \leq 2\epsilon$ where $\hat{p}$ denotes the underlying distribution of generated samples, and $\overline{p}_*$ approximates the data distribution in 2-Wasserstein distance, i.e., $W_2(p_*, \overline{p}_*) \leq 2\epsilon$.

The detailed description and proof will be deferred to Appendix E.4. As summarized in Table 2, our results simultaneously achieve fast convergence and operate under mild assumptions. Prior continuous diffusion samplers either attain only polynomial complexity in ϵ—ranging from $\tilde{\mathcal{O}}(d/\epsilon^2)$ to $\tilde{\mathcal{O}}(d^{5/4}/\epsilon^{1/2})$—or, when targeting logarithmic dependence on $1/\epsilon$, impose strong structural conditions such as smooth score requirements on the entire forward OU process (Chen et al., 2026) or the restriction that the target distribution takes the form of a bounded distribution convolved with a Gaussian (Gatmiry et al., 2026). In contrast, our method achieves a score estimation complexity of $\mathcal{O}(d\ln^2(d/\epsilon))$—the best among all listed approaches—while requiring neither the Euclidean score estimation error assumption **[A1]'** nor the **smooth score** condition on the forward marginals. Instead, our assumptions pertain only to the data distribution p_* itself and can be further relaxed via early stopping (Theorem 7), making our guarantees broadly applicable.

5. Conclusion and Limitation

In conclusion, we introduce a novel approach, QTD, which first quantizes the continuous data distribution into a discrete counterpart, and then applies a truncated uniformization procedure to

Table 2: Comparison with prior works simulating reverse particle SDEs, where **[A1]'** denotes the score estimation error trained in Euclidean space and **smooth score** denotes the smooth score assumption for the whole OU process (p_t) starting from p_*. Note that Assumptions **[A3]** is only about p_* and can be replaced by the early stopping trick. All complexities for TV convergence are achieved by assuming $\epsilon_{\text{score}} = \tilde{o}(\epsilon)$.

Results	Early-Stopping	Assumptions	Complexity (for TV)
Chen et al. (2023b)	No	**[A1]'** , **[A2]**, **smooth score**	$\tilde{\mathcal{O}}(d\epsilon^{-2})$
Chen et al. (2023a)	No	**[A1]'**, **[A2]**, **[A3]**	$\tilde{\mathcal{O}}(d^2\epsilon^{-2})$
Huang et al. (2024b)	No	**[A1]'** , **[A2]**, **smooth score**	$\tilde{\mathcal{O}}(d^{1/2}\epsilon^{-1})$
Gatmiry et al. (2026)	No	$p_* = p_0$(R-bounded) $* \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$	$\mathcal{O}((R/\sigma)^2 \ln^2(1/\epsilon))$
Chen et al. (2026)	No	**[A1]'** , **[A2]**, **smooth score**	$\mathcal{O}(d \ln^3(d/\epsilon))$
Our Theorem 6	No	**[A1]**, **[A2]**, **[A3]**, **[A4]**	$\mathcal{O}(d \ln^2(d/\epsilon))$
Chen et al. (2023a)	Yes	**[A1]'**, **[A2]**	$\tilde{\mathcal{O}}(d^2\epsilon^{-2})$
Benton et al. (2024a)	Yes	**[A1]'**, **[A2]**	$\tilde{\mathcal{O}}(d\epsilon^{-2})$
Li and Yan (2024)	Yes	**[A1]'**, **[A2]**	$\tilde{\mathcal{O}}(d\epsilon^{-1})$
Li and Cai (2024)	Yes	**[A1]'** ,**[A2]**	$\tilde{\mathcal{O}}(d^{5/4}\epsilon^{-1/2})$
Our Theorem 7	Yes	**[A1]**, **[A2]**, **[A4]**	$\mathcal{O}(d \ln^2(d/\epsilon))$

achieve unbiased inference with improved score-evaluation complexity for continuous data generation. Beyond its state-of-the-art theoretical complexity, the truncated uniformization framework is of independent interest as an inference algorithm for discrete diffusion models, where it also attains top-tier theoretical complexity under minimal assumptions.

A key limitation of our approach is that achieving accelerated convergence without degrading generation quality requires the discrete score estimation error to be on par with the continuous score estimation error outlined by (Chen et al., 2023b; Benton et al., 2024a). While some works (Meng et al., 2022; Lou et al., 2024) have introduced discrete training objectives such as concrete score matching and denoising score entropy, no direct comparison between discrete and continuous score training has been conducted.

Acknowledgments

The research is supported by NSF Award CCF-2112665 (TILOS) and CDC-RFA-FT-23-0069 from the CDC's Center for Forecasting and Outbreak Analytics. Difan Zou also acknowledges the support from Hong Kong GRF award 17307425 and ECS award 27309624.

References

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024a.

Joe Benton, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. From denoising diffusions to denoising markov models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):286–301, 2024b.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Fan Chen, Sinho Chewi, Constantinos Daskalakis, and Alexander Rakhlin. High-accuracy sampling for diffusion models and log-concave distributions. *arXiv preprint arXiv:2602.01338*, 2026.

Hongrui Chen and Lexing Ying. Convergence analysis of discrete diffusion model: Exact implementation through uniformization. *arXiv preprint arXiv:2402.08095*, 2024.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023b.

Khashayar Gatmiry, Sitan Chen, and Adil Salim. High-accuracy and dimension-free sampling with diffusions. *arXiv preprint arXiv:2601.10708*, 2026.

Ye He, Kevin Rojas, and Molei Tao. Zeroth-order sampling methods for non-log-concave distributions: Alleviating metastability by denoising diffusion. *Advances in Neural Information Processing Systems*, 37:71122–71161, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xunpeng Huang, Hanze Dong, Yifan Hao, Yi An Ma, and Tong Zhang. Reverse diffusion monte carlo. In *12th International Conference on Learning Representations, ICLR 2024*, 2024a.

Xunpeng Huang, Difan Zou, Hanze Dong, Zhang Zhang, Yian Ma, and Tong Zhang. Reverse transition kernel: A flexible framework to accelerate diffusion inference. *Advances in Neural Information Processing Systems*, 37:95515–95578, 2024b.

Gen Li and Changxiao Cai. Provable acceleration for diffusion models under minimal assumptions. *arXiv preprint arXiv:2410.23285*, 2024.

Gen Li and Yuchen Jiao. Improved convergence rate for diffusion probabilistic models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Gen Li and Yuling Yan. O (d/t) convergence theory for diffusion probabilistic models under minimal assumptions. *arXiv preprint arXiv:2409.18959*, 2024.

Yuchen Liang, Renxiang Huang, Lifeng Lai, Ness Shroff, and Yingbin Liang. Absorb and converge: Provable convergence guarantee for absorbing discrete diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VsZzTSyk5p.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32819–32848, 2024.

Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. *Advances in Neural Information Processing Systems*, 35: 34532–34545, 2022.

Yinuo Ren, Haoxuan Chen, Yuchen Zhu, Wei Guo, Yongxin Chen, Grant M Rotskoff, Molei Tao, and Lexing Ying. Fast solvers for discrete diffusion models: Theory and applications of high-order algorithms. *arXiv preprint arXiv:2502.00234*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Nico M van Dijk. Approximate uniformization for continuous-time markov chains with an application to performability analysis. *Stochastic processes and their applications*, 40(2):339–357, 1992.

Nico M van Dijk, Sem PJ van Brummelen, and Richard J Boucherie. Uniformization: Basics, extensions and applications. *Performance evaluation*, 118:8–32, 2018.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

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

Zikun Zhang, Zixiang Chen, and Quanquan Gu. Convergence of score-based discrete diffusion models: A discrete-time analysis. *arXiv preprint arXiv:2410.02321*, 2024.

Contents

Appendix A. Notation Summary

We summarize all notations used in the main paper and appendix in Table 3.

Table 3: Summary of key notations used in the paper.

Symbol	Description
$\mathrm{Cube}\,(L)$	Bounded cube $[-L, L]^d$ covering high-probability mass of p_*
$\mathrm{Cell}\,(i_0, \ldots, i_{d-1})$	Quantization cell (hypercubes) defined by coordinate bins, Eq. (10)
$\mathcal{Y}$	Binary discrete space $\{0,1\}^{d \log_2 K}$
$\overline{\mathcal{Y}}$	Grid index space $\{0, \ldots, K-1\}^d$
$\mathrm{vBin}(\cdot)$	Mapping from grid index $\overline{\mathcal{Y}}$ to binary code $\mathcal{Y}$
$p_* \propto \exp(-f_*)$	Target continuous distribution in $\mathbb{R}^d$
$\tilde{p}_*$	Truncated and renormalized version of p_* over $\mathrm{Cube}\,(L)$, Eq. (9)
$\overline{p}_*$	Histogram approximation to $\tilde{p}_*$ over $\mathrm{Cube}\,(L)$, Eq. (11)
$\overline{q}_*$	Discrete distribution on $\overline{\mathcal{Y}} = \{0, \ldots, K-1\}^d$ induced by $\overline{p}_*$, Eq. (12)
q_*	Discrete distribution on $\mathcal{Y} = \{0,1\}^{d \log_2 K}$, $q_* = \overline{q}_* \circ \mathrm{vBin}^{-1}$
$\mathbf{y}_t^{\rightarrow}$	Forward-time CTMC on $\mathcal{Y}$
$q_t^{\rightarrow}$	Marginal distribution of forward process at time t, i.e., $\mathbf{y}_t^{\rightarrow} \sim q_t^{\rightarrow}$
$q_{t',t}^{\rightarrow}$	Joint distribution of $(\mathbf{y}_{t'}^{\rightarrow}, \mathbf{y}_t^{\rightarrow})$
$q_{\infty}^{\rightarrow}$	Stationary distribution of the forward CTMC (uniform distribution)
$q_{t'\|t}^{\rightarrow}(\boldsymbol{y}'\|\boldsymbol{y})$	Conditional transition probability in forward process, Eq. (1)
$\mathbf{y}_t^{\leftarrow}$	Reverse-time CTMC defined by $q_t^{\leftarrow} := q_{T-t}^{\rightarrow}$, $\mathbf{y}_t^{\leftarrow} \sim q_t^{\leftarrow}$
$q_t^{\leftarrow}$	Marginal distribution of reverse process at time t, $q_t^{\leftarrow} = q_{T-t}^{\rightarrow}$
$q_{t',t}^{\leftarrow}$	Joint distribution of $(\mathbf{y}_{t'}^{\leftarrow}, \mathbf{y}_t^{\leftarrow})$
$q_{t'\|t}^{\leftarrow}(\boldsymbol{y}'\|\boldsymbol{y})$	Conditional transition probability of the ideal reverse process
$\hat{q}_{t+\Delta t\|t}(\boldsymbol{y}'\|\boldsymbol{y})$	Practical reverse conditional probability, Eq. (8)
$R^{\rightarrow}(\boldsymbol{y}, \boldsymbol{y}')$	Forward transition rate from state $\boldsymbol{y}'$ to $\boldsymbol{y}$, Eq. (2), and Eq. (13). This follows the ordering of the conditional distribution $p(\boldsymbol{y}\|\boldsymbol{y}')$, which is the *transpose* of the convention used in some other works.
$R_t^{\leftarrow}(\boldsymbol{y}, \boldsymbol{y}')$	Reverse transition rate at time t from state $\boldsymbol{y}'$ to $\boldsymbol{y}$, $R_t^{\leftarrow}(\boldsymbol{y}, \boldsymbol{y}') := R^{\rightarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \frac{q_t^{\leftarrow}(\boldsymbol{y})}{q_t^{\leftarrow}(\boldsymbol{y}')}$, Eq. (3)
$\tilde{R}_t(\boldsymbol{y}, \boldsymbol{y}')$	Estimated reverse transition rate using the learned density ratio, $\tilde{R}_t(\boldsymbol{y}, \boldsymbol{y}') = R^{\rightarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \tilde{v}_{t,\boldsymbol{y}'}(\boldsymbol{y})$, Eq. (5)
$\hat{R}_t(\cdot, \cdot)$	Truncated version of $\tilde{R}_t(\cdot, \cdot)$ with threshold β_t, Eq. (16)
$R_t^{\leftarrow}(\boldsymbol{y}),\ \tilde{R}_t(\boldsymbol{y}),\ \hat{R}_t(\boldsymbol{y})$	Total reverse transition rate out of state $\boldsymbol{y}$ for each rate type, defined as $R(\boldsymbol{y}) := \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R(\boldsymbol{y}', \boldsymbol{y})$ with $R \in \{R_t^{\leftarrow},\ \tilde{R}_t,\ \hat{R}_t\}$
β_t	Upper bound on $R_t^{\leftarrow}(\boldsymbol{y})$, $\beta_t = 2d \log_2 K \max\{1, (T-t)^{-1}\}$, Eq. (14)
$v_{t,\boldsymbol{y}'}(\boldsymbol{y})$	Density ratio $q_t^{\leftarrow}(\boldsymbol{y}) / q_t^{\leftarrow}(\boldsymbol{y}')$
$\tilde{v}_{t,\boldsymbol{y}'}(\boldsymbol{y})$	Learned approximation to $v_{t,\boldsymbol{y}'}(\boldsymbol{y}) = q_t^{\leftarrow}(\boldsymbol{y}) / q_t^{\leftarrow}(\boldsymbol{y}')$
$L_{\mathrm{SE}}(\hat{v})$	Score entropy loss used to train $\tilde{v}$, Eq. (5)
$\boldsymbol{e}_i$	One-hot vector with a 1 at position i and 0 elsewhere
l	Width of each quantization cell
$K = 2L/l$	Number of quantization bins per dimension

Appendix B. Forward and Reverse Processes of Discrete Diffusion Models

In order to simplify the notation in this section, we introduce some new notations as supplementary to Section 2. Since we consider the discrete diffusion on $\mathcal{Y}$, we defined the inner product on this discrete space for two functions as

$$\langle f, g\rangle_{\mathcal{Y}} \coloneqq \sum_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{y})\cdot g(\boldsymbol{y}).$$

Besides, the delta on $\mathcal{Y}$ is defined as

$$\delta_{\boldsymbol{y}}(\boldsymbol{y}') = \begin{cases} 1 & \boldsymbol{y}' = \boldsymbol{y} \\ 0 & \text{otherwise} \end{cases}.$$

B.1. The Forward Process of Discrete Diffusion Models

In this section, we refine the introduction about the forward process of discrete diffusion in Section 2 with the same notations. In general, the time-homogeneous CTMC can be described by a Markov semigroup $\mathcal{Q}_t^{\rightarrow}$ defined as:

$$\mathcal{Q}_t^{\rightarrow}[f](\boldsymbol{y}) = \mathbb{E}\left[f(\mathbf{y}_t)|\mathbf{y}_0 = \boldsymbol{y}\right] = \left\langle f, q_{t|0}^{\rightarrow}(\cdot|\boldsymbol{y})\right\rangle_{\mathcal{Y}} \tag{18}$$

where the function $f\colon \mathcal{Y}\to\mathbb{R}$. Due to the definition, the infinitesimal operator $\mathcal{L}^{\rightarrow}$ of the time homogeneous $\mathcal{Q}_t^{\rightarrow}$ is denoted as

$$\mathcal{L}^{\rightarrow}[f](\boldsymbol{y}) = \lim_{t\to 0}\left[\frac{\mathcal{Q}_t^{\rightarrow}[f]-f}{t}\right](\boldsymbol{y}) = \left\langle f, \partial_t q_{t|0}^{\rightarrow}(\cdot|\boldsymbol{y})\Big|_{t=0}\right\rangle_{\mathcal{Y}} \coloneqq \langle f, R^{\rightarrow}(\cdot,\boldsymbol{y})\rangle_{\mathcal{Y}} \tag{19}$$

where

$$R^{\rightarrow}(\boldsymbol{y}',\boldsymbol{y}) \coloneqq \partial_t q_{t|0}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y})\Big|_{t=0} = \lim_{t\to 0}\left[\frac{q_{t|0}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y}) - \delta_{\boldsymbol{y}}(\boldsymbol{y}')}{t}\right]. \tag{20}$$

According to the time-homogeneous property, we have

$$q_{t+\Delta t|t}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y}) = \delta_{\boldsymbol{y}}(\boldsymbol{y}') + \Delta t\cdot R^{\rightarrow}(\boldsymbol{y}',\boldsymbol{y}) + o(\Delta t)$$

for any t. Here, the transition rate function $R^{\rightarrow}$ must satisfy

$$R^{\rightarrow}(\boldsymbol{y},\boldsymbol{y}') \geq 0 \text{ when } \boldsymbol{y}'\neq\boldsymbol{y} \quad \text{and} \quad R^{\rightarrow}(\boldsymbol{y}',\boldsymbol{y}') = -\sum_{\boldsymbol{y}\neq\boldsymbol{y}'} R^{\rightarrow}(\boldsymbol{y},\boldsymbol{y}') \leq 0 \tag{21}$$

due to the definition Eq. (20). Under this setting, we can provide the dynamic of $q_{t|0}$ for any t. Specifically, we have

$$\partial_t\mathcal{Q}_t^{\rightarrow}[f](\boldsymbol{y}) = \mathcal{Q}_t^{\rightarrow}\left[\mathcal{L}f\right](\boldsymbol{y}) = \left\langle \mathcal{L}^{\rightarrow}f, q_{t|0}^{\rightarrow}(\cdot|\boldsymbol{y})\right\rangle_{\mathcal{Y}} = \sum_{\boldsymbol{y}'\in\mathcal{Y}} \mathcal{L}^{\rightarrow}[f](\boldsymbol{y}')\cdot q_{t|0}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y})$$

$$= \sum_{y'\in\mathcal{Y}}\left[\sum_{\tilde{\boldsymbol{y}}\in\mathcal{Y}} f(\tilde{\boldsymbol{y}})\cdot R^{\rightarrow}(\tilde{\boldsymbol{y}},\boldsymbol{y}')\cdot q_{t|0}(\boldsymbol{y}'|\boldsymbol{y})\right] = \sum_{\tilde{\boldsymbol{y}}\in\mathcal{Y}}\left[f(\tilde{\boldsymbol{y}})\cdot\sum_{\boldsymbol{y}'\in\mathcal{Y}} R^{\rightarrow}(\tilde{\boldsymbol{y}},\boldsymbol{y}')\cdot q_{t|0}(\boldsymbol{y}'|\boldsymbol{y})\right],$$

where the first inequality follows from the semigroup property. Combining with the fact

$$\partial_t \mathcal{Q}_t^{\rightarrow}[f](\boldsymbol{y}) = \left\langle f, \partial_t q_{t|0}^{\rightarrow}(\cdot|\boldsymbol{y}) \right\rangle_{\mathcal{Y}}$$

derived from Eq. (18), we have

$$\partial_t q_{t|0}^{\rightarrow}(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}} R(\tilde{\boldsymbol{y}}, \boldsymbol{y}') \cdot q_{t|0}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y}) = \left\langle R(\tilde{\boldsymbol{y}}, \cdot), q_{t|0}^{\rightarrow}(\cdot|\boldsymbol{y}) \right\rangle_{\mathcal{Y}}.$$

