FROM SCORES TO GIBBS CORRECTORS: ACCELERATING UNIFORM-RATE DISCRETE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete diffusion models have achieved strong empirical performance in text and other symbolic domains, but, especially for uniform-rate models, they often require many steps to generate a single sample. Existing acceleration methods either rely on training additional quantities or suffer from slow mixing. In this work, we propose a novel Gibbs-based corrector for discrete diffusion models, termed Gibbs-Accelerated Discrete Diffusion (GADD). GADD leverages the structure of the concrete score function to construct Gibbs posterior likelihoods directly, without requiring any additional training beyond standard score estimation. We show that GADD achieves an overall sampling complexity of $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$, yielding the first such rate for diffusion-based samplers for uniform-rate discrete diffusion models. We also conduct numerical experiments demonstrating the practical advantages of GADD across synthetic data, zero-shot text sampling, and zero-shot conditional music generation. These results corroborate the theory and show that GADD consistently improves sample quality and wall-clock efficiency over standard baselines, including vanilla Euler methods and CTMC correctors. Beyond this, our theoretical analysis introduces a novel framework for analyzing predictor-corrector methods in discrete diffusion models, which may be of independent interest. Unlike existing approaches that rely on the Girsanov change-of-measure technique, our method is based on an induction argument that tracks error propagation across predictor iterations while accounting for inaccuracies in the corrector updates.

1 INTRODUCTION

Generative modeling is a central problem in deep learning, aiming to learn distributions that closely match observed data and enable high-quality sample generation. In recent years, diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Austin et al., 2021) have emerged as a powerful and flexible generative framework, achieving state-of-the-art performance across a wide range of applications. Among them, **discrete** diffusion models operate directly over discrete sample spaces and have proven highly effective for discrete-data domains. They have achieved strong empirical performance in natural language processing (Nie et al., 2025), graph generation (Vignac et al., 2023; Liu et al., 2023), musical note generation (Campbell et al., 2022), and molecular and drug design (Alakhdar et al., 2024; Nisonoff et al., 2025), among others.

Despite their empirical success, diffusion models face a major bottleneck in generation efficiency. Unlike classical generative models such as GANs, diffusion models typically require a large number of iterative steps to generate a single sample. A substantial body of work has explored accelerated sampling methods for *continuous* diffusion models, including predictor-corrector schemes (Song et al., 2021), Runge-Kutta methods (Wu et al., 2024), Hessian-based approaches (Liang et al., 2025a), and, more recently, successive refinement techniques (Li et al., 2025). In contrast, relatively few studies have focused on accelerating *discrete* diffusion models (see Section 1.2 for details).

Among the two canonical classes of *discrete* diffusion models, *masked (i.e., absorbing-rate)* diffusion models (MDMs) exhibit unique structure induced by the masking dynamics: once a token reaches the absorbing masked state it remains fixed thereafter, which facilitates the development of faster samplers through structured unmasking steps. This built-in structure can be exploited to design efficient samplers. For example, Zheng et al. (2025) showed that MDMs admit an analytic characterization of sampling times, requiring at most d steps to generate high-quality samples, and Zhao et al. (2025) exploited informed correctors that are specially designed for MDMs. Despite the strong language modeling performance of masked diffusion models (MDMs), *uniform-rate* discrete diffusion models

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

remain widely used in domains such as graph and molecule generation, where masking-based approaches can suffer from state-clashing and produce invalid structures (e.g., molecules violating chemical rules) (Seo et al., 2025; Zhang et al., 2025a). More recently, Deschenaux et al. (2026) showed that uniform-rate models augmented with predictor-corrector methods can outperform both autoregressive (AR) and mask diffusion models in few-step text generation.