According to the time-homogeneous property, the above equation can be easily extended to

$$\partial_t q_{t|s}^{\rightarrow}(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}} R(\tilde{\boldsymbol{y}}, \boldsymbol{y}') \cdot q_{t|s}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y}) = \left\langle R(\tilde{\boldsymbol{y}}, \cdot), q_{t|s}^{\rightarrow}(\cdot|\boldsymbol{y}) \right\rangle_{\mathcal{Y}}. \tag{22}$$

Combining with Bayes' Theorem, the transition of the marginal distribution is

$$\frac{\mathrm{d}q_t^{\rightarrow}}{\mathrm{d}t}(\boldsymbol{y}) = \langle R(\boldsymbol{y}, \cdot), q_t^{\rightarrow} \rangle_{\mathcal{Y}}. \tag{23}$$

Matrix Presentation. Suppose the support set $\mathcal{Y}$ of $q_t^{\rightarrow}$ be written as $\mathcal{Y} = \{\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{|\mathcal{X}|}\}$, we may consider the marginal distribution $q_s^{\rightarrow}$ to be a vector, i.e.,

$$\boldsymbol{q}_t^{\rightarrow} = \left[q_t(\boldsymbol{y}_0), q_t(\boldsymbol{y}_1), \ldots, q_t(\boldsymbol{y}_{|\mathcal{Y}|-1})\right],$$

conditional transition probability function $q_{t|s}^{\rightarrow}$ to be a matrix, i.e.,

$$\boldsymbol{Q}_{t|s}^{\rightarrow} = \begin{bmatrix} q_{t|s}^{\rightarrow}(\boldsymbol{y}_0|\boldsymbol{y}_0) & q_{t|s}^{\rightarrow}(\boldsymbol{y}_0|\boldsymbol{y}_1) & \cdots & q_{t|s}^{\rightarrow}(\boldsymbol{y}_0|\boldsymbol{y}_{|\mathcal{Y}|-1}) \\ q_{t|s}^{\rightarrow}(\boldsymbol{y}_1|\boldsymbol{y}_0) & q_{t|s}^{\rightarrow}(\boldsymbol{y}_1|\boldsymbol{y}_1) & \cdots & q_{t|s}^{\rightarrow}(\boldsymbol{y}_1|\boldsymbol{y}_{|\mathcal{Y}|-1}) \\ \cdots & \cdots & \cdots & \cdots \\ q_{t|s}^{\rightarrow}(\boldsymbol{y}_{|\mathcal{Y}|-1}|\boldsymbol{y}_0) & q_{t|s}^{\rightarrow}(\boldsymbol{y}_{|\mathcal{Y}|-1}|\boldsymbol{y}_1) & \cdots & q_{t|s}^{\rightarrow}(\boldsymbol{y}_{|\mathcal{Y}|-1}|\boldsymbol{y}_{|\mathcal{Y}|-1}) \end{bmatrix}.$$

Similarly, the function R can also be presented as

$$\boldsymbol{R}^{\rightarrow} = \begin{bmatrix} R^{\rightarrow}(\boldsymbol{y}_0, \boldsymbol{y}_0) & R^{\rightarrow}(\boldsymbol{y}_0, \boldsymbol{y}_1) & \cdots & R^{\rightarrow}(\boldsymbol{y}_0, \boldsymbol{y}_{|\mathcal{Y}|-1}) \\ R^{\rightarrow}(\boldsymbol{y}_1, \boldsymbol{y}_0) & R^{\rightarrow}(\boldsymbol{y}_1, \boldsymbol{y}_1) & \cdots & R^{\rightarrow}(\boldsymbol{y}_1, \boldsymbol{y}_{|\mathcal{Y}|-1}) \\ \cdots & \cdots & \cdots & \cdots \\ R^{\rightarrow}(\boldsymbol{y}_{|\mathcal{Y}|-1}, \boldsymbol{y}_0) & R^{\rightarrow}(\boldsymbol{y}_{|\mathcal{Y}|-1}, \boldsymbol{y}_1) & \cdots & R^{\rightarrow}(\boldsymbol{y}_{|\mathcal{Y}|-1}, \boldsymbol{y}_{|\mathcal{Y}|-1}) \end{bmatrix}. \tag{24}$$

Under this condition, Eq. (23) can be written as

$$\mathrm{d}\boldsymbol{q}_t^{\rightarrow}/\mathrm{d}t = \boldsymbol{R}^{\rightarrow} \cdot \boldsymbol{q}_t^{\rightarrow} \tag{25}$$

matching the usual presentation shown in Chen and Ying (2024); Zhang et al. (2024). Besides, Eq. (21) shown in Section 2 can also be presented as $\mathbf{1} \cdot \boldsymbol{R} = \mathbf{0}$.

The following lemma gives the closed-form expression for the probability transition kernel of the forward process, which also suggests an efficient implementation.

Lemma 8 (Forward transition kernel) *Consider the forward CTMC, i.e., $\{\mathbf{y}_t\}_{t=0}^T$ with the infinitesimal operator $\mathbb{R}^{\rightarrow}$ given in Eq. (13). Then for any two timestamps $s \leq t$, the forward transition probability satisfies*

$$q^{\rightarrow}_{t|s}(\boldsymbol{y}|\boldsymbol{y}') = 2^{-d\log_2 K} \cdot \prod_{i=0}^{d\log_2 K - 1} \left[1 + (-1)^{|\boldsymbol{y}_i - \boldsymbol{y}'_i|} \cdot e^{-2(t-s)}\right].$$

Remark 9 *The transition probability in Lemma 8 factorizes across coordinates. This means that the forward transition can be implemented as $d\log_2 K$ independent bit-wise updates. Specifically, for each coordinate i, flip $\boldsymbol{y}'_i$ with probability $\frac{1-e^{-2(t-s)}}{2}$ to obtain $\boldsymbol{y}_i$.*

Proof Combining Eq. (24) and Eq. (25), the dynamic of marginal distribution $q_t^{\rightarrow}$ can be written as a matrix-vector product, i.e.,

$$\mathrm{d}\boldsymbol{q}_t^{\rightarrow}/\mathrm{d}t = \boldsymbol{R}^{\rightarrow} \cdot \boldsymbol{q}_t^{\rightarrow}$$

where

$$\boldsymbol{R}^{\rightarrow} = \begin{bmatrix} R^{\rightarrow}(\boldsymbol{y}_0, \boldsymbol{y}_0) & R^{\rightarrow}(\boldsymbol{y}_0, \boldsymbol{y}_1) & \dots & R^{\rightarrow}(\boldsymbol{y}_0, \boldsymbol{y}_{|\mathcal{Y}|-1}) \\ R^{\rightarrow}(\boldsymbol{y}_1, \boldsymbol{y}_0) & R^{\rightarrow}(\boldsymbol{y}_1, \boldsymbol{y}_1) & \dots & R^{\rightarrow}(\boldsymbol{y}_1, \boldsymbol{y}_{|\mathcal{Y}|-1}) \\ \dots & \dots & \dots & \dots \\ R^{\rightarrow}(\boldsymbol{y}_{|\mathcal{Y}|-1}, \boldsymbol{y}_0) & R^{\rightarrow}(\boldsymbol{y}_{|\mathcal{Y}|-1}, \boldsymbol{y}_1) & \dots & R^{\rightarrow}(\boldsymbol{y}_{|\mathcal{Y}|-1}, \boldsymbol{y}_{|\mathcal{Y}|-1}) \end{bmatrix}.$$

Here, $\boldsymbol{R}^{\rightarrow}$ can be decomposed into the sum

$$\boldsymbol{R}^{\rightarrow} = \sum_{i=0}^{d\log_2 K - 1} \boldsymbol{R}_i^{\rightarrow},$$

we first note that the state space is $\{0,1\}^{d\log_2 K}$, where each coordinate can flip independently. Hence, each coordinate contributes its own "flip" component to the overall generator $\boldsymbol{R}^{\rightarrow}$. Concretely, let us label the coordinates $0, \dots, d\log_2 K - 1$, and consider the generator corresponding to a single coordinate i. Such a generator only acts nontrivially on the ith coordinate, which can flip from 0 to 1 or 1 to 0, while all other coordinates remain unchanged.

Each "flip" for coordinate i can be represented by a 2×2 generator matrix (reflecting the two possible states, 0 or 1). Moreover, since the flipping of different coordinates occurs independently, we adopt the tensor-product (or Kronecker-product) structure, placing the 2×2 flip matrix in the ith position and 2×2 identity matrices in all other positions. Hence, each $\boldsymbol{R}_i^{\rightarrow}$ is of the form

$$\boldsymbol{R}_i^{\rightarrow} = \boldsymbol{I} \otimes \cdots \otimes \boldsymbol{A} \otimes \cdots \otimes \boldsymbol{I},$$

where

$$\boldsymbol{A} \coloneqq \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

is a generator of the flip in the ith coordinate, and $\boldsymbol{I}$ is the 2×2 identity in all coordinates. By the Kolmogorov forward equation, we have

$$\boldsymbol{Q}^{\rightarrow}_{t|s} = \exp\left((t-s)\boldsymbol{R}^{\rightarrow}\right) = \exp\left((t-s)\boldsymbol{A}\right)^{\otimes d} = \begin{bmatrix} \frac{1+e^{-2(t-s)}}{2} & \frac{1-e^{-2(t-s)}}{2} \\ \frac{1-e^{-2(t-s)}}{2} & \frac{1+e^{-2(t-s)}}{2} \end{bmatrix}^{\otimes d},$$

which implies

$$q^{\rightarrow}_{t|s}(\boldsymbol{y}|\boldsymbol{y}') = 2^{-d\log_2 K} \cdot \prod_{i=0}^{d\log_2 K - 1} \left[1 + (-1)^{|\boldsymbol{y}_i - \boldsymbol{y}'_i|} \cdot e^{-2(t-s)}\right] \quad \text{and} \quad \boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}.$$

Hence, the proof is completed. ■

Figure 3 visualizes the evolution of transition probabilities under different forward processes. The tridiagonal CTMC (second row) can be viewed as a discrete analogue of the normalized Gaussian transition (first row), where the domain $[0, 1]$ is quantized into 8 bins. The tridiagonal structure results in slow mixing, as transitions are restricted to immediate neighbors. At small time steps (e.g., $t = 0.01$, first column), the transition kernel satisfies $\boldsymbol{Q}^{\rightarrow}_{t+\Delta t|t} \approx \boldsymbol{I} + \Delta t \cdot \boldsymbol{R}^{\rightarrow}$, so the sparsity of the transition kernel closely reflects that of the rate matrix $\boldsymbol{R}^{\rightarrow}$. For efficient simulation of the reverse process, defined by $R^{\leftarrow}_t(\boldsymbol{y}, \boldsymbol{y}') := R^{\rightarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \frac{q^{\leftarrow}_t(\boldsymbol{y})}{q^{\leftarrow}_t(\boldsymbol{y}')}$ as Eq. (3), it is essential that $\boldsymbol{R}^{\rightarrow}$ remains sparse. While the dense forward process (third row) mixes rapidly, it incurs high computational cost per step when simulating the reverse process. In contrast, the hypercube structure (fourth row) achieves a favorable balance: it enables efficient long-range transitions for fast mixing while preserving an $\mathcal{O}(\log |\mathcal{Y}|)$ sparse structure for efficient implementation.

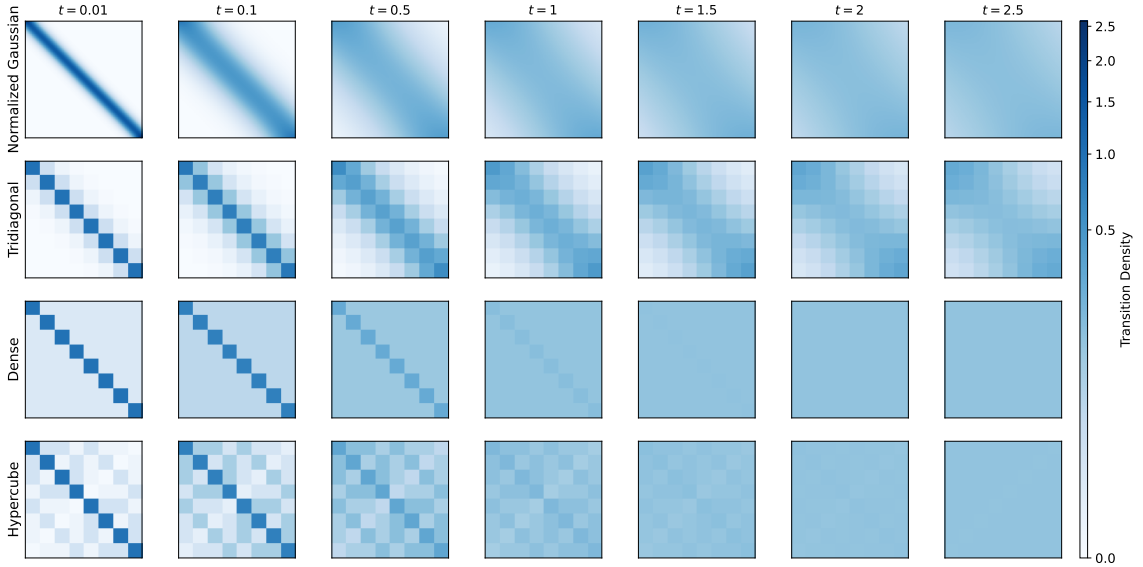

Figure 3: Heatmaps of the probability transition at different time steps t for four diffusion processes: a continuous normalized Gaussian kernel on $[0, 1]$ (top row), and discrete CTMCs over $|\mathcal{Y}| = 8$ states based on tridiagonal, dense, and hypercube transition rate matrices (bottom three rows).

B.2. Proof of Eq. (3)

Proof For any $t \in [0, T]$, the marginal, joint, and conditional distribution w.r.t. $\{\mathbf{y}^{\leftarrow}_t\}$ are denoted as

$$\mathbf{y}^{\leftarrow}_t \sim q^{\leftarrow}_t, \quad (\mathbf{y}^{\leftarrow}_t, \mathbf{y}^{\leftarrow}_{t'}) \sim q^{\leftarrow}_{t,t'}, \quad \text{and} \quad q^{\leftarrow}_{t'|t} = q^{\leftarrow}_{t',t}/q^{\leftarrow}_t,$$

which have $q_t^{\leftarrow} = q_{T-t}^{\rightarrow}$. Considering the reverse CTMC, we have

$$\frac{\mathrm{d}q_t^{\leftarrow}}{\mathrm{d}t}(\boldsymbol{y}) = \lim_{\Delta t \to 0} \frac{q_{t+\Delta t}^{\leftarrow}(\boldsymbol{y}) - q_t^{\leftarrow}(\boldsymbol{y})}{\Delta t} = \lim_{\Delta t \to 0} \frac{\sum_{\boldsymbol{y}'} q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}|\boldsymbol{y}') \cdot q_t^{\leftarrow}(\boldsymbol{y}') - q_t^{\leftarrow}(\boldsymbol{y})}{\Delta t}$$
$$= \lim_{\Delta t \to 0} \frac{\sum_{\boldsymbol{y}'} \left(q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}|\boldsymbol{y}') - \delta_{\boldsymbol{y}'}(\boldsymbol{y}) \right) \cdot q_t^{\leftarrow}(\boldsymbol{y}')}{\Delta t}.$$

Therefore, due to the following definition

$$R_t^{\leftarrow}(\boldsymbol{y}, \boldsymbol{y}') \coloneqq \lim_{\Delta t \to 0} \left[\frac{q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}|\boldsymbol{y}') - \delta_{\boldsymbol{y}'}(\boldsymbol{y})}{\Delta t} \right],$$

we have

$$\frac{\mathrm{d}\, q_t^{\leftarrow}}{\mathrm{d}\, t}(\boldsymbol{y}) \;=\; \langle R_t^{\leftarrow}(\boldsymbol{y}, \cdot),\, q_t^{\leftarrow}(\cdot) \rangle_{\mathcal{Y}}.$$

Then, we start to check the property of $\boldsymbol{R}_t^{\leftarrow}(\boldsymbol{y}, \boldsymbol{y}')$ when $\boldsymbol{y} \neq \boldsymbol{y}'$ in the following, which has

$$R_t^{\leftarrow}(\boldsymbol{y}, \boldsymbol{y}') = \lim_{\Delta t \to 0} \frac{q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}|\boldsymbol{y}')}{\Delta t} = \lim_{s \to t} \partial_t q_{t|s}^{\leftarrow}(\boldsymbol{y}|\boldsymbol{y}').$$

Specifically, with the Bayes Theorem, it has

$$\begin{aligned}
&\lim_{s \to t} \partial_t q_{t|s}^{\leftarrow}(\boldsymbol{y}|\boldsymbol{y}') = -1 \cdot \lim_{s \to t} \partial_{T-t} q_{T-t|T-s}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}') \\
&= -1 \cdot \lim_{s \to t} \partial_{T-t} \left[\frac{q_{T-s|T-t}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y}) \cdot q_{T-t}^{\rightarrow}(\boldsymbol{y})}{q_{T-s}^{\rightarrow}(\boldsymbol{y}')} \right] \\
&= \underbrace{- \lim_{s \to t} \partial_{T-t} q_{T-s|T-t}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y}) \cdot \frac{q_{T-t}^{\rightarrow}(\boldsymbol{y})}{q_{T-s}^{\rightarrow}(\boldsymbol{y}')}}_{\text{Term 1}} \underbrace{- \lim_{s \to t} \frac{q_{T-s|T-t}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y})}{q_{T-s}^{\rightarrow}(\boldsymbol{y}')} \cdot \partial_{T-t} q_{T-t}^{\rightarrow}(\boldsymbol{y})}_{\text{Term 2}}.
\end{aligned} \tag{26}$$

For Term 1 of Eq. (26), we have

$$\text{Term 1} = -R^{\rightarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \frac{q_{T-t}^{\rightarrow}(\boldsymbol{y})}{q_{T-t}^{\rightarrow}(\boldsymbol{y}')},$$

which follows from the Kolmogorov backward theorem (Lemma 19) and Eq. (19):

$$\partial_{T-t} q_{T-s|T-t}^{\rightarrow}(\boldsymbol{y}'|\boldsymbol{y}) = -\mathcal{L}^{\rightarrow}[q_{T-s|T-t}^{\rightarrow}(\boldsymbol{y}'|\cdot)](\boldsymbol{y}) = -\left\langle q_{T-s|T-t}^{\rightarrow}(\boldsymbol{y}'|\cdot), R^{\rightarrow}(\cdot, \boldsymbol{y}) \right\rangle_{\mathcal{Y}} = R^{\rightarrow}(\boldsymbol{y}', \boldsymbol{y}).$$

Here, the last equation follows from the fact

$$\lim_{s \to t} q_{T-s|T-t}^{\rightarrow}(\boldsymbol{y}'|\tilde{\boldsymbol{y}}) = 1 \quad \text{only when} \quad \boldsymbol{y}' = \tilde{\boldsymbol{y}}.$$

Additionally, with the fact $\boldsymbol{y}' \neq \boldsymbol{y}$, it has

$$\lim_{s \to t} q_{T-s|T-t}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}') = 0.$$

That means Term 2 of Eq. (26) will definitely be 0. Hence, the proof is completed. ■

B.3. Proof of Eq. (4)

Proof [Adapted from Proposition 1 of Campbell et al. (2022)] The RHS of Eq. (4) satisfies

$$\lim_{\Delta t\to 0}\left[\frac{q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}|\boldsymbol{y}')-\delta_{\boldsymbol{y}'}(\boldsymbol{y})}{\Delta t}\right]=\lim_{s\to t}\partial_t q^{\leftarrow}_{t|s}(\boldsymbol{y}|\boldsymbol{y}').$$

Besides, we have

$$\begin{aligned}&\lim_{s\to t}\partial_t q^{\leftarrow}_{t|s}(\boldsymbol{y}|\boldsymbol{y}')=\lim_{s\to t}\partial_t\left[q^{\rightarrow}_{T-s|T-t}(\boldsymbol{y}'|\boldsymbol{y})\cdot\frac{q^{\rightarrow}_{T-t}(\boldsymbol{y})}{q^{\rightarrow}_{T-s}(\boldsymbol{y}')}\right]\\&=\lim_{s\to t}\left[\partial_t(q^{\rightarrow}_{T-s|T-t}(\boldsymbol{y}'|\boldsymbol{y}))\cdot\frac{q^{\rightarrow}_{T-t}(\boldsymbol{y})}{q^{\rightarrow}_{T-s}(\boldsymbol{y}')}+q^{\rightarrow}_{T-s|T-t}(\boldsymbol{y}'|\boldsymbol{y})\cdot\frac{\partial_t q^{\rightarrow}_{T-t}(\boldsymbol{y})}{q^{\rightarrow}_{T-s}(\boldsymbol{y}')}\right].\end{aligned}$$

When $\boldsymbol{y}\neq\boldsymbol{y}'$, we have

$$\lim_{s\to t}q^{\rightarrow}_{T-s|T-t}(\boldsymbol{y}'|\boldsymbol{y})=0,$$

which implies

$$\lim_{s\to t}\partial_t q^{\leftarrow}_{t|s}(\boldsymbol{y}|\boldsymbol{y}')=\lim_{s\to t}\partial_t(q^{\rightarrow}_{T-s|T-t}(\boldsymbol{y}'|\boldsymbol{y}))\cdot\frac{q^{\rightarrow}_{T-t}(\boldsymbol{y})}{q^{\rightarrow}_{T-s}(\boldsymbol{y}')}=R^{\rightarrow}(\boldsymbol{y}',\boldsymbol{y})\cdot\frac{q^{\rightarrow}_{T-t}(\boldsymbol{y})}{q^{\rightarrow}_{T-t}(\boldsymbol{y}')}.$$

The last equation follows from the Kolmogorov backward theorem, i.e., Lemma 19 and Eq. (19)

$$\partial_{T-t}q^{\rightarrow}_{T-s|T-t}(\boldsymbol{y}'|\boldsymbol{y})=-\mathcal{L}^{\rightarrow}[q^{\rightarrow}_{T-s|T-t}(\boldsymbol{y}'|\cdot)](\boldsymbol{y})=-\left\langle q^{\rightarrow}_{T-s|T-t}(\boldsymbol{y}'|\cdot),R^{\rightarrow}(\cdot,\boldsymbol{y})\right\rangle_{\mathcal{Y}}=R^{\rightarrow}(\boldsymbol{y}',\boldsymbol{y}).$$

Combining with Eq. (3), we have

$$\lim_{\Delta t\to 0}\left[\frac{q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}|\boldsymbol{y}')-\delta_{\boldsymbol{y}'}(\boldsymbol{y})}{\Delta t}\right]=\lim_{s\to t}\partial_t q^{\leftarrow}_{t|s}(\boldsymbol{y}|\boldsymbol{y}')=R^{\rightarrow}(\boldsymbol{y}',\boldsymbol{y})\cdot\frac{q^{\rightarrow}_{T-t}(\boldsymbol{y})}{q^{\rightarrow}_{T-t}(\boldsymbol{y}')}=R^{\leftarrow}_t(\boldsymbol{y},\boldsymbol{y}')\quad(27)$$

when $\boldsymbol{y}'\neq\boldsymbol{y}$. Besides, we have

$$\begin{aligned}&\sum_{\boldsymbol{y}\in\mathcal{Y}}R^{\leftarrow}_t(\boldsymbol{y},\boldsymbol{y}')=\sum_{\boldsymbol{y}\in\mathcal{Y}}R^{\rightarrow}(\boldsymbol{y}',\boldsymbol{y})\cdot\frac{q^{\rightarrow}_{T-t}(\boldsymbol{y})}{q^{\rightarrow}_{T-t}(\boldsymbol{y}')}\\&=\sum_{\boldsymbol{y}\in\mathcal{Y}}\lim_{\Delta t\to 0}\left[\frac{q^{\rightarrow}_{T-t+\Delta t|T-t}(\boldsymbol{y}'|\boldsymbol{y})-\delta_{\boldsymbol{y}}(\boldsymbol{y}')}{\Delta t}\right]\cdot\frac{q^{\rightarrow}_{T-t}(\boldsymbol{y})}{q^{\rightarrow}_{T-t}(\boldsymbol{y}')}=\sum_{\boldsymbol{y}\in\mathcal{Y}}\lim_{\Delta t\to 0}\left[\frac{q^{\rightarrow}_{T-t+\Delta t|T-t}(\boldsymbol{y}|\boldsymbol{y}')-\delta_{\boldsymbol{y}'}(\boldsymbol{y})}{\Delta t}\right]=0,\end{aligned}$$

which means

$$R^{\leftarrow}_t(\boldsymbol{y}',\boldsymbol{y}')=-\sum_{\boldsymbol{y}\neq\boldsymbol{y}'}R^{\leftarrow}_t(\boldsymbol{y},\boldsymbol{y}')=\lim_{\Delta t\to 0}-\left[\frac{1-\sum_{\boldsymbol{y}\neq\boldsymbol{y}'}q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}|\boldsymbol{y}')}{\Delta t}\right],$$

where the last inequality follows from Eq. (27). Hence, the proof is completed. ■

Appendix C. Proof of Lemma 1

Lemma 10 *Suppose the data distribution p_* is σ sub-Gaussian, by choosing $L \geq \sigma \cdot \sqrt{2\ln(2d/\epsilon)}$, the TV distance between p_* and $\tilde{p}_*$ defined in Eq. (9) will be smaller than ϵ, i.e., $\mathrm{TV}\left(p_*, \tilde{p}_*\right) \leq \epsilon$.*

Proof When p_* satisfies σ sub-Gaussian properties, i.e.,

$$\mathbb{E}_{\mathbf{x}\sim p_*}\left[\exp\left(l\left\langle \mathbf{x}, \boldsymbol{u}\right\rangle\right)\right] \leq \exp\left(\frac{\sigma^2 l^2 \cdot \|\boldsymbol{u}\|^2}{2}\right).$$

By choosing $\boldsymbol{u} = \boldsymbol{e}_i$, we can easily found that each dimension of $\mathbf{x}$ will be σ sub-Gaussian, i.e.,

$$\mathbb{E}_{\mathbf{x}_i\sim p_{*,i}}\left[\exp\left(l\mathbf{x}_i\boldsymbol{u}_i\right)\right] \leq \exp\left(\frac{\sigma^2 l^2 \cdot \|\boldsymbol{u}_i\|^2}{2}\right).$$

According to the sub-Gaussian properties for each coordinate, we have

$$\mathbb{P}_i\left[|\mathbf{x}_i| \geq l\right] \leq 2\exp\left(-\frac{l^2}{2\sigma^2}\right).$$

With the union bound, we have

$$\mathbb{P}_*\left[\max_{1\leq i\leq d}|\mathbf{x}_i| > L\right] \leq \sum_{i=1}^{d}\mathbb{P}_*\left[|\mathbf{x}_i| > L\right] \leq 2d\cdot\exp\left(-\frac{L^2}{2\sigma^2}\right).$$

Under this condition, by supposing

$$2d\cdot\exp\left(-\frac{L^2}{2\sigma^2}\right) \leq \epsilon \quad \Leftrightarrow \quad L \geq \sigma\cdot\sqrt{2\ln\frac{2d}{\epsilon}}, \tag{28}$$

we have $\mathbb{P}_*\left[\max_{1\leq i\leq d}|\mathbf{x}_i| \geq L\right] \leq \epsilon$. Under this condition, the total variation distance between $\tilde{p}_*$ and p_* can be upper bounded by

$$\begin{aligned}
&\mathrm{TV}\left(p_*, \tilde{p}_*\right) = \frac{1}{2}\int_{\mathbb{R}^d}\left|p_*(\boldsymbol{x}) - \tilde{p}_*(\boldsymbol{x})\right|\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\int_{\boldsymbol{x}\in\mathrm{Cube}(L)}\left(\tilde{p}_*(\boldsymbol{x}) - p_*(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\boldsymbol{x}\notin\mathrm{Cube}(L)}p_*(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\left[1 - \int_{\boldsymbol{x}\in\mathrm{Cube}(L)}p_*(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right] + \frac{1}{2}\int_{\boldsymbol{x}\notin\mathrm{Cube}(L)}p_*(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int_{\boldsymbol{x}\notin\mathrm{Cube}(L)}p_*(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq \epsilon
\end{aligned} \tag{29}$$

where the last inequality follows from Eq. (28). Hence, the proof is completed. ■

Lemma 11 *Suppose the distribution $\tilde{p}_*$ defined in Eq. (9) satisfies H–smoothness, by choosing*

$$l \leq \left(2HL + \|\nabla f_*(\mathbf{0})\|\right)^{-1}\cdot d^{-1/2}\epsilon,$$

the TV distance satisfies $\mathrm{TV}\left(\tilde{p}_*, \overline{p}_*\right) \leq 2\epsilon$ where $\overline{p}_*$ is defined in Eq. (11).

Proof By Lagrange's mean value theorem, for each cell $\mathrm{Cell}\,(i_0, i_1, \ldots, i_{d-1})$, there exists a point $\bar{\boldsymbol{x}}_{i_0,i_1,\ldots,i_{d-1}} \in \mathrm{Cell}\,(i_0, i_1, \ldots, i_{d-1})$ such that

$$\tilde{p}_*(\bar{\boldsymbol{x}}_{i_0,i_1,\ldots,i_{d-1}}) = \frac{\int_{\boldsymbol{u}\in\mathrm{Cell}(i_0,i_1,\ldots,i_{d-1})} \tilde{p}_*(\boldsymbol{u})\mathrm{d}\boldsymbol{u}}{l^d}.$$

Therefor, the piecewise constant density $\overline{p}_*$ satisfies $\overline{p}_*(\boldsymbol{x}) = \tilde{p}_*(\bar{\boldsymbol{x}}_{i_0,i_1,\ldots,i_{d-1}})$, for any $\boldsymbol{x} \in \mathrm{Cell}\,(i_0, i_1, \ldots, i_{d-1})$.

We now aim to bound the difference $|\tilde{p}_*(\boldsymbol{u}) - \tilde{p}_*(\boldsymbol{x})|$ for any $\boldsymbol{u}, \boldsymbol{x} \in \mathrm{Cell}\,(i_0, i_1, \ldots, i_{d-1})$, using H–smoothness. Later, we will choose $\boldsymbol{u} = \bar{\boldsymbol{x}}_{i_0,i_1,\ldots,i_{d-1}}$ to bound the total variation distance between $\tilde{p}_*$ and $\overline{p}_*$.

According to the construction of $\tilde{p}_*$, i.e., Eq. (9), we have

$$\frac{\tilde{p}_*(\boldsymbol{u})}{\tilde{p}_*(\boldsymbol{x})} = \frac{p_*(\boldsymbol{u})}{p_*(\boldsymbol{x})} = \exp\left(f_*(\boldsymbol{x}) - f_*(\boldsymbol{u})\right). \tag{30}$$

With H–smoothness, i.e., $\left\|\nabla^2 f_*\right\| \le H$, we have

$$\begin{aligned} f_*(\boldsymbol{x}) - f_*(\boldsymbol{u}) &\le \nabla f_*(\boldsymbol{u}) \cdot (\boldsymbol{x} - \boldsymbol{u}) + \frac{H}{2} \cdot \|\boldsymbol{u} - \boldsymbol{x}\|^2 \\ &\le \|\nabla f_*(\boldsymbol{u})\| \cdot \|\boldsymbol{x} - \boldsymbol{u}\| + \frac{H}{2} \cdot \|\boldsymbol{x} - \boldsymbol{u}\|^2. \end{aligned} \tag{31}$$

Since $\boldsymbol{u}, \boldsymbol{x} \in \mathrm{Cell}\,(i_0, i_1, \ldots, i_{d-1})$, and each cell is an axis-aligned hypercube of side length l, we have

$$\|\boldsymbol{x} - \boldsymbol{u}\|^2 = \sum_{i=1}^{d} \|\boldsymbol{x}_i - \boldsymbol{u}_i\|^2 \le dl^2.$$

Let $G_0 \coloneqq \|\nabla f_*(\boldsymbol{0})\|$. Then we have

$$\|\nabla f_*(\boldsymbol{u})\| \le \|\nabla f_*(\boldsymbol{u}) - \nabla f_*(\boldsymbol{0})\| + G_0 \le H \cdot 2L + G_0,$$

where the last inequality follows from $\boldsymbol{u} \in \mathrm{Cube}\,(L)$. Therefore, by requiring

$$l \le \frac{\epsilon}{\sqrt{d} \cdot (2HL + G_0)},$$

and $\epsilon \le 8HL^2$ without loss of generality, we will have $l \le \sqrt{2\epsilon/(dH)}$, which means

$$\|\nabla f_*(\boldsymbol{u})\| \cdot \|\boldsymbol{x} - \boldsymbol{u}\| + \frac{H}{2} \cdot \|\boldsymbol{x} - \boldsymbol{u}\|^2 \le (2HL + G_0) \cdot \sqrt{d}l + \frac{H}{2} \cdot dl^2 \le 2\epsilon. \tag{32}$$

Plugging Eq. (31) and Eq. (32) into Eq. (30), we have

$$\frac{\tilde{p}_*(\boldsymbol{u})}{\tilde{p}_*(\boldsymbol{x})} \le \exp(2\epsilon) \le (1 + 4\epsilon). \tag{33}$$

With a similar technique, we have

$$-(f_*(\boldsymbol{x}) - f_*(\boldsymbol{u})) = f_*(\boldsymbol{u}) - f_*(\boldsymbol{x}) \le \|\nabla f_*(\boldsymbol{x})\| \cdot \|\boldsymbol{x} - \boldsymbol{u}\| + \frac{H}{2} \cdot \|\boldsymbol{x} - \boldsymbol{u}\|^2.$$

Under the same setting, it implies

$$\frac{\tilde{p}_*(\boldsymbol{x})}{\tilde{p}_*(\boldsymbol{u})} \leq \exp(2\epsilon) \quad \Leftrightarrow \quad \frac{\tilde{p}_*(\boldsymbol{u})}{\tilde{p}_*(\boldsymbol{x})} \geq \exp(-2\epsilon) \geq 1 - 2\epsilon. \tag{34}$$