Despite their promise, accelerating uniform-rate models remains challenging and comparatively underexplored, as, unlike masking-based designs, they lack salient structure that can be readily exploited for acceleration. From a theoretical perspective, random-step samplers (e.g., uniformization) can simulate the reverse process exactly up to score-estimation error, but their realized step counts can be unbounded. For deterministic-step samplers, existing total-variation complexity bounds scale only polynomially in ε, where ε is the target accuracy (see Table 1 for details): $\mathcal{O}(\varepsilon^{-1})$ for exact and τ-leaping methods (Campbell et al., 2022; Zhang et al., 2025b), $\mathcal{O}(\varepsilon^{-2})$ for Euler and Tweedie τ-leaping (Liang et al., 2025c). More recently, Ren et al. (2025b) proposed an accelerated approach achieving an improved rate of $\mathcal{O}(\varepsilon^{-1})$ by leveraging higher-order numerical solvers to more accurately approximate the reverse process, inspired by advances in higher-order numerical solvers (Wu et al., 2024). Overall, existing samplers for uniform-rate models achieve only *polynomial* dependence on ε, which naturally raises the following question:

Question: Can we break the $\mathcal{O}(\mathrm{poly}(\varepsilon^{-1}))$ barrier and achieve a logarithmic dependence, namely $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$, for uniform-rate discrete diffusion models?

Our study provides affirmative answer to this question.

1.1 Our Contributions

In this paper, we propose the *first* uniform-rate discrete-diffusion-based sampler that achieves $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$ convergence rate. Our detailed contributions are as follows.

Algorithmic Design: Our key contribution is to propose a novel Gibbs-based corrector, the Gibbs-Accelerated Discrete Diffusion (GADD) algorithm, for sampling from uniform-rate discrete diffusion models. Notably, leveraging properties of the existing concrete score function, our GADD algorithm *does not need extra-training* beyond the score functions. Instead, the posterior likelihood required for the Gibbs update can be directly obtained from the score estimator.

Theoretical Analysis: We theoretically show that GADD achieves an overall sampling complexity of $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$, improving upon existing diffusion-based sampling methods for uniform-rate discrete diffusion models. In comparison, all previous samplers only achieve the convergence rate of $\mathcal{O}(\mathrm{poly}(\varepsilon^{-1}))$. In particular, we show that the diffusion process naturally provides a *warm-start* to apply Gibbs correctors; such a warm-start simultaneously reduces both the Gibbs convergence error (under perfect per-step update) and the error from inaccurate score estimation.

Numerical Experiments: We further conduct numerical experiments demonstrating the practical advantages of GADD across synthetic data, zero-shot unconditional text sampling, and zero-shot conditional music generation. These results corroborate our theoretical findings and show that GADD achieves superior efficiency and robustness compared to previous approaches. In particular, GADD consistently improves sample quality at the same NFE budget, performs well on challenging spiky target distributions, and yields favorable wall-clock efficiency in realistic discrete-generation settings.

General Framework for Predictor-Corrector Sampling: Overall, we develop a general framework for analyzing predictor-corrector methods in discrete diffusion models, based on an induction argument that tracks error propagation across iterations while accounting for inaccurate corrector updates. By decomposing the error into initialization, mixing, and estimation components, the framework provides explicit conditions for controlling global error without stronger score accuracy requirements than in vanilla diffusion models. Beyond GADD, it also applies to the classical CTMC corrector (Campbell et al., 2022), offering a unified view of its convergence under approximate updates.

1.2 Related Works

In this subsection, we focus on related works that provide convergence guarantees for uniform-rate discrete diffusion models. Please see Appendix A for additional related works.

Convergence Theory on Uniform-Rate Discrete Diffusion Samplers. A growing body of work studies the number of steps required for samplers to achieve ε-TV guarantees (see Table 1). The

Algorithm	Number of steps	Paper Reference
Exact	$\widetilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\varepsilon}\right)$	Zhang et al. (2025b)
τ-leaping	$\widetilde{\mathcal{O}}\left(\frac{d^4}{\varepsilon}\right)$	Campbell et al. (2022)
	$\widetilde{\mathcal{O}}\left(\frac{d^2}{\varepsilon^2}\right)$	Ren et al. (2025a)
	$\widetilde{\mathcal{O}}\left(\frac{d}{\varepsilon^2}\right)$	Dmitriev et al. (2026)
Euler Method and Tweedie τ-leaping	$\widetilde{\mathcal{O}}\left(\frac{d^2}{\varepsilon^2}\right)$	Liang et al. (2025c)
DMPM (and its variants)	$\widetilde{\mathcal{O}}\left(\frac{d}{\varepsilon^4}\right)$	Conforti et al. (2025) Pham et al. (2025)
θ-RK-2 and θ-Trapezoidal	$\widetilde{\mathcal{O}}\left(\frac{\mathrm{poly}(d)}{\varepsilon}\right)^{\dagger}$	Ren et al. (2025b)
GADD (ours)	$\widetilde{\mathcal{O}}\left(d\rho_*^{-1}\log^2(dS/\varepsilon)\right)$	(This paper, Thm 1)