Combining Eq. (33) with Eq. (34), we have

$$1 - 4\epsilon \leq \frac{\tilde{p}_*(\boldsymbol{u})}{\tilde{p}_*(\boldsymbol{x})} \leq 1 + 4\epsilon. \tag{35}$$

Hence we are able to control the TV distance between $\tilde{p}_*$ and $\overline{p}_*$, i.e.,

$$\begin{aligned}
&\mathrm{TV}\left(\overline{p}_*, \tilde{p}_*\right) = \frac{1}{2}\int_{\boldsymbol{x}\in\mathrm{Cube}(L)} \left|\overline{p}_*(\boldsymbol{x}) - \tilde{p}_*(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\sum_{i_0,i_1,\ldots,i_{d-1}} \int_{\boldsymbol{x}\in\mathrm{Cell}(i_0,i_1,\ldots,i_{d-1})} \left|\overline{p}_*(\boldsymbol{x}) - \tilde{p}_*(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\sum_{i_0,i_1,\ldots,i_{d-1}} \int_{\boldsymbol{x}\in\mathrm{Cell}(i_0,i_1,\ldots,i_{d-1})} \left|\tilde{p}_*(\bar{\boldsymbol{x}}_{i_0,i_1,\ldots,i_{d-1}}) - \tilde{p}_*(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\sum_{i_0,i_1,\ldots,i_{d-1}} \int_{\boldsymbol{x}\in\mathrm{Cell}(i_0,i_1,\ldots,i_{d-1})} \tilde{p}_*(\boldsymbol{x}) \left|\frac{\tilde{p}_*(\bar{\boldsymbol{x}}_{i_0,i_1,\ldots,i_{d-1}})}{\tilde{p}_*(\boldsymbol{x})} - 1\right| \mathrm{d}\boldsymbol{x} \\
&\leq \frac{1}{2}\sum_{i_0,i_1,\ldots,i_{d-1}} \int_{\boldsymbol{x}\in\mathrm{Cell}(i_0,i_1,\ldots,i_{d-1})} \tilde{p}_*(\boldsymbol{x}) 4\epsilon \mathrm{d}\boldsymbol{x} \\
&= 2\epsilon,
\end{aligned} \tag{36}$$

where the last inequality follows from Eq. (35). Hence, the proof is completed. ■

Lemma 12 *Suppose the data distribution p_* satisfy Assumption [A2]–[A3], we have*

$$\left\|\nabla f_*(\mathbf{0})\right\|^2 \leq 2Hd + 2H^2 m_0$$

Proof We start with the following inequality

$$\begin{aligned}
&\left\|\nabla f_*(\mathbf{0})\right\|^2 = \int_{\boldsymbol{x}\in\mathbb{R}^d} p_*(\boldsymbol{x}) \left\|\nabla f_*(\mathbf{0})\right\|^2 \mathrm{d}\boldsymbol{x} \\
&\leq 2\int_{\boldsymbol{x}\in\mathbb{R}^d} p_*(\boldsymbol{x}) \left\|\nabla f_*(\boldsymbol{x})\right\|^2 \mathrm{d}\boldsymbol{x} + 2\int_{\boldsymbol{x}\in\mathbb{R}^d} p_*(\boldsymbol{x}) \left\|\nabla f_*(\mathbf{0}) - \nabla f_*(\boldsymbol{x})\right\|^2 \mathrm{d}\boldsymbol{x} \\
&\leq 2Hd + 2H^2 \int_{\boldsymbol{x}\in\mathbb{R}^d} p_*(\boldsymbol{x}) \left\|\boldsymbol{x}\right\|^2 \mathrm{d}\boldsymbol{x} = 2Hd + 2H^2 m_0
\end{aligned}$$

where the second inequality follows from Lemma 20 and Assumption **[A3]** and the last inequality follows from Assumption **[A2]**. Hence, the proof is completed. ■

Proof [Proof of Lemma 1] The TV distance between the original data distribution p_* and the histogram approximation $\overline{p}_*$ can be written as

$$\mathrm{TV}\left(p_*, \overline{p}_*\right) \leq \mathrm{TV}\left(p_*, \tilde{p}_*\right) + \mathrm{TV}\left(\tilde{p}_*, \overline{p}_*\right).$$

Following from Lemma 10, we will have $\mathrm{TV}\left(p_*, \tilde{p}_*\right)$ by choosing

$$L \geq \sigma \cdot \sqrt{2\ln(2d/\epsilon)}. \tag{37}$$

Moreover, with the quantization shown in Eq. (11), it has $\mathrm{TV}\left(\tilde{p}_*, \overline{p}_*\right) \leq 2\epsilon$ by choosing

$$l \leq \left(2HL + \|\nabla f_*(\mathbf{0})\|\right)^{-1} \cdot d^{-1/2}\epsilon, \tag{38}$$

which follows from Lemma 11. Combining Eq. (37) with Eq. (38), if we set

$$L = \sigma \cdot \sqrt{2\ln(2d/\epsilon)} \quad \text{and} \quad l \coloneqq \frac{\epsilon}{2H(L\sqrt{d} + d + \sqrt{dm_0})}$$

and l satisfies

$$\begin{aligned} l &\leq \frac{\epsilon}{\left(2HL + 2\sqrt{Hd} + 2H\sqrt{m_0}\right)\sqrt{d}} \leq \frac{\epsilon}{\left(2HL + \sqrt{2Hd + 2H^2 m_0}\right)\sqrt{d}} \\ &\leq \left(2HL + \|\nabla f_*(\mathbf{0})\|\right)^{-1} \cdot d^{-1/2}\epsilon \end{aligned}$$

where the last inequality follows from Lemma 12. That means

$$l = \Omega\left(\left[2H \cdot \left(\sigma\sqrt{2d\ln(2d/\epsilon)} + d + \sqrt{dm_0}\right)\right]^{-1} \cdot \epsilon\right),$$

it will have $\mathrm{TV}\left(p_*, \overline{p}_*\right) \leq 3\epsilon$. Hence, the proof is completed. ■

Appendix D. Proof of Lemma 2

To make our analysis clear, we define the variables, the random variables, and the marginal density derived by a specific ordered set $S \subseteq \{0, 1, \ldots, d\log_2 K - 1\}$. Specifically, we have

$$\boldsymbol{y}_S = \sum_{i=0}^{|S|-1} \boldsymbol{e}_i \cdot y_{S_i} \quad \text{and} \quad \mathbf{y}_{t,S} = \sum_{i=0}^{|S|-1} \boldsymbol{e}_i \cdot \mathrm{y}_{t,S_i}$$

where there are

$$\boldsymbol{y} = [y_0, y_1, \ldots, y_{d\log_2 K - 1}] \quad \text{and} \quad \mathbf{y}_t = [\mathrm{y}_{t,0}, \mathrm{y}_{t,1}, \ldots, \mathrm{y}_{t,d\log_2 K - 1}].$$

Suppose $\mathbf{y}_t \sim q_t$ The underlying distribution of $\mathbf{y}_{t,S}$ is denoted as

$$q_{t,S}(\boldsymbol{y}_S) = \sum_{\tilde{\boldsymbol{y}} \in \mathcal{Y}} q_t(\tilde{\boldsymbol{y}}) \cdot \mathbf{1}_{\boldsymbol{y}_S}(\tilde{\boldsymbol{y}}_S).$$

Lemma 13 (Modified log-Sobolev inequality for the forward process) *Suppose the transition rate function $R^{\rightarrow}$ of the CTMC $\{\mathbf{y}_t^{\rightarrow}\}_{t=0}^{T}$ be defined as Eq. (13). CTMC satisfies modified log-Sobolev inequality with a constant 2, that is to say, for any $f \in \mathbb{L}_2(q_\infty^{\rightarrow})$, it has*

$$\mathrm{Ent}_{q_\infty^{\rightarrow}}[f] \leq \mathcal{E}(f, \ln f)$$

where Ent *and $\mathcal{E}$ denote the entropy and the Dirichlet functional.*

Proof We start from the setting of the transition rate matrix of the forward process shown in Eq. (13). Combining with the Eq. (19), the infinitesimal generator for the forward process can be obtained, i.e.,

$$\mathcal{L}^{\rightarrow}[f](\boldsymbol{y}) = \langle f, R^{\rightarrow}(\cdot, \boldsymbol{y})\rangle_{\mathcal{Y}}\,. \tag{39}$$

To verify the modified log-Sobolev inequality, we first require to calculate the Dirichlet functional $\mathcal{E}(f, \ln f)$. Here $\mathcal{E}$ denotes the Dirichlet functional

$$\mathcal{E}(f, g) \coloneqq \int \Gamma(f, g)\mathrm{d}q_{\infty}^{\rightarrow},$$

where q_{∞} denotes the invariant measure of this forward process and Γ denotes the carré du champ operator, i.e.,

$$\Gamma(f, g) \coloneqq \frac{1}{2}\left(\mathcal{L}[f \cdot g] - f \cdot \mathcal{L}[g] - g \cdot \mathcal{L}[f]\right).$$

Specifically, presenting the transition rate matrix to be a matrix version Eq. (25), we have

$$\mathrm{d}\boldsymbol{q}_t^{\rightarrow}/\mathrm{d}t = \boldsymbol{R}^{\rightarrow} \cdot \boldsymbol{q}_t^{\rightarrow}$$

Combining the fact $\mathbf{1} \cdot \boldsymbol{R} = \mathbf{0}$ and $\boldsymbol{R}$ is symmetric, the RHS of the above equation satisfies

$$\boldsymbol{R}^{\rightarrow} \cdot 2^{-d\log_2 K} \cdot \mathbf{1} = \mathbf{0},$$

which implies the uniform distribution coincides with the invariant measure of $q_{\infty}^{\rightarrow}$. Then, for the Dirichlet functional, it has

$$\begin{aligned}
\mathcal{E}(f, \ln f) &= \frac{1}{2}\int \mathcal{L}[f \cdot \ln f](\boldsymbol{y}) - f(\boldsymbol{y}) \cdot \mathcal{L}[\ln f](\boldsymbol{y}) - \ln f(\boldsymbol{y}) \cdot \mathcal{L}[f](\boldsymbol{y})\mathrm{d}q_{\infty}(\boldsymbol{y}) \\
&= \frac{1}{2}\sum_{\boldsymbol{y}\in\mathcal{Y}} q_{\infty}(\boldsymbol{y}) \cdot \Bigg[\sum_{\boldsymbol{y}'\in\mathcal{Y}} f(\boldsymbol{y}')\ln f(\boldsymbol{y}') \cdot R(\boldsymbol{y}', \boldsymbol{y}) - f(\boldsymbol{y}) \cdot \sum_{\boldsymbol{y}'\in\mathcal{Y}} \ln f(\boldsymbol{y}') \cdot R(\boldsymbol{y}', \boldsymbol{y}) \\
&\qquad - \ln f(\boldsymbol{y}) \cdot \sum_{\boldsymbol{y}'\in\mathcal{Y}} f(\boldsymbol{y}') \cdot R(\boldsymbol{y}', \boldsymbol{y}) + f(\boldsymbol{y}) \cdot \ln f(\boldsymbol{y}) \cdot \underbrace{\sum_{\boldsymbol{y}'\in\mathcal{Y}} R(\boldsymbol{y}', \boldsymbol{y})}_{=0}\Bigg] \\
&= \frac{1}{2}\sum_{\boldsymbol{y}\in\mathcal{Y}}\sum_{\boldsymbol{y}'\in\mathcal{Y}} q_{\infty}(\boldsymbol{y})\left(f(\boldsymbol{y}) - f(\boldsymbol{y}')\right) \cdot R(\boldsymbol{y}', \boldsymbol{y}) \cdot (\ln f(\boldsymbol{y}) - \ln f(\boldsymbol{y}')).
\end{aligned}$$

Plugging the definition of R into the above equation, we have

$$\begin{aligned}
\mathcal{E}(f, \ln f) =& \frac{1}{2} \cdot \sum_{\boldsymbol{y}\in\mathcal{Y}} q_{\infty}(\boldsymbol{y}) \cdot \sum_{i=0}^{d\log_2 K - 1} \sum_{\tilde{y}_i \in \{0,1\}} (f(\boldsymbol{y}) - f(\boldsymbol{y} + (\tilde{y}_i - y_i) \cdot \boldsymbol{e}_i)) \\
&\cdot (\ln f(\boldsymbol{y}) - \ln f(\boldsymbol{y} + (\tilde{y}_i - y_i) \cdot \boldsymbol{e}_i))\,.
\end{aligned} \tag{40}$$

Then, we consider $\mathrm{Ent}_{q_\infty^{\rightarrow}}[f]$, which satisfies

$$\begin{aligned}
&\mathrm{Ent}_{q_\infty^{\rightarrow}}[f] = \mathbb{E}_{\mathbf{y}\sim q_\infty^{\rightarrow}}\left[f(\mathbf{y})\ln f(\mathbf{y})\right] - \mathbb{E}_{\mathbf{y}\sim q_\infty^{\rightarrow}}\left[f(\mathbf{y})\right]\ln\left(\mathbb{E}_{\mathbf{y}\sim q_\infty^{\rightarrow}}\left[f(\mathbf{y})\right]\right)\\
&\leq \sum_{i=0}^{d\log_2 K-1} \mathbb{E}_{\mathbf{y}_{[0:i-1,i+1:d\log_2 K-1]}}\left[\underbrace{\mathbb{E}_{\mathrm{y}_i}\left[f(\mathbf{y})\ln f(\mathbf{y})\right] - \mathbb{E}_{\mathrm{y}_i}\left[f(\mathbf{y})\right]\ln\left(\mathbb{E}_{\mathrm{y}_i}\left[f(\mathbf{y})\right]\right)}_{\text{Term 1}}\right].
\end{aligned} \tag{41}$$

due to the sub-additivity of the entropy, i.e., Lemma 17. Term 1 of Eq. (41) satisfies

$$\begin{aligned}
\text{Term 1} =& \sum_{y_i\in\{0,1\}} q_{\infty,i}^{\rightarrow}(y_i)\cdot f(\boldsymbol{y}_{0:i-1},y_i,\boldsymbol{y}_{i+1:d\log_2 K-1})\ln f(\boldsymbol{y}_{0:i-1},y_i,\boldsymbol{y}_{i+1:d\log_2 K-1})\\
&- \sum_{y_i\in\{0,1\}} q_{\infty,i}^{\rightarrow}(\boldsymbol{y}_i) f(\boldsymbol{y}_{0:i-1},y_i,\boldsymbol{y}_{i+1:d\log_2 K-1})\\
&\qquad\cdot\ln\left(\sum_{\tilde{y}_i\in\{0,1\}} q_{\infty,i}^{\rightarrow}(\tilde{y}_i) f(\boldsymbol{y}_{0:i-1},\tilde{y}_i,\boldsymbol{y}_{i+1:d\log_2 K-1})\right)\\
\leq& \sum_{y_i\in\{0,1\}} q_{\infty,i}^{\rightarrow}(y_i)\cdot f(\boldsymbol{y}_{0:i-1},y_i,\boldsymbol{y}_{i+1:d\log_2 K-1})\\
&\qquad\cdot\sum_{\tilde{y}_i\in\{0,1\}}\left[\frac{\ln f(\boldsymbol{y}_{0:i-1},y_i,\boldsymbol{y}_{i+1:d\log_2 K-1})}{2} - \frac{\ln f(\boldsymbol{y}_{0:i-1},\tilde{y}_i,\boldsymbol{y}_{i+1:d\log_2 K-1})}{2}\right]\\
\leq& \frac{1}{2}\sum_{y_i,\tilde{y}_i\in\{0,1\}} q_{\infty,i}^{\rightarrow}(y_i)\cdot\left(f(\boldsymbol{y}_{0:i-1},y_i,\boldsymbol{y}_{i+1:d\log_2 K-1}) - f(\boldsymbol{y}_{0:i-1},\tilde{y}_i,\boldsymbol{y}_{i+1:d\log_2 K-1})\right)\\
&\qquad\cdot\left(\ln f(\boldsymbol{y}_{0:i-1},y_i,\boldsymbol{y}_{i+1:d\log_2 K-1}) - \ln f(\boldsymbol{y}_{0:i-1},\tilde{y}_i,\boldsymbol{y}_{i+1:d\log_2 K-1})\right),
\end{aligned}$$

where the first inequality follows from the concavity of the logarithm function, and the last inequality follows from

$$\begin{aligned}
&\sum_{\boldsymbol{y},\tilde{\boldsymbol{y}}}\cdot f(\boldsymbol{y})\cdot(\ln f(\boldsymbol{y}) - \ln f(\tilde{\boldsymbol{y}})) = \sum_{\tilde{\boldsymbol{y}}} f(\tilde{\boldsymbol{y}})\cdot(\ln f(\tilde{\boldsymbol{y}}) - \ln f(\boldsymbol{y}))\\
&= \frac{1}{2}\sum_{\boldsymbol{y},\tilde{\boldsymbol{y}}}\cdot(f(\boldsymbol{y}) - f(\tilde{\boldsymbol{y}}))\cdot(\ln f(\boldsymbol{y}) - \ln f(\tilde{\boldsymbol{y}}))
\end{aligned}$$

and $q_\infty^{\rightarrow}(\cdot)$ is a constant function. Then, plugging this inequality into Eq. (41), we have

$$\begin{aligned}
\mathrm{Ent}_{q_\infty^{\rightarrow}}[f] \leq& \frac{1}{2}\cdot\sum_{i=0}^{d\log_2 K-1}\left[\sum_{\boldsymbol{y}_{[0:i-1,i+1:d\log_2 K-1]}} q_{\infty,[0:i-1,i+1:d\log_2 K-1]}^{\rightarrow}(\mathbf{y}_{[0:i-1,i+1:d\log_2 K-1]})\right.\\
&\sum_{y_i} q_{\infty,i}^{\rightarrow}(y_i)\sum_{\tilde{\boldsymbol{y}}_i}\left(f(\boldsymbol{y}_{0:i-1},y_i,\boldsymbol{y}_{i+1:d\log_2 K-1}) - f(\boldsymbol{y}_{0:i-1},\tilde{y}_i,\boldsymbol{y}_{i+1:d\log_2 K-1})\right)\\
&\left.\cdot\left(\ln f(\boldsymbol{y}_{0:i-1},y_i,\boldsymbol{y}_{i+1:d\log_2 K-1}) - \ln f(\boldsymbol{y}_{0:i-1},\tilde{y}_i,\boldsymbol{y}_{i+1:d\log_2 K-1})\right)\right]\\
=& \frac{1}{2}\cdot\sum_{\boldsymbol{y}} q_\infty^{\rightarrow}(\boldsymbol{y})\cdot\sum_{i=0}^{d\log_2 K-1}\sum_{\tilde{\boldsymbol{y}}_i}(f(\boldsymbol{y}) - f(\boldsymbol{y}+(\tilde{y}_i-y_i)\cdot\boldsymbol{e}_i))\\
&\cdot(\ln f(\boldsymbol{y}) - \ln f(\boldsymbol{y}+(\tilde{y}_i-y_i)\cdot\boldsymbol{e}_i))
\end{aligned} \tag{42}$$

Comparing Eq. (42) and Eq. (40), it satisfies

$$\mathrm{Ent}_{q_\infty^{\rightarrow}}[f] \leq \frac{C_{\mathrm{LSI}}}{2} \cdot \mathcal{E}(f, \ln f)$$

by choosing $C_{\mathrm{LSI}} = 2$. ■

Proof [Proof of Lemma 2] We investigate the dynamic of KL divergence between $q_t^{\rightarrow}$ and $q_\infty^{\rightarrow}$ in the forward process. Specifically, we have

$$\begin{aligned}
&\frac{\mathrm{dKL}\left(q_t^{\rightarrow} \middle\| q_\infty^{\rightarrow}\right)}{\mathrm{d}t} = \sum_{\boldsymbol{y} \in \mathcal{Y}} \frac{\mathrm{d}q_t^{\rightarrow}(\boldsymbol{y})}{\mathrm{d}t} \cdot \ln \frac{q_t^{\rightarrow}(\boldsymbol{y})}{q_\infty^{\rightarrow}(\boldsymbol{y})} = \sum_{\boldsymbol{y} \in \mathcal{Y}} \ln \frac{q_t^{\rightarrow}(\boldsymbol{y})}{q_\infty^{\rightarrow}(\boldsymbol{y})} \left(\sum_{\boldsymbol{y}_0 \in \mathcal{Y}} R(\boldsymbol{y}, \boldsymbol{y}_0) \cdot q_t^{\rightarrow}(\boldsymbol{y}_0) \right) \\
&= \sum_{\boldsymbol{y}_0} q_\infty^{\rightarrow}(\boldsymbol{y}_0) \cdot \frac{q_t^{\rightarrow}(\boldsymbol{y}_0)}{q_\infty^{\rightarrow}(\boldsymbol{y}_0)} \cdot \sum_{\boldsymbol{y}} \ln \frac{q_t^{\rightarrow}(\boldsymbol{y})}{q_\infty^{\rightarrow}(\boldsymbol{y})} \cdot R^{\rightarrow}(\boldsymbol{y}, \boldsymbol{y}_0) \\
&= \sum_{\boldsymbol{y}_0} q_\infty^{\rightarrow}(\boldsymbol{y}_0) \cdot \frac{q_t^{\rightarrow}(\boldsymbol{y}_0)}{q_\infty^{\rightarrow}(\boldsymbol{y}_0)} \cdot \mathcal{L}[\ln \frac{q_t^{\rightarrow}}{q_\infty^{\rightarrow}}](\boldsymbol{y}') = -\mathcal{E}\left(\frac{q_t^{\rightarrow}}{q_\infty^{\rightarrow}}, \ln \frac{q_t^{\rightarrow}}{q_\infty^{\rightarrow}} \right)
\end{aligned}$$

Due to Lemma 13, we have

$$\frac{\mathrm{dKL}\left(q_t^{\rightarrow} \middle\| q_\infty^{\rightarrow}\right)}{\mathrm{d}t} = -\mathcal{E}\left(\frac{q_t^{\rightarrow}}{q_\infty^{\rightarrow}}, \ln \frac{q_t^{\rightarrow}}{q_\infty^{\rightarrow}} \right) \leq \mathrm{Ent}_{q_\infty}\left[\frac{q_t^{\rightarrow}}{q_\infty^{\rightarrow}} \right] = -\mathrm{KL}\left(q_t^{\rightarrow} \middle\| q_\infty^{\rightarrow}\right).$$

According to the Gronwall's theorem, we have

$$\mathrm{KL}\left(q_t^{\rightarrow} \middle\| q_\infty^{\rightarrow}\right) \leq e^{-t} \cdot \mathrm{KL}\left(q_0^{\rightarrow} \middle\| q_\infty^{\rightarrow}\right).$$

Combining with the following initialization error bound,

$$\mathrm{KL}\left(q_0^{\rightarrow} \middle\| q_\infty^{\rightarrow}\right) = \sum_{\boldsymbol{y} \in \mathcal{Y}} q_0^{\rightarrow}(\boldsymbol{y}) \ln \frac{q_0^{\rightarrow}(\boldsymbol{y})}{2^{-d \log_2 K}} \leq d \log_2 K.$$

Hence, the proof is completed. ■

Appendix E. Supplementary Proofs for the Discrete Reverse Process

E.1. Proof of Lemma 3

Proof [Proof of Lemma 3 (adapted from Proposition 5 of Chen and Ying (2024))] Suppose the transition rate function $R^{\rightarrow}$ of the CTMC $\{\mathbf{y}_t^{\rightarrow}\}_{t=0}^T$ be defined as Eq. (13), the marginal distribution at time t can be written as

$$q_t^{\rightarrow}(\boldsymbol{y}) = \sum_{\boldsymbol{y}_0 \in \mathcal{Y}} q_0^{\rightarrow}(\boldsymbol{y}_0) \cdot q_{t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}_0).$$

Define the plus operator as follows

$$\boldsymbol{y} \oplus \boldsymbol{e}_i = \left[y_0, y_1, \ldots, y_{i-1}, (y_i + 1) \bmod 2, y_{i+1}, \ldots, y_{d \log_2 K - 1}\right],$$

then we have

$$\frac{q_t^{\rightarrow}(\boldsymbol{y}\oplus\boldsymbol{e}_i)}{q_t^{\rightarrow}(\boldsymbol{y})}=\frac{\sum_{\boldsymbol{y}_0\in\mathcal{Y}}q_0^{\rightarrow}(\boldsymbol{y}_0)\cdot q_{t|0}^{\rightarrow}(\boldsymbol{y}+\boldsymbol{e}_i|\boldsymbol{y}_0)}{\sum_{\boldsymbol{y}_0\in\mathcal{Y}}q_0^{\rightarrow}(\boldsymbol{y}_0)\cdot q_{t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}_0)}=\frac{\sum_{\boldsymbol{y}_0\in\mathcal{Y}}q_0^{\rightarrow}(\boldsymbol{y}_0)\cdot q_{t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}_0)\cdot\frac{q_{t|0}^{\rightarrow}(\boldsymbol{y}+\boldsymbol{e}_i|\boldsymbol{y}_0)}{q_{t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}_0)}}{\sum_{\boldsymbol{y}_0\in\mathcal{Y}}q_0^{\rightarrow}(\boldsymbol{y}_0)\cdot q_{t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}_0)}.$$

According to Bayes Theorem, we have

$$q_{0|t}^{\rightarrow}(\boldsymbol{y}_0|\boldsymbol{y})\cdot q_t^{\rightarrow}(\boldsymbol{y})=q_{t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}_0)\cdot q_0^{\rightarrow}(\boldsymbol{y}_0)\quad\Leftrightarrow\quad q_{0|t}^{\rightarrow}(\boldsymbol{y}_0|\boldsymbol{y})\propto q_{t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}_0)\cdot q_0^{\rightarrow}(\boldsymbol{y}_0),$$

which implies

$$\frac{q_t^{\rightarrow}(\boldsymbol{y}\oplus\boldsymbol{e}_i)}{q_t^{\rightarrow}(\boldsymbol{y})}=\mathbb{E}_{\mathbf{y}_0\sim q_{0|t}^{\rightarrow}(\cdot|\boldsymbol{y})}\left[\frac{q_{t|0}^{\rightarrow}(\boldsymbol{y}+\boldsymbol{e}_i|\boldsymbol{y}_0)}{q_{t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}_0)}\right].$$

With Lemma 8, we have

$$\frac{q_{t|0}^{\rightarrow}(\boldsymbol{y}+\boldsymbol{e}_i|\boldsymbol{y}_0)}{q_{t|0}^{\rightarrow}(\boldsymbol{y}|\boldsymbol{y}_0)}=\frac{1+(-1)^{|(y_i+1-y_{0,i})\bmod 2|}\cdot e^{-2t}}{1+(-1)^{|(y_i-y_{0,i})\bmod 2|}\cdot e^{-2t}}\leq\frac{1+e^{-2t}}{1-e^{-2t}},$$

which means

$$\frac{q_t^{\rightarrow}(\boldsymbol{y}\oplus\boldsymbol{e}_i)}{q_t^{\rightarrow}(\boldsymbol{y})}\leq\frac{1+e^{-2t}}{1-e^{-2t}}\leq 1+t^{-1}.$$

Therefore, if we consider the transition rate matrix of the reverse process, i.e.,

$$R_t^{\leftarrow}(\boldsymbol{y}',\boldsymbol{y})\coloneqq R^{\rightarrow}(\boldsymbol{y},\boldsymbol{y}')\cdot\frac{q_t^{\leftarrow}(\boldsymbol{y}')}{q_t^{\leftarrow}(\boldsymbol{y})}$$

provided by Eq (3), it has

$$\sum_{\boldsymbol{y}'\neq\boldsymbol{y}}R_t^{\leftarrow}(\boldsymbol{y}',\boldsymbol{y})=\sum_{i=0}^{d\log_2 K-1}\frac{q_t^{\leftarrow}(\boldsymbol{y}\oplus\boldsymbol{e}_i)}{q_t^{\leftarrow}(\boldsymbol{y})}=\sum_{i=0}^{d\log_2 K-1}\frac{q_{T-t}^{\rightarrow}(\boldsymbol{y}\oplus\boldsymbol{e}_i)}{q_{T-t}^{\rightarrow}(\boldsymbol{y})}\leq(d\log_2 K)\cdot(1+(T-t)^{-1}).$$

Hence, the proof is completed. ■

E.2. Proof of Theorem 6

The ultimate target of Alg. 2 is to generate sample $\hat{\mathbf{x}}$ and require its underlying distribution $\hat{p}$ to be close to the continuous data distribution p_*. However, Alg. 2 can be divided into two parts:

1. **Truncated Uniformization**: Generate a discrete sample following $\hat{q}_{T-\delta}=\hat{q}_{t_W}$ which approximates q_*, which is from Step. 2 to Step. 10.

2. Mapping the generated discrete data to the corresponding cell in Euclidean space and uniformly drawing a sample from the cell, which is from Step. 12

All the following notations correspond to those mentioned in Alg. 2.

Lemma 14 *Suppose we have a timestamp sequence satisfying*

$$t_0 = 0 \quad \textit{and} \quad t_{w+1} - t_w = 0.5 \cdot (T - t_{w+1}),$$

then we know the sequence $\{t_w\}_{w=0}^{W}$ *is strict increasing and* $t_W < T$ *for any* W.

Proof According to the timestamp setting, i.e.,

$$t_0 = 0 \quad \text{and} \quad t_{w+1} - t_w = 0.5 \cdot (T - t_{w+1}),$$

solve for t_{k+1}, we have

$$t_{w+1} = \frac{0.5T + t_w}{1.5} = \frac{T + 2t_w}{3}.$$

Then, we consider the difference:

$$t_{w+1} - t_w = \frac{T + 2\,t_w}{3} - t_w = \frac{T + 2t_w - 3t_w}{3} = \frac{T - t_w}{3}.$$

If $T - t_w > 0$, then we have

$$t_{w+1} - t_w = \frac{T - t_w}{3} > 0,$$

which shows $t_{w+1} > t_w$. Thus, as long as $t_w < T$, the sequence is strictly increasing.

Moreover, due to the fact $t_0 = 0 < T$, we can prove that $t_w < T$ for all w. Specifically, assume $t_w < T$; then

$$t_{w+1} = \frac{T + 2t_w}{3} < \frac{T + 2T}{3} = T.$$

Therefore, $t_{w+1} < T$ as well, completing the induction. Hence t_w remains below T for all w, and the sequence $\{t_w\}$ is strictly increasing. ■

Lemma 15 *Suppose the reverse process is divided into* W *segments with endpoints* $\{t_w\}_{w=0}^{W}$ *satisfying*

$$t_0 = 0, \quad t_{w+1} - t_w = 0.5 \cdot (T - t_{w+1}) \quad \textit{and} \quad t_W = T - \delta,$$

if we set

$$\beta_{t_w} \coloneqq 2d \log_2 K / \min\{1, T - t_w\}$$

then we have

$$\sum_{k=1}^{W} \beta_{t_w} \cdot (t_w - t_{w-1}) \leq 2d \log_2 K \cdot (T + \ln(1/\delta))$$

Proof [Adapted from Theorem 6 of Chen and Ying (2024)] Suppose there exist time steps $t_0, t_1, \ldots, t_W$ such that $T - t_w = s_w$ for each $w = 0, \ldots, W$. According to Lemma 14, we know $\{t_w\}_{w=0}^{W}$ is a increasing sequence, if we set

$$s_w \coloneqq T - t_w,$$

then it can be expected that $s_0 > s_1 > \cdots > s_W \geq \delta > 0$. According to the choice of β_w, it has

$$\beta_w = \frac{Cd \log_2 K}{\min(1,\, s_w)}, \quad \text{and} \quad s_{w-1} - s_w > 0.$$

For w such that $\delta \leq s_w < 1$, notice that $\min(1, s_w) = s_w$, we have $\beta_w = Cd \log_2 K / s_w$ and

$$\sum_{w:\delta \leq s_w < 1} \beta_w \cdot (t_w - t_{w-1}) = \sum_{w:\delta \leq s_w < 1} \beta_w \left(s_{w-1} - s_w\right) = \sum_{w:\delta \leq s_w < 1} \frac{Cd \log_2 K}{s_w} \left(s_{w-1} - s_w\right).$$

Because $1/s$ is a decreasing function for $s > 0$, we have

$$\frac{1}{s_w} \leq \frac{1}{s} \quad \text{for all } s \in [s_w, s_{w-1}],$$

which implies

$$\frac{Cd}{s_w} \left(s_{w-1} - s_w\right) \leq Cd \log_2 K \int_{s_w}^{s_{w-1}} \frac{1}{s} \mathrm{d}s.$$

Hence,

$$\sum_{w:\delta \leq s_w < 1} \frac{Cd \log_2 K}{s_w} \left(s_{w-1} - s_w\right) \leq Cd \log_2 K \sum_{w:\delta \leq s_w < 1} \int_{s_w}^{s_{w-1}} \frac{1}{s} \, ds = Cd \log_2 K \int_{\delta}^{1} \frac{1}{s} \, ds.$$

Evaluating the integral on the right gives

$$Cd \log_2 K \int_{\delta}^{1} \frac{1}{s} \mathrm{d}s = Cd \log_2 K \left[\ln(s)\right]_{\delta}^{1} = Cd \log_2 K \ln(1/\delta).$$

Therefore, we have established the exact upper bound

$$\sum_{k:\delta \leq s_k < 1} \lambda_k \left(s_{k-1} - s_k\right) \leq Cd \log_2 K \ln(1/\delta).$$

For $s_w \geq 1$, , notice that $\min(1, s_w) = 1$, we have $\beta_w = Cd \log_2 K$.

$$\sum_{w:1 \leq s_w \leq T} \beta_w \cdot (t_w - t_{w-1}) = \sum_{w:1 \leq s_w \leq T} \beta_w \left(s_{w-1} - s_w\right)$$
$$= \sum_{w:1 \leq s_w \leq T} Cd \log_2 K \cdot \left(s_{w-1} - s_w\right) \leq Cd \log_2 K \cdot (T - 1).$$

Combining the two parts, we have

$$\sum_{w=1}^{W} \beta_w \cdot (t_w - t_{w-1}) = \sum_{w=1}^{W} \beta_w \cdot (s_{w-1} - s_w)$$
$$= \sum_{w:\delta \leq s_w < 1} \beta_w \left(s_{w-1} - s_w\right) + \sum_{w:1 \leq s_w \leq T} \beta_w \left(s_{w-1} - s_w\right) \leq Cd \log_2 K \cdot \left(T + \ln(1/\delta)\right).$$

Hence, the proof is completed. ■

Lemma 16 *Following the notations shown in Section 2, we have*

$$\lim_{\Delta t \to 0} \left[\Delta t^{-1} \cdot \ln \frac{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}\right] = \hat{R}_t(\boldsymbol{y}) - R^{\leftarrow}_t(\boldsymbol{y}).$$

Proof Since we have required $\Delta t \to 0$, that is to say

$$\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \to \hat{q}_{t|t}(\boldsymbol{y}'|\boldsymbol{y}) = 0 \quad \text{and} \quad q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \to q^{\leftarrow}_{t|t}(\boldsymbol{y}'|\boldsymbol{y}) = 0 \quad \forall \boldsymbol{y}' \neq \boldsymbol{y},$$

which automatically makes

$$\left| \frac{\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \left(\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) - q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \right)}{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})} \right| \leq \frac{1}{2} < 1.$$