Table 1: *Summary of results for uniform-rate discrete diffusion samplers in terms of the number of steps needed to achieve $\mathcal{O}(\varepsilon)$ accuracy in $\mathrm{TV}(q_\delta, \hat{p}_{T-\delta})$ (or $\mathcal{O}(\varepsilon^2)$ for $\mathrm{KL}(q_\delta \| p_{T-\delta})$), where q_δ is a perturbed distribution satisfying $\mathrm{TV}(q_0, q_\delta) \lesssim d\delta$. We only display those results having deterministic number of step-sizes. Here, d denotes the dimension (e.g., length of the generated sentence), and ρ_* is the path-wise worst-case spectral gap for Gibbs samplers (see discussions in Sections 3.3 and 3.4). † The $\mathrm{poly}(d)$ dependence is what we infer, which is not provided in their paper. Overall, our result achieves the first $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$ convergence rate when q_0 is connected, whereas all prior works with deterministic step-complexity achieve polynomial dependence on ε^{-1}.*

earliest such result, Campbell et al. (2022), analyzed τ-leaping under the TV metric. Subsequent studies considered two classes of samplers. For random-step methods, Ren et al. (2025a); Chen & Ying (2024) established guarantees for the uniformization sampler, though the realized number of steps may be unbounded. For deterministic-step methods, Zhang et al. (2025b) obtained strong guarantees under the assumption of access to a perfect per-step solver, while Conforti et al. (2025); Pham et al. (2025) achieved low dependence on d via the DMPM sampler at the cost of a high ε dependence. Ren et al. (2025a); Dmitriev et al. (2026) further improved guarantees for τ-leaping, and more recently, Liang et al. (2025c) provided the first analysis of the Euler method and Tweedie τ-leaping. Notably, all these analyses of deterministic-step samplers yield $\mathcal{O}(\mathrm{poly}(\varepsilon^{-1}))$ convergence rates. Concurrent with our work, Kan et al. (2026) derives an S-independent upper bound on the error under the assumption of exact simulation of the continuous-time sampling process; however, the corresponding step complexity when there is discretization remains unclear.

Acceleration of Uniform-Rate Discrete Diffusion Samplers. Compared to the extensive literature on standard samplers, relatively few works study the acceleration of uniform-rate discrete diffusion samplers. Several empirical approaches have been proposed to accelerate discrete diffusion models. Campbell et al. (2022) first introduced a predictor-corrector scheme with a CTMC corrector that moves the particle toward the target distribution, and Gat et al. (2024) generalized this forward-backward idea and developed a unified corrector especially for discrete flow-matching. More recently, Deschenaux et al. (2026) developed a family of Ψ-samplers that achieve strong performance in both language and image modeling. Concurrent work by Gourevitch et al. (2026) studies leave-one-out parameterizations of uniform diffusion and their application to Gibbs-based predictor-corrector sampling. These works formulate their correctors through posterior parameterizations (Deschenaux et al., 2026; Gat et al., 2024; Gourevitch et al., 2026), whereas our work constructs Gibbs conditional estimators directly from concrete scores and establishes non-asymptotic convergence guarantees under score-estimation error. On the theoretical side, Ren et al. (2025b) accelerated discrete diffusion models using higher-order numerical methods based on Runge-Kutta schemes, and achieves a convergence rate of $\mathcal{O}(\varepsilon^{-1})$ in terms of the target accuracy ε.

2 Preliminaries

In this section, we review both discrete diffusion models and Gibbs samplers, the two major components in the work.