Under this condition, we have

$$\begin{aligned}
&\ln \frac{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})} = \ln \left[1 + \frac{\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \left(\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) - q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \right)}{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})} \right] \\
&= \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} \cdot \left[\frac{\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \left(\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) - q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \right)}{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})} \right]^i,
\end{aligned}$$

which implies (with the dominated convergence theorem)

$$\begin{aligned}
&\lim_{\Delta t \to 0} \left[\Delta t^{-1} \cdot \ln \frac{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})} \right] \\
&= \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} \cdot \lim_{\Delta t \to 0} \frac{\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \left(\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) - q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \right)}{\Delta t} \\
&\quad \cdot \lim_{\Delta t \to 0} \frac{\left(\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \left(\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) - q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \right) \right)^{i-1}}{\left(1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \right)^i}.
\end{aligned}$$

Only when $i = 1$, we have

$$\lim_{\Delta t \to 0} \frac{\left(\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \left(\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) - q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \right) \right)^{i-1}}{\left(1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \right)^i} = 1,$$

otherwise it will be equivalent to 0. Therefore, we have

$$\begin{aligned}
&\lim_{\Delta t \to 0} \left[\Delta t^{-1} \cdot \ln \frac{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}{1 - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})} \right] = \lim_{\Delta t \to 0} \frac{\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \left(\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) - q^{\leftarrow}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y}) \right)}{\Delta t} \\
&= \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \left(\hat{R}_t(\boldsymbol{y}', \boldsymbol{y}) - R^{\leftarrow}_t(\boldsymbol{y}', \boldsymbol{y}) \right) = \hat{R}_t(\boldsymbol{y}) - R^{\leftarrow}_t(\boldsymbol{y}).
\end{aligned}$$

Hence, the proof is completed. ■

E.3. The Proof of Lemma 4

Proof We start from the dynamic of KL divergence with the time growth in the reverse process, i.e.,

$$\frac{\mathrm{dKL}\left(q_t^{\leftarrow}\middle\|\hat{q}_t\right)}{\mathrm{d}t} = \lim_{\Delta t\to 0}\left[\frac{\mathrm{KL}\left(q_{t+\Delta t}^{\leftarrow}\middle\|\hat{q}_{t+\Delta t}\right)-\mathrm{KL}\left(q_t^{\leftarrow}\middle\|\hat{q}_t\right)}{\Delta t}\right]$$
$$\leq \lim_{\Delta t\to 0}\left[\frac{\mathbb{E}_{\mathbf{y}\sim q_t^{\leftarrow}}\left[\mathrm{KL}\left(q_{t+\Delta t|t}^{\leftarrow}(\cdot|\mathbf{y})\middle\|\hat{q}_{t+\Delta t|t}(\cdot|\mathbf{y})\right)\right]}{\Delta t}\right]$$

where the inequality follows from the chain rule of KL divergence. Under this condition, we have

$$\frac{\mathrm{dKL}\left(q_t^{\leftarrow}\middle\|\hat{q}_t\right)}{\mathrm{d}t} \leq \sum_{\boldsymbol{y}\in\mathcal{Y}} q_t^{\leftarrow}(\boldsymbol{y})\cdot \underbrace{\lim_{\Delta t\to 0}\left[\frac{\mathrm{KL}\left(q_{t+\Delta t|t}^{\leftarrow}(\cdot|\boldsymbol{y})\middle\|\hat{q}_{t+\Delta t|t}(\cdot|\boldsymbol{y})\right)}{\Delta t}\right]}_{\text{Term 1}}. \tag{43}$$

For each $\boldsymbol{y}\in\mathcal{Y}$, we focus on Term 1 of Eq. (43), and have

$$\begin{aligned}\text{Term 1} &= \lim_{\Delta t\to 0}\left[\Delta t^{-1}\cdot\sum_{\boldsymbol{y}'\in\mathcal{Y}} q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})\cdot\ln\frac{q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})}{\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}\right]\\ &= \underbrace{\lim_{\Delta t\to 0}\left[\sum_{\boldsymbol{y}'\neq\boldsymbol{y}}\frac{q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})}{\Delta t}\cdot\ln\frac{q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})}{\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}\right]}_{\text{Term 1.1}}+\\ &\quad\underbrace{\lim_{\Delta t\to 0}\left[\Delta t^{-1}\cdot\left(1-\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})\right)\cdot\ln\frac{1-\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})}{1-\sum_{\boldsymbol{y}'\neq\boldsymbol{y}}\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}\right]}_{\text{Term 1.2}}.\end{aligned} \tag{44}$$

For Term 1.1, we have

$$\begin{aligned}\text{Term 1.1} &= \sum_{\boldsymbol{y}'\neq\boldsymbol{y}}\lim_{\Delta t\to 0}\left[\frac{q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})}{\Delta t}\right]\cdot\lim_{\Delta t\to 0}\left[\ln\frac{q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})}{\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}\right]\\ &= \sum_{\boldsymbol{y}'\neq\boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}',\boldsymbol{y})\cdot\ln\left[\lim_{\Delta t\to 0}\left(\frac{q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})}{\Delta t}\cdot\frac{\Delta t}{\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}\right)\right]\\ &= \sum_{\boldsymbol{y}'\neq\boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}',\boldsymbol{y})\cdot\ln\frac{R_t^{\leftarrow}(\boldsymbol{y}',\boldsymbol{y})}{\hat{R}_t(\boldsymbol{y}',\boldsymbol{y})},\end{aligned} \tag{45}$$

where the second equation follows from the composition rule of the limit calculation. For Term 1.2, we have

$$\begin{aligned}&\text{Term 1.2} = \lim_{\Delta t\to 0}\left[1-\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})\right]\cdot\lim_{\Delta t\to 0}\left[\Delta t^{-1}\cdot\ln\frac{1-\sum_{\boldsymbol{y}'\neq\boldsymbol{y}} q_{t+\Delta t|t}^{\leftarrow}(\boldsymbol{y}'|\boldsymbol{y})}{1-\sum_{\boldsymbol{y}'\neq\boldsymbol{y}}\hat{q}_{t+\Delta t|t}(\boldsymbol{y}'|\boldsymbol{y})}\right]\\ &= \sum_{\boldsymbol{y}'\neq\boldsymbol{y}}\left(\hat{R}_t(\boldsymbol{y}',\boldsymbol{y})-R_t^{\leftarrow}(\boldsymbol{y}',\boldsymbol{y})\right) = \hat{R}_t(\boldsymbol{y})-R_t^{\leftarrow}(\boldsymbol{y})\end{aligned} \tag{46}$$

where the first inequality follows from Lemma 16. Plugging Eq. (45), Eq. (46) and Eq. (44), into Eq. (43) we have

$$\frac{\mathrm{dKL}\left(q_t^{\leftarrow} \| \hat{q}_t\right)}{\mathrm{d}t} \leq \sum_{\boldsymbol{y} \in \mathcal{Y}} q_t^{\leftarrow}(\boldsymbol{y}) \cdot \left(\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \ln \frac{R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y})}{\hat{R}_t(\boldsymbol{y}', \boldsymbol{y})} + \hat{R}_t(\boldsymbol{y}) - R_t^{\leftarrow}(\boldsymbol{y}) \right). \tag{47}$$

For any $\boldsymbol{y} \in \mathcal{Y}$, we have

$$\begin{aligned} &\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \ln \frac{R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y})}{\hat{R}_t(\boldsymbol{y}', \boldsymbol{y})} + \hat{R}_t(\boldsymbol{y}) - R_t^{\leftarrow}(\boldsymbol{y}) \\ &= \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y}) \ln \frac{R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y})}{\tilde{R}_t(\boldsymbol{y}', \boldsymbol{y})} + \tilde{R}_t(\boldsymbol{y}) - R_t^{\leftarrow}(\boldsymbol{y}) \\ &\quad \underbrace{+ \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y}) \ln \frac{\tilde{R}_t(\boldsymbol{y}', \boldsymbol{y})}{\hat{R}_t(\boldsymbol{y}', \boldsymbol{y})} + \hat{R}_t(\boldsymbol{y}) - \tilde{R}_t(\boldsymbol{y})}_{\text{Term 2}}. \end{aligned} \tag{48}$$

When $\tilde{R}_t(\boldsymbol{y}) \leq \beta_t$, we have

$$\hat{R}_t(\boldsymbol{y}', \boldsymbol{y}) = \tilde{R}_t(\boldsymbol{y}', \boldsymbol{y}) \quad \text{and} \quad \hat{R}(\boldsymbol{y}) = \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \hat{R}(\boldsymbol{y}', \boldsymbol{y}) = \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \tilde{R}(\boldsymbol{y}', \boldsymbol{y}) = \tilde{R}(\boldsymbol{y})$$

which implies Term $2 = 0$ in Eq. (48). Otherwise, we have

$$\frac{\hat{R}_t(\boldsymbol{y}', \boldsymbol{y})}{\tilde{R}_t(\boldsymbol{y}', \boldsymbol{y})} = \frac{\beta_t}{\tilde{R}_t(\boldsymbol{y})} \quad \text{and} \quad \frac{\hat{R}_(\boldsymbol{y})}{\tilde{R}_t(\boldsymbol{y})} = \frac{\beta_t}{\tilde{R}_t(\boldsymbol{y})},$$

which implies

$$\begin{aligned} \text{Term } 2 &= \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \ln \frac{\tilde{R}_t(\boldsymbol{y})}{\beta_t} + \beta_t - \tilde{R}_t(\boldsymbol{y}) \\ &= R_t^{\leftarrow}(\boldsymbol{y}) \cdot \ln \left[1 + \frac{\tilde{R}_t(\boldsymbol{y}) - \beta_t}{\beta_t} \right] + \beta_t - \tilde{R}_t(\boldsymbol{y}) \leq \beta_t \cdot \left[\frac{\tilde{R}_t(\boldsymbol{y}) - \beta_t}{\beta_t} \right] + \beta_t - \tilde{R}_t = 0. \end{aligned}$$

Combining with Eq. (48) and Eq. (47), we have

$$\begin{aligned} &\frac{\mathrm{dKL}\left(q_t^{\leftarrow} \| \hat{q}_t\right)}{\mathrm{d}t} \leq \sum_{\boldsymbol{y} \in \mathcal{Y}} q_t^{\leftarrow}(\boldsymbol{y}) \cdot \left(\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \ln \frac{R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y})}{\tilde{R}_t(\boldsymbol{y}', \boldsymbol{y})} + \tilde{R}_t(\boldsymbol{y}) - R_t^{\leftarrow}(\boldsymbol{y}) \right) \\ &= \sum_{\boldsymbol{y} \in \mathcal{Y}} q_t^{\leftarrow}(\boldsymbol{y}) \cdot \left(\sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y}) \cdot \ln \frac{R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y})}{\tilde{R}_t(\boldsymbol{y}', \boldsymbol{y})} + \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} \tilde{R}_t(\boldsymbol{y}', \boldsymbol{y}) - \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R_t^{\leftarrow}(\boldsymbol{y}', \boldsymbol{y}) \right) \\ &= \sum_{\boldsymbol{y} \in \mathcal{Y}} q_t^{\leftarrow}(\boldsymbol{y}) \cdot \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R^{\rightarrow}(\boldsymbol{y}, \boldsymbol{y}') \cdot \left[-\frac{q_t^{\leftarrow}(\boldsymbol{y}')}{q_t^{\leftarrow}(\boldsymbol{y})} + \hat{v}_{t,\boldsymbol{y}}(\boldsymbol{y}') + \frac{q_t^{\leftarrow}(\boldsymbol{y}')}{q_t^{\leftarrow}(\boldsymbol{y})} \ln \frac{q_t^{\leftarrow}(\boldsymbol{y}')}{q_t^{\leftarrow}(\boldsymbol{y}) \hat{v}_{t,\boldsymbol{y}}(\boldsymbol{y}')} \right] \\ &= \sum_{\boldsymbol{y} \in \mathcal{Y}} q_t^{\leftarrow}(\boldsymbol{y}) \cdot \sum_{\boldsymbol{y}' \neq \boldsymbol{y}} R^{\rightarrow}(\boldsymbol{y}, \boldsymbol{y}') D_\phi \left(\frac{q_t^{\leftarrow}(\boldsymbol{y}')}{q_t^{\leftarrow}(\boldsymbol{y})} \| \hat{v}_{t,\boldsymbol{y}}(\boldsymbol{y}') \right), \end{aligned} \tag{49}$$

where D_ϕ is the Bregman divergence with $\phi(c) = c \ln c$ (as Eq. (5)), and the last equation follows from the definition of Bregman divergence:

$$D_\phi(u\|v) = \phi(u) - \phi(v) - \langle \nabla\phi(v), u - v\rangle = u \ln \frac{u}{v} - u + v.$$

Then, by Eq. (5) and Assumption **[A1]**, we have

$$\int_0^{T-\delta} \mathrm{dKL}\left(q_t^{\leftarrow} \middle\| \hat{q}_t\right) \le (T-\delta)\epsilon_{\text{score}}^2.$$

Hence, the proof is completed. ■

Bounding $\mathrm{TV}\left(q_*, q_\delta^{\rightarrow}\right)$ We adopt the proof strategy of Theorem 6 in Chen and Ying (2024). Consider the forward process $(X_t)_{t\ge 0}$. By the coupling characterization of the total variation distance, we have

$$\mathrm{TV}\left(q_*, q_\delta^{\rightarrow}\right) \coloneqq \inf_{\gamma\in\Gamma(q_*, q_\delta^{\rightarrow})} \mathbb{P}_{(u,v)\sim\gamma}[u \ne v] \le \mathbb{P}(X_0 \ne X_\delta),$$

where $\Gamma(q_*, q_\delta^{\rightarrow})$ is the set of all couplings of $(q_*, q_\delta^{\rightarrow})$, and the inequality holds because (X_0, X_δ) gives a coupling of $(q_*, q_\delta^{\rightarrow})$.

By the transition kernel given in (Chen and Ying, 2024, Proposition 3), we have

$$\mathbb{P}(X_0 = X_\delta) = \frac{1}{2^{d\log_2 K}} \Pi_{i=1}^{d\log_2 K}(1 + (-1)^0 e^{-2\delta})^{d\log_2 K} = \left(\frac{1+e^{-2\delta}}{2}\right)^{d\log_2 K} \ge e^{-\delta d \log_2 K},$$

where the inequality holds due to the convexity of the exponential function. Thus,

$$\mathrm{TV}\left(q_*, q_\delta^{\rightarrow}\right) \le 1 - e^{-\delta d\log_2 K} \tag{50}$$

Proof [Proof of Theorem 6] We start from the quantization algorithm, i.e., Alg. 1. Since the data distribution p_* is supposed to satisfy Assumption **[A2]**–**[A4]**, by introducing Lemma 1, the histogram-like approximation $\overline{p}_*$ will be close to p_*, i.e.,

$$\mathrm{TV}\left(\overline{p}_*, p_*\right) \le 3\epsilon$$

by choosing

$$L = \sigma \cdot \sqrt{2\ln(2d/\epsilon)} \quad \text{and} \quad l = \left[2H \cdot \left(\sigma\sqrt{2d\ln(2d/\epsilon)} + d + \sqrt{dm_0}\right)\right]^{-1} \cdot \epsilon.$$

Under this condition, we have

$$\begin{aligned} K = \frac{2L}{l} &= 4H \cdot \left[2\sigma^2 d^{1/2} \cdot \ln\frac{2d}{\epsilon} + \sigma d \cdot \sqrt{2\ln\frac{2d}{\epsilon}} + d^{1/2} m_0^{1/2} \cdot \sqrt{2\ln\frac{2d}{\epsilon}}\right] \cdot \epsilon^{-1} \\ &\le 24 H \sigma^2 d m_0 \epsilon^{-1} \cdot \ln(2d/\epsilon) \end{aligned}$$

where the last inequality follows from $\sigma \ge 1$ and $m_0 \ge 1$ without loss of generality. Then, after the training, the implementation of Alg. 2 requires $\overline{N} \sim \mathrm{Poisson}(\overline{\beta})$ steps.

Proof of bound of the expectation of $\overline{N}$, i.e., $\overline{\beta}$. According to Lemma 15, if we set

$$t_0 = 0, \quad t_{w+1} - t_w = 0.5 \cdot (T - t_{w+1}) \quad \text{and} \quad t_W = T - \delta,$$

for the time partitions,

$$\beta_{t_w} \coloneqq 2d \log_2 K / \min\{1, T - t_w\}$$

for the intermediate Poisson, then it has

$$\begin{aligned}\overline{\beta} &= \sum_{k=1}^{W} \beta_{t_w} \cdot (t_w - t_{w-1}) \leq 2d \log_2 K \cdot (T + \ln(1/\delta)) \\ &\leq 2d \cdot \left[\log_2(24H\sigma^2) + \log_2(dm_0/\epsilon) + \log_2[\ln(2d/\epsilon)]\right] \cdot (T + \ln(1/\delta)) \,.\end{aligned}$$

Proof of the TV distance bound. Since our truncated uniformization, i.e., Alg. 2, exactly simulates the reversed process, from Lemma 4, the KL divergence gap between $q^{\leftarrow}_{T-\delta} = q^{\rightarrow}_{\delta}$ and $\hat{q}_{T-\delta}$ is bounded by the KL divergence as follows:

$$\begin{aligned}&\mathrm{KL}\left(q^{\leftarrow}_{T-\delta} \middle\| \hat{q}_{T-\delta}\right) \leq \mathrm{KL}\left(q^{\leftarrow}_{0} \middle\| \hat{q}_{0}\right) + (T - \delta)\epsilon^2_{\text{score}} \\ &\leq e^{-T} \cdot d \log_2 K + (T - \delta)\epsilon^2_{\text{score}} = \epsilon^2 + (\ln(d/\epsilon) + \ln \log_2 K)^2 \cdot \epsilon^2_{\text{score}} \leq 2\epsilon^2\end{aligned} \tag{51}$$

where the second inequality follows from Lemma 2, the third inequality establishes when T is chosen as

$$T = \ln(d/\epsilon) + \ln \log_2 K,$$

and the last inequality is established when we have

$$\epsilon_{\text{score}} = \frac{\epsilon}{\ln(d/\epsilon) + \ln \log_2 K} = \tilde{O}(\epsilon).$$

Under this condition, due to Pinsker's inequality, Eq. (51) can be relaxed to

$$\mathrm{TV}\left(q^{\leftarrow}_{T-\delta}, \hat{q}_{T-\delta}\right) \leq \sqrt{\frac{\mathrm{KL}\left(q^{\leftarrow}_{T-\delta} \middle\| \hat{q}_{T-\delta}\right)}{2}} \leq \epsilon.$$

Then we have

$$\begin{aligned}&\mathrm{TV}\left(q_*, \hat{q}_{T-\delta}\right) \leq \mathrm{TV}\left(q_*, q^{\leftarrow}_{T-\delta}\right) + \mathrm{TV}\left(q^{\leftarrow}_{T-\delta}, \hat{q}_{T-\delta}\right) \\ &= \mathrm{TV}\left(q_*, q^{\rightarrow}_{\delta}\right) + \mathrm{TV}\left(q^{\leftarrow}_{T-\delta}, \hat{q}_{T-\delta}\right) = 1 - e^{-\delta d \log_2 K} + \epsilon \leq 2\epsilon\end{aligned}$$

where the second equation follows from Eq. (50) and the last inequality is established by requiring

$$\delta \leq \frac{\epsilon}{d \cdot \log_2 K} \quad \Leftrightarrow \quad \delta d \log_2 K \leq \epsilon.$$

Under this condition, we have

$$\delta d \log_2 K \leq \epsilon \leq \ln \frac{1}{1 - \epsilon} \quad \Rightarrow \quad 1 - e^{-\delta d \log_2 K} \leq \epsilon.$$

Suppose the underlying distributions of $\overline{\mathbf{y}}, \hat{\mathbf{x}}$ are $\overline{q}, \hat{p}$ respectively, due to the connection between $\hat{p}, \overline{p}_*$ and $\overline{q}, \overline{q}_*$ shown in Eq. (12), we have

$$\mathrm{TV}\left(\overline{p}_*, \hat{p}\right) = \int \left|\overline{p}_*(\boldsymbol{x}) - \hat{p}(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x} = \sum_{\overline{\boldsymbol{y}} \in \overline{\mathcal{Y}}} \left|\overline{q}(\overline{\boldsymbol{y}}) - \overline{q}_*(\overline{\boldsymbol{y}})\right| = \sum_{\boldsymbol{y} \in \mathcal{Y}} \left|\hat{q}_{T-\delta}(\boldsymbol{y}) - \hat{q}_*(\boldsymbol{y})\right| \leq 2\epsilon.$$

Combining this result with Lemma 1, we have

$$\mathrm{TV}\left(p_*, \hat{p}\right) \leq \mathrm{TV}\left(p_*, \overline{p}_*\right) + \mathrm{TV}\left(\overline{p}_*, \hat{p}\right) \leq 3\epsilon + 2\epsilon \leq 5\epsilon.$$

Hence, the proof is completed. ■

E.4. Proof of Theorem 7.

Proof We start from proving $W_2(p_*, \overline{p}_*) \leq 2\epsilon$ by analyzing the quantization Alg. 1, and then prove $\mathrm{TV}\left(\overline{p}_*, \hat{p}\right) \leq 2\epsilon$. We emphasize that the whole analysis does not require the Lipschitzness assumption.

Proof of $W_2(p_*, \overline{p}_*) \leq 2\epsilon$: We prove $W_2(p_*, \tilde{p}_*) \leq \epsilon$ and $W_2(\overline{p}_*, \tilde{p}_*) \leq \epsilon$ both by considering the coupling, and then use triangle inequality to conclude. First, under the tail truncation step, i.e., Eq. (9) in our paper to construct $\tilde{p}_*$, i.e.,

$$\tilde{p}_*(x) := \frac{p_*(x)}{\int_{x \in \mathrm{Cube}(L)} p_*(x) dx} \quad \forall x \in \mathrm{Cube}(L).$$

To control the Wasserstein 2 distance between $\tilde{p}_*$ and p_*, we construct the explicit coupling $(\mathbf{x}, \mathbf{y})$ in the following procedure: Let $\mathbf{x} \sim p_*$, and independently sample an auxiliary randan variable $\mathbf{z} \sim \tilde{p}_*$, where $\mathbf{z} \perp\!\!\!\perp \mathbf{x}$. Let $\mathbf{y} = \mathbf{x} \cdot \mathbf{1}(\mathbf{x} \in \mathrm{Cube}(L)) + \mathbf{z} \cdot \mathbf{1}(\mathbf{x} \notin \mathrm{Cube}(L))$. Then, for any measurable $A \subseteq \mathrm{Cube}(L)$, we have

$$\begin{aligned}
&\mathbb{P}(\mathbf{y} \in A) = \mathbb{P}(\mathbf{x} \in A, \mathbf{x} \in \mathrm{Cube}(L)) + \mathbb{P}(\mathbf{x} \notin \mathrm{Cube}(L), \mathbf{z} \in A) \\
&= \mathbb{P}(\mathbf{x} \in A) + \mathbb{P}(\mathbf{x} \notin \mathrm{Cube}(L)) \cdot \mathbb{P}(\mathbf{z} \in A) \\
&= p_*(A) + (1 - p_*(\mathrm{Cube}(L))) \cdot \tilde{p}_*(A) \\
&= p_*(\mathrm{Cube}(L)) \cdot \tilde{p}_*(A) + (1 - p_*(\mathrm{Cube}(L))) \cdot \tilde{p}_*(A) = \tilde{p}_*(A).
\end{aligned}$$

Therefore, the marginal distribution of $\mathbf{y}$ is $\tilde{p}_*$.

Hence $(\mathbf{x}, \mathbf{y}) \sim \gamma$ gives a coupling of $(p_*, \tilde{p}_*)$. Then, we have

$$W_2^2(\tilde{p}_*, p_*) \leq \mathbb{E}_\gamma\left[\|\mathbf{x} - \mathbf{y}\|^2\right] \leq \underbrace{\mathbb{E}_\gamma\left[\|\mathbf{x} - \mathbf{y}\|^2 \mathbf{1}\{\mathbf{x} \in \mathrm{Cube}(L)\}\right]}_{=0} + \mathbb{E}_\gamma\left[\|\mathbf{x} - \mathbf{y}\|^2 \mathbf{1}\{\mathbf{x} \notin \mathrm{Cube}(L)\}\right]. \tag{52}$$

For the last term, since it has $\|\mathbf{y}\|^2 \leq dL^2 \leq \|\mathbf{x}\|^2$, we have

$$\mathbb{E}_\gamma\left[\|\mathbf{x} - \mathbf{y}\|^2 \mathbf{1}\{\mathbf{x} \notin \mathrm{Cube}(L)\}\right] \leq 2\mathbb{E}\left[(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2) \mathbf{1}\{\mathbf{x} \notin \mathrm{Cube}(L)\}\right] \leq 4\mathbb{E}\left[\|\mathbf{x}\|^2 \mathbf{1}\{\|\mathbf{x}\| \geq L\}\right].$$

Therefore, to upper bound $W_2^2(\tilde{p}_*, p_*)$, we bound $\mathbb{E}\left[\|\mathbf{x}\|^2 \mathbf{1}\{\|\mathbf{x}\| \geq L\}\right]$. We use the general identity (i.e., Layer cake representation) for non-negative random variables $W > 0$:

$$\mathbb{E}[W] = \int_0^\infty \mathbb{P}(W \geq t)\, dt.$$

By taking $W = \|\mathbf{x}\|^2 \mathbf{1}\{\|\mathbf{x}\| \geq L\}$, we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}\|^2 \mathbf{1}\{\|\mathbf{x}\| \geq L\}\right] &= \int_0^\infty \mathbb{P}(\|\mathbf{x}\|^2 \mathbf{1}\{\|\mathbf{x}\| \geq L\} \geq t) dt = \int_0^\infty \mathbb{P}(\|\mathbf{x}\|^2 \geq t, \|\mathbf{x}\| \geq L) dt \\ &= \int_0^{L^2} \mathbb{P}(\|\mathbf{x}\|^2 \geq L^2) dt + \int_{L^2}^\infty \mathbb{P}(\|\mathbf{x}\|^2 \geq t) dt.\end{aligned} \tag{53}$$

By choosing $u = e_i$ in Assumption **[A4]**, we have that each dimension of x is σ sub-Gaussian, i.e.,

$$\mathbb{E}_{\mathbf{x}_i \sim p_{*,i}}\left[\exp\left(r\mathbf{x}_i\right)\right] \leq \exp\left(\frac{\sigma^2 r^2}{2}\right).$$

Thus we have $\mathbb{P}_i\left[|\mathbf{x}_i| \geq r\right] \leq 2\exp\left(-\frac{r^2}{2\sigma^2}\right)$. With the union bound, we have for any $t > 0$,

$$\mathbb{P}_{\mathbf{x}\sim p_*}(\|\mathbf{x}\|^2 \geq t) \leq \mathbb{P}_{\mathbf{x}\sim p_*}\left[\max_{1\leq i\leq d} |\mathbf{x}_i| > \sqrt{t/d}\right] \leq \sum_{i=1}^d \mathbb{P}_{\mathbf{x}\sim p_*}\left[|\mathbf{x}_i| > \sqrt{t/d}\right] \leq 2d\cdot\exp\left(-\frac{t}{2d\sigma^2}\right).$$

Therefore, the upper bound of Eq. (53) can be written as

$$\mathbb{E}\left[\|\mathbf{x}\|^2 \mathbf{1}\{\|x\| \geq L\}\right] = \int_0^{L^2} \mathbb{P}(\|\mathbf{x}\|^2 \geq L^2) dt + \int_{L^2}^\infty \mathbb{P}(\|\mathbf{x}\|^2 \geq t) dt \leq 2d(L^2 + 2d\sigma^2)\exp(-\frac{L^2}{2d\sigma^2}).$$

Plugging this result into Eq. (52), we have

$$W_2^2(\tilde{p}_*, p_*) \leq 8d(L^2 + 2d\sigma^2)\exp(-\frac{L^2}{2d\sigma^2}).$$

Therefore, by choosing

$$L \geq \sqrt{2d}\sigma \cdot \sqrt{4\ln\frac{2d\sigma}{\epsilon}},$$

we have

$$8d(L + 2d\sigma^2)\exp(-\frac{L^2}{2d\sigma^2}) \leq \epsilon^2.$$

Thus, we have $W_2(\tilde{p}_*, p_*) \leq \epsilon$.

Next, we consider the construction of $\overline{p}_*$ in Eq. (11): $\overline{p}_*$ is a piecewise constant function that averages the $\tilde{p}_*$ over each quantization cell. Therefore, the total mass of each cell remains the same, i.e., $\overline{p}_*(C_k) = \tilde{p}_*(C_k)$ for any cell C_k. Therefore, we construct the coupling of $(\overline{p}_*, \tilde{p}_*)$ by considering each cell C_k: within the cell C_k, choose the independent coupling γ_k that transports the mass distribution from $\tilde{p}_*(\cdot|C_k)$ to the uniform measure $\overline{p}_*(\cdot|C_k)$ only within C_k. Specifically, construct the coupling (X, Y) as follows:

1. Sample the cell index according to the PMF $\mathbb{P}(\text{index} = k) = \tilde{p}_*(C_k)$.
2. Conditioned on $K = k$, sample X from the conditional law $\overline{p}_*(\cdot \mid C_k)$; sample Y, independently of X, from the conditional law $\tilde{p}_*(\cdot \mid C_k)$.

Since $\overline{p}_*(C_k) = \tilde{p}_*(C_k)$, we have

$$p_X(x) = \overline{p}_*(x \mid C_k)\tilde{p}_*(C_k) = \overline{p}_*(x \mid C_k)\overline{p}_*(C_k) = \overline{p}_*(x)$$
$$\text{and} \quad p_Y(y) = \tilde{p}_*(y \mid C_k)\tilde{p}_*(C_k) = \tilde{p}_*(y).$$

Hence, $(\mathbf{x}, \mathbf{y})$ gives a coupling of $(\overline{p}_*, \tilde{p}_*)$. Conditioned on each cell C_k, we have

$$\|\mathbf{x} - \mathbf{y}\| \le \sqrt{d} \cdot l, \quad \mathbf{x} \sim \tilde{p}_*(\cdot|C_k) \quad \text{and} \quad \mathbf{y} \sim \overline{p}_*(\cdot|C_k),$$

where l is the diameter of cell. Since all probability mass stays in its original cell, we get

$$\mathbb{E}[\|X - Y\|^2] \le dl^2 \quad \Rightarrow \quad W_2^2(\tilde{p}_*, \overline{p}) \le dl^2,$$

which implies $W_2(\tilde{p}_*, \overline{p}) \le \epsilon$ by requiring $l = \epsilon/\sqrt{d}$. Such a choice is independent of the smoothness of p_*. Due to the triangle inequality, we have $W_2(p_*, \overline{p}_*) \le 2\epsilon$.

Proof of the TV distance bound. Since our truncated uniformization, i.e., Alg. 2, exactly simulates the reversed process, from Lemma 4, the KL divergence gap between $q^{\leftarrow}_{T-\delta} = q^{\rightarrow}_{\delta}$ and $\hat{q}_{T-\delta}$ is bounded by the KL divergence as follows:

$$\begin{aligned} &\mathrm{KL}\left(q^{\leftarrow}_{T-\delta} \middle\| \hat{q}_{T-\delta}\right) \le \mathrm{KL}\left(q^{\leftarrow}_{0} \middle\| \hat{q}_{0}\right) + (T-\delta)\epsilon^2_{\text{score}} \\ &\le e^{-T} \cdot d \log_2 K + (T-\delta)\epsilon^2_{\text{score}} = \epsilon^2 + (\ln(d/\epsilon) + \ln\log_2 K)^2 \cdot \epsilon^2_{\text{score}} \le 2\epsilon^2, \end{aligned} \tag{54}$$

where the second inequality follows from Lemma 2, the third inequality establishes when T is chosen as

$$T = \ln(d/\epsilon) + \ln\log_2 K,$$

and the last inequality is established when we have

$$\epsilon_{\text{score}} = \frac{\epsilon}{\ln(d/\epsilon) + \ln\log_2 K} = \tilde{O}(\epsilon).$$

Under this condition, due to Pinsker's inequality, Eq. (54) can be relaxed to

$$\mathrm{TV}\left(q^{\leftarrow}_{T-\delta}, \hat{q}_{T-\delta}\right) \le \sqrt{\frac{\mathrm{KL}\left(q^{\leftarrow}_{T-\delta} \middle\| \hat{q}_{T-\delta}\right)}{2}} \le \epsilon.$$

Then we have

$$\begin{aligned} &\mathrm{TV}\left(q_*, \hat{q}_{T-\delta}\right) \le \mathrm{TV}\left(q_*, q^{\leftarrow}_{T-\delta}\right) + \mathrm{TV}\left(q^{\leftarrow}_{T-\delta}, \hat{q}_{T-\delta}\right) \\ &= \mathrm{TV}\left(q_*, q^{\rightarrow}_{\delta}\right) + \mathrm{TV}\left(q^{\leftarrow}_{T-\delta}, \hat{q}_{T-\delta}\right) = 1 - e^{-\delta d \log_2 K} + \epsilon \le 2\epsilon \end{aligned}$$

where the second equation follows from Eq. (50) and the last inequality is established by requiring

$$\delta \le \frac{\epsilon}{d \cdot \log_2 K} \quad \Leftrightarrow \quad \delta d \log_2 K \le \epsilon.$$

Under this condition, we have

$$\delta d \log_2 K \le \epsilon \le \ln \frac{1}{1-\epsilon} \quad \Rightarrow \quad 1 - e^{-\delta d \log_2 K} \le \epsilon.$$

Suppose the underlying distributions of $\overline{y}, \hat{x}$ are $\overline{q}, \hat{p}$ respectively, due to the connection between $\hat{p}, \overline{p}_*$ and $\overline{q}, \overline{q}_*$ shown in Eq. (12), we have

$$\mathrm{TV}\left(\overline{p}_*, \hat{p}\right) = \int |\overline{p}_*(\boldsymbol{x}) - \hat{p}(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x} = \sum_{\overline{\boldsymbol{y}} \in \overline{\mathcal{Y}}} |\overline{q}(\overline{\boldsymbol{y}}) - \overline{q}_*(\overline{\boldsymbol{y}})| = \sum_{\boldsymbol{y} \in \mathcal{Y}} |\hat{q}_{T-\delta}(\boldsymbol{y}) - \hat{q}_*(\boldsymbol{y})| \le 2\epsilon.$$

Proof of bound of the expectation of $\overline{N}$, i.e., $\overline{\beta}$. After the training, the implementation of Alg. 2 requires $\overline{N} \sim \text{Poisson}(\overline{\beta})$ steps. According to Lemma 15, if we set

$$t_0 = 0, \quad t_{w+1} - t_w = 0.5 \cdot (T - t_{w+1}) \quad \text{and} \quad t_W = T - \delta,$$

for the time partitions,

$$\beta_{t_w} := 2d \log_2 K / \min\{1, T - t_w\}$$

for the intermediate Poisson, then it has

$$\begin{aligned}\overline{\beta} &= \sum_{k=1}^{W} \beta_{t_w} \cdot (t_w - t_{w-1}) \leq 2d \log_2 K \cdot (T + \ln(1/\delta)) \\ &= 2d \cdot \left[\log_2(4\sqrt{2d}\sigma \cdot \sqrt{\ln(2d\sigma/\epsilon)} \cdot \sqrt{d}/\epsilon)\right] \cdot (T + \ln(1/\delta))\end{aligned}$$

where the last equation is due to $L = 2\sqrt{2d}\sigma \cdot \sqrt{\ln(2d\sigma/\epsilon)}$, $l = \epsilon/\sqrt{d}$, and $K = 2L/l$. Therefore, the proof is completed. ■