2.1 Discrete Diffusion Sampling

Discrete diffusion models are defined by a forward noising process and a corresponding reverse denoising process involving *discrete* data. The forward process is formulated as a continuous-time

Markov chain (CTMC) on the discrete state space $\mathcal{X} = [S]^d$, where d denotes the number of tokens and each token takes values from a vocabulary of size S. Let $x_0 \in [S]^d$ denote the initial data sample, distributed according to the probability mass function (p.m.f.) q_0. Let $R_t \in \mathbb{R}^{S^d \times S^d}$ be the (time-dependent) rate matrix of the CTMC. For any two states $x, y \in [S]^d$, the entry $R_t(x, y)$ specifies the instantaneous transition rate from x to y at time t. The conditional distribution of the state at time $t + \Delta t$ given the state at time t satisfies

$$q_{t+\Delta t|t}(y \mid x) = \mathbb{1}\{y = x\} + R_t(x, y)\,\Delta t + o(\Delta t), \tag{1}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. Validity of the CTMC requires that $R_t(x, y) \geq 0$ for all $x \neq y$ and $\sum_y R_t(x, y) = 0$ for all x. Among the many choices of R_t, one typical choice is the "uniform-rate" matrix (Lou et al., 2024). Specifically, the "uniform-rate" is defined as

$$R_t(x, y) = \begin{cases} \frac{1}{S} & \text{if } \mathrm{Ham}\,(x, y) = 1 \\ 0 & \text{if } \mathrm{Ham}\,(x, y) \geq 2 \end{cases},$$

where $\mathrm{Ham}\,(x, y)$ denotes the Hamming distance between x and y. With such an R_t, the tokens would evolve independently and homogeneously for each token (Campbell et al., 2022; Lou et al., 2024). Also, one can show that $q_T \approx \mathrm{Uniform}([S]^d)$ and the mixing speed is exponentially fast (Zhang et al., 2025b).

To sample from the diffusion model, one typical approach is to walk through the reverse process, which is defined as the exact time reversal of the forward CTMC with initial distribution equal to q_T (Campbell et al., 2022; Kelly, 2011). By Campbell et al. (2022), the reverse process is itself a CTMC whose transition rate is $\overleftarrow{R}_t(x, y) := R_t(y, x)q_t(y)/q_t(x)$, which depends on the density ratio for all Hamming-distance-1 pairs (x, y), a quantity commonly referred to as the *(concrete) score*. Since this quantity is generally intractable, we estimate it using a neural network, with one popular choice of loss function: the score-entropy loss (Lou et al., 2024). After we have obtained an approximate score, for practical sampling, we can discretize the continuous-time reverse process and employ approximate sampling methods for CTMC, such as the Euler method (Lou et al., 2024). More specifically, define the estimated reverse rate as $\hat{R}_t(x, y) := R_t(y, x)s_t(y, x)$. Upon initializing $x_{t_N} \sim \mathrm{Uniform}([S]^d)$ (with $t_N = T$), the Euler method is given by, for each $k = N - 1, \ldots, 0$,

$$x^i_{t_k} = \begin{cases} a, & \text{w.p. } \hat{R}^i_{k+1}(x^i_{t_{k+1}}, a)(t_{k+1} - t_k),\ \forall a \neq x^i_{t_{k+1}} \\ x^i_{t_{k+1}}, & \text{w.p. } 1 + \hat{R}^i_{k+1}(x^i_{t_{k+1}}, x^i_{t_{k+1}})(t_{k+1} - t_k) \end{cases}, \tag{2}$$

where $\hat{R}^i_{k+1}(x^i, a) := \hat{R}_{t_{k+1}}(x, x^{-i} \oplus_i a),\ \forall a \neq x^i$ and $\hat{R}^i_{k+1}(x^i, x^i) := -\sum_{a \neq x^i} \hat{R}^i_{k+1}(x^i, a)$. To guarantee the validity of Euler steps, whenever $-(t_{k+1} - t_k)\hat{R}^i_{k+1}(x^i, x^i)$ exceeds a small probability $c \in (0, 1)$, we threshold it to c by proportionally rescaling the off-diagonal probabilities.

2.2 Random-Scan Gibbs Samplers

In the sampling literature, Gibbs sampling is among the most popular methods due to its simplicity, exactness of conditional updates, and ease of implementation. One typical Gibbs samplers is the (single-site) random-scan Gibbs sampler, where an update index is drawn randomly at each step. For a target distribution π supported on $\mathcal{X}$, define the one-step Markov transition kernel for random-scan Gibbs sampler as

$$P(x, y) := \textstyle\sum_{i=1}^d w_i \mathbb{1}\left\{x^{-i} = y^{-i}\right\} \pi^i(y^i | x^{-i}). \tag{3}$$

Here $w_i > 0$ is some fixed (i.e., independent of the current state and the target accuracy level) weight function over the indices that sums up to 1. For random-scan Gibbs samplers, such a kernel enjoys many regularity conditions of Markov operators. To name a few: first, given its probabilistic nature, P is aperiodic. Also, if $\mathrm{supp}(\pi)$ is connected in Hamming distance, P is irreducible on that support. Given these two properties, running such a Markov chain would yield a unique stationary distribution, which can be shown as π itself. Note here that on a diffusion path, $q_t(x) > 0$ for all $t > 0$.