Appendix F. Technical Lemmas

Lemma 17 (Theorem 4.10 of Boucheron et al. (2003)) *Let $\Phi(x) = x \ln x$ for $x > 0$ and $\Phi(0) = 0$. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be independent random variables taking values in a countable set $\mathcal{X}$ and let $f \colon \mathcal{X} \to [0, \infty)$. We have*

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n}\left[\Phi\left(f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)\right)\right] - \Phi\left(\mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n}\left[f\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\right)\right]\right) \\ &\leq \sum_{i=1}^{n} \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_{i-1},\mathbf{x}_{i+1},\ldots,\mathbf{x}_n}\left[\mathbb{E}_{\mathbf{x}_i}\left[\Phi\left(f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)\right)\right] - \Phi\left(\mathbb{E}_{\mathbf{x}_i}\left[f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)\right]\right)\right].\end{aligned}$$

Lemma 18 (Chain rule of TV) *Consider four random variables, $\mathbf{x}, \mathbf{z}, \tilde{\mathbf{x}}, \tilde{\mathbf{z}}$, whose underlying distributions are denoted as p_x, p_z, q_x, q_z. Suppose $p_{x,z}$ and $q_{x,z}$ denotes the densities of joint distributions of $(\mathbf{x}, \mathbf{z})$ and $(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$, which we write in terms of the conditionals and marginals as*

$$\begin{aligned}p_{x,z}(\boldsymbol{x}, \boldsymbol{z}) &= p_{x|z}(\boldsymbol{x}|\boldsymbol{z}) \cdot p_z(\boldsymbol{z}) = p_{z|x}(\boldsymbol{z}|\boldsymbol{x}) \cdot p_x(\boldsymbol{x}) \\ q_{x,z}(\boldsymbol{x}, \boldsymbol{z}) &= q_{x|z}(\boldsymbol{x}|\boldsymbol{z}) \cdot q_z(\boldsymbol{z}) = q_{z|x}(\boldsymbol{z}|\boldsymbol{x}) \cdot q_x(\boldsymbol{x}).\end{aligned}$$

then we have

$$\begin{aligned}\text{TV}\left(p_{x,z}, q_{x,z}\right) \leq \min\big\{&\text{TV}\left(p_z, q_z\right) + \mathbb{E}_{\mathbf{z}\sim p_z}\left[\text{TV}\left(p_{x|z}(\cdot|\mathbf{z}), q_{x|z}(\cdot|\mathbf{z})\right)\right], \\ &\text{TV}\left(p_x, q_x\right) + \mathbb{E}_{\mathbf{x}\sim p_x}\left[\text{TV}\left(p_{z|x}(\cdot|\mathbf{x}), q_{z|x}(\cdot|\mathbf{x})\right)\right]\big\}.\end{aligned}$$

Besides, we have

$$\text{TV}\left(p_x, q_x\right) \leq \text{TV}\left(p_{x,z}, q_{x,z}\right).$$

Lemma 19 (Backward Kolmogorov equation) *Suppose the infinitesimal operator of a Markov semigroup is $\mathcal{L}$, If we denote the transition density from $\mathbf{y}_s = \boldsymbol{y}$ to $\mathbf{y}_t = \boldsymbol{y}'$ as $p_{t|s}(\boldsymbol{y}'|\boldsymbol{y})$, then it solves the backward Kolmogorov equation*

$$-\frac{\partial p_{t|s}(\boldsymbol{y}'|\boldsymbol{y})}{\partial s} = \mathcal{L}\left[p_{t|s}(\boldsymbol{y}'|\cdot)\right](\boldsymbol{y}), \quad p_{s|s}(\boldsymbol{y}'|\boldsymbol{y}) = \delta(\boldsymbol{y}' - \boldsymbol{y}).$$

Lemma 20 (Lemma 11 in Vempala and Wibisono (2019)) *Suppose the density function satisfies* $p \propto \exp(-f)$ *where* f *is* H*-smooth, i.e., [A3]. Then, it has*

$$\mathbb{E}_{\mathbf{x}\sim p}\left[\|\nabla f(\mathbf{x})\|^2\right] \leq Hd.$$

Appendix G. Empirical Results

To evaluate the efficiency and accuracy of QTD relative to DDPM, we conduct a synthetic experiment on a 1D Gaussian Mixture Model (GMM) in Section G.1. On the other hand, although Table 2 indicates that the complexity of most diffusion inference methods scales with d, in practice the number of inference iterations is typically much smaller than the particle dimension. To demonstrate the practical effectiveness of QTD, we introduce a multi-flip trick that yields an inexact variant of QTD, allowing the number of iterations to be smaller than d. The implementation details and corresponding empirical results are presented in Section G.2.

G.1. Experiments on Synthetic Data

Data Setup. We define the target distribution as a 2-component GMM, discretized into $K = 128$ categories over the range $[-4, 4]$. For each of the 10 independent experimental seeds, we randomly sample the target parameters: component means $\mu_1 \sim \mathcal{U}[-1.5, -1.0]$ and $\mu_2 \sim \mathcal{U}[1.0, 1.5]$, standard deviations $\sigma_1, \sigma_2 \sim \mathcal{U}[0.4, 1.0]$, and mixing weights $w_1 \sim \mathcal{U}[0.3, 0.7]$ (with $w_2 = 1 - w_1$).

Experimental Setup and Architecture. We employ three-layer MLP backbones with comparable model capacity and ReLU activations to parameterize the discrete and continuous scores, respectively. QTD is trained by minimizing the discrete score entropy (Eq. 5), while DDPM is trained using the standard denoising diffusion objective. For the forward processes, we set the mixing time $T = 2.0$ for QTD and a 1000-step linear schedule for DDPM. We train both models for 20000 iterations using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 256.

Inference and Evaluation. During inference, we draw 5000 independent samples from each model. To achieve a similar generation performance, DDPM is executed with 1000 denoising steps (constant step size $\Delta t = 10^{-3}$), while QTD utilizes an early stopping threshold of $\delta = 10^{-4}$. The performance is quantified using the Total Variation (TV) distance between the empirical distribution $\hat{p}$ and the true discretized GMM p:

$$\mathrm{TV}(\hat{p}, p) = \frac{1}{2}\sum_{k=1}^{K} |\hat{p}_k - p_k|.$$

The TV distance is calculated across the 5000 samples for each of the 10 target distribution seeds.

Results. As shown in Figure 4, QTD demonstrates significant computational advantages. While both methods eventually converge to similar TV distances, QTD achieves rapid stabilization within the first 100–150 Number of Function Evaluations (NFE)*. In contrast, DDPM requires the full 1000 denoising steps to reach a comparable level of accuracy. This behavior is further illustrated in

*. Since QTD sampling terminates adaptively, the NFE varies across trajectories. The notation `mean max NFE` in Figure 4 denotes the mean (over 10 seeds) of the maximum NFE (over 5000 trajectories per seed) as the computational budget for each run.

the histogram evolution (Figure 5), where the QTD empirical distribution matches the gray target GMM much earlier in the generative process than the DDPM baseline.

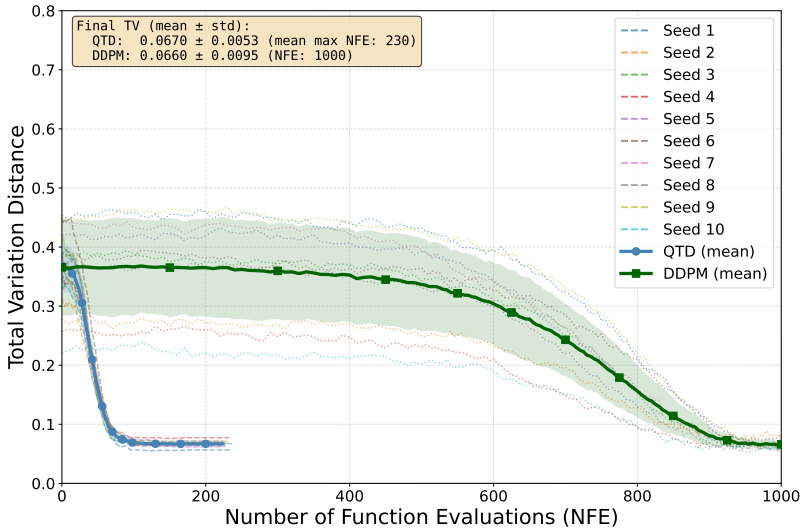

Figure 4: Convergence comparison between QTD and DDPM on a 1D Gaussian Mixture Model. We plot TV distance from the target distribution against NFE. Dashed lines show individual runs across 10 random seeds; solid lines show the mean with shaded standard deviation. QTD (blue) converges rapidly within the first 100 NFE and stabilizes around 150 NFE, whereas DDPM (green) requires the full 1000 denoising steps. Both methods achieve similar final TV distances, but QTD demonstrates significantly improved computational efficiency.

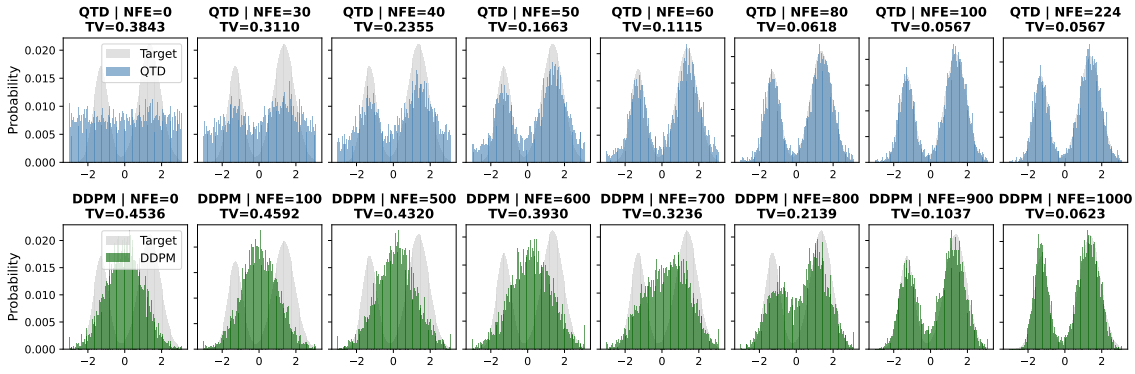

Figure 5: Histogram evolution comparison between QTD (first row) and DDPM (second row). Histograms show the empirical distribution (colored) versus the target Gaussian mixture model (gray) at various numbers of function evaluation (NFE) checkpoints. QTD achieves rapid convergence within around 100 NFE, while DDPM requires the full 1000 denoising steps to match the target distribution.

Algorithm 3 INFERENCE PROCESS WITH MULTIFLIP TRUNCATED UNIFORMIZATION

Require: Total time T, time partition $0 = t_0 < \ldots < t_W = T - \delta$, parameters $\beta_{t_1}, \ldots, \beta_{t_W}$ set as Eq. (14), reverse transition rate $\hat{R}_t^{\leftarrow}$ from score $\tilde{v}_{t,\boldsymbol{y}'}(\cdot)$, Poisson factor λ, flip candidates M.

Draw $\hat{\boldsymbol{y}}_{t_0} \sim \mathrm{Uniform}\big(\{0,1\}^{d \log_2 K}\big)$.
for $w = 1$ **to** W **do**
  Draw $N \sim \mathrm{Poisson}(\lambda \cdot \beta_{t_w}(t_w - t_{w-1}))$;
  Sample N points i.i.d. from $[t_{w-1}, t_w]$, sort as $\tau_1 < \tau_2 < \cdots < \tau_N$;
  Set $\mathbf{z}_0 = \hat{\mathbf{y}}_{t_{w-1}}$;
  for $n = 1$ **to** N **do**
    Evaluate $\hat{R}_{\tau_n}(\boldsymbol{z}_{n-1}+\boldsymbol{e}_i,\ \boldsymbol{z}_{n-1})$ for all i using reference state $\boldsymbol{z}_{n-1}$;
    Let $\hat{R}_{\tau_n}(\boldsymbol{z}_{n-1}) = \sum_i \hat{R}_{\tau_n}(\boldsymbol{z}_{n-1}+\boldsymbol{e}_i,\ \boldsymbol{z}_{n-1})$ and $p_n = \min\Big(1,\ \beta_{t_w}^{-1} \cdot \hat{R}_{\tau_n}(\boldsymbol{z}_{n-1})\Big)$;
    Sample M bit positions i.i.d. $i_1, \ldots, i_M \sim \mathrm{Cat}\left(\dfrac{\hat{R}_{\tau_n}(\boldsymbol{z}_{n-1}+\boldsymbol{e}_i,\ \boldsymbol{z}_{n-1})}{\hat{R}_{\tau_n}(\boldsymbol{z}_{n-1})}\right)$;
    Draw $A_1, \ldots, A_M \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(p_n)$;
    Collect accepted flip set $\mathcal{F}_n = \big\{i_m : A_m = 1,\ m = 1, \ldots, M\big\}$;
    Set $\mathbf{z}_n = \mathbf{z}_{n-1} \oplus \boldsymbol{f}_n$, where $\boldsymbol{f}_n[j] = \mathbf{1}[j \in \mathcal{F}_n]$;
  end for
  Set $\hat{\mathbf{y}}_{t_w} = \mathbf{z}_N$;
end for
Recover $\bar{\mathbf{y}} = \mathrm{vBin}^{-1}(\hat{\mathbf{y}}_{t_W})$ and draw $\hat{\mathbf{x}} \sim \mathrm{Uniform}(\mathrm{Cell}(\bar{\mathbf{y}}))$.
return $\hat{\mathbf{x}}$.

G.2. Experiments on MNIST

Data Setup. We use the standard MNIST training split (60000 images) for training and the test split (10000 images) for FID evaluation. Each 28×28 grayscale image is first downsampled to 16×16 via bilinear interpolation and then keep $K = 256$ discrete gray levels using uniformly spaced thresholds over $[0, 1]$. For training data construction, QTD uses the quantized pixel values directly as discrete tokens, whereas DDPM re-normalizes them to $[-1, 1]$ as continuous inputs. During FID evaluation, all generated images are passed through the same quantization pipeline prior to feature extraction to ensure a fair comparison.

Experimental Setup and Architecture. Both models share the same UNet2DModel as the backbone with channel widths $[128, 256, 512]$, two residual blocks per level, a time-embedding dimension of 256, and a single self-attention layer at the penultimate encoder resolution (4×4). QTD parameterizes per-bit log-scores and is trained by minimizing the discrete score entropy loss, while DDPM is trained with the standard MSE ϵ-prediction objective. For the forward process, QTD uses a discrete uniform noising process with mixing time $T = 1.0$ and start time $t_0 = 0.0$; DDPM uses a 7200-step linear Gaussian noise schedule. Both models are trained for 10 epochs using the AdamW optimizer with a learning rate of 2×10^{-4}, a cosine warm-up of 500 steps, a batch size of 256, and gradient clipping at 1.0.

Inference and Evaluation. During inference, we draw 10,000 independent samples from each model. DDPM is executed with 1200, 2400, and 3600 denoising steps. QTD is evaluated under QTD-MultiFlip (Alg. 3), which uses multi-flip Truncated Uniformization with M simultaneous bit

flips and a Poisson acceleration factor of λ; use an early-stopping threshold of δ. Image quality is quantified using the Fréchet Inception Distance (FID) between the empirical generated distribution $\hat{p}$ and the real MNIST test distribution p:

$$\mathrm{FID}(\hat{p},\, p) \;=\; \left\|\mu_{\hat{p}} - \mu_p\right\|^2 \;+\; \mathrm{Tr}\Big(\Sigma_{\hat{p}} + \Sigma_p - 2(\Sigma_{\hat{p}}\Sigma_p)^{1/2}\Big)\,, \tag{55}$$

where μ and Σ are the mean and covariance of InceptionV3 pool_3 activations (dimension 2048) computed over 10000 images.

Results. Table 4 reports FID scores computed from 10000 generated samples at convergence. QTD-MultiFlip (Alg. 3) consistently achieves the lowest FID across all inference budgets, demonstrating that the multi-flip proposal accelerates convergence to the target distribution. Even with a modest number of inference steps, QTD matches or surpasses DDPM in generation quality while requiring significantly fewer network function evaluations, consistent with the computational advantages observed in the 1D synthetic setting.

Table 4: FID comparison on MNIST between DDPM and QTD (lower is better). All scores are computed from 10000 generated samples.

	FID (↓)	
Inference Steps	DDPM (Ho et al., 2020)	QTD (Alg. 3)
1200	4.358	**4.147**
2400	4.397	**4.134**
3600	4.330	**4.081**

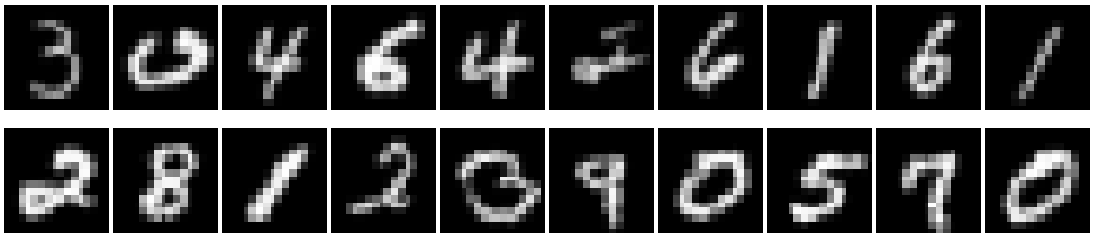

Figure 6: Random samples generated on MNIST. Top row: DDPM. Bottom row: QTD.