One way to characterize the convergence of such a Gibbs operator is through its spectral gap. Define p_ℓ as the probability at the ℓ-th Gibbs iteration (with $\ell = 1, \ldots, L$). Indeed, the convergence rate can be characterized as

$$\chi^2(p_\ell \| \pi) \leq (1 - \rho)^\ell \chi^2(p_0 \| \pi), \tag{4}$$

where $\chi^2(\cdot\|\cdot)$ is the χ^2 divergence. For a full review of Gibbs samplers (Glauber dynamics), we refer readers to Levin et al. (2017). Finally, in the context of diffusion models when $\pi = q_t$, we write the spectral gap of the underlying Gibbs operator as ρ_t.

2.3 List of Notations

Let $x^i (1 \le i \le d)$ denote the i-th element of a vector $x \in [S]^d$ and $x^{-i} \in [S]^{d-1}$ denote the vector with the i-th element removed. Define $\mathrm{Ham}(x, y)$ as the Hamming distance between two vectors x and y. For a positive integer n, $[n] := \{1, \dots, n\}$. Write $\mathrm{supp}(q)$ as the support of a distribution q.

3 Discrete Diffusion Sampling with Gibbs Correctors

In this section, we propose and analyze a novel Gibbs corrector for sampling from the uniform-rate CTMC, which we call the Gibbs-accelerated Discrete Diffusion (GADD) sampler.

3.1 The GADD Algorithm

We summarize our algorithm in Algorithm 1 below.

Algorithm 1 consists of two components: an outer loop and an inner loop. The outer loop defines a sequence of target distributions q_{t_k}, each corresponding to a perturbed data distribution induced by the forward diffusion process. For each t_k, the inner loop runs a Gibbs sampler for L_k steps targeting q_{t_k}. In practice, setting $L_k = 0$ corresponds to performing no correction at the current step. In particular, when $L_k = 0$ for all k and the Euler update is applied throughout, Algorithm 1 reduces to the standard Euler sampler defined in (2). We also note that the weights w_i in the algorithm can be chosen in an informed manner, similar to Zhao et al. (2025). Finally, we provide a variant of Algorithm 1 in Appendix B, which employs systematic-scan Gibbs updates rather than random-scan.

Algorithm 1: Gibbs-Accelerated Discrete Diffusion (GADD)

Input: initialization $x_{t_N} \sim p_{\text{init}} = \mathrm{Uniform}([S]^d)$, discretization points $\{t_k\}_{k=0}^N$ (with $t_N = T$ and $t_0 = \delta$), estimated score s_t, correction steps $\{L_k\}_{k=0}^N$

for $k = N-1$ **to** 0 **do**
  $z_0 =$ the output of the Euler update given $x_{t_{k+1}}$ (using (2)) ;
  for $\ell = 1$ **to** L_k **do**
    Sample i w.r.t. weight w_i over $[d]$;
    Construct posterior $\hat{q}^i_{t_k}(\cdot | z^{-i}_{\ell-1})$ (using (5));
    Update $z^i_\ell \sim \hat{q}^i_{t_k}$ while fixing $z^{-i}_\ell = z^{-i}_{\ell-1}$;
  end
  $x_{t_k} = z_{L_k}$;
end
Return: x_{t_0}

Intuitively, the predictors and correctors in GADD benefit each other. On the one hand, Gibbs corrector steps geometrically contract the sampling error, whereas such contraction is not generally guaranteed for vanilla predictor steps (Zhang et al., 2025b; Liang et al., 2025c; Ren et al., 2025b). On the other hand, Euler predictors update all coordinates in parallel and transport samples through intermediate targets, providing a *warm-up* mechanism that can reduce the initial mismatch and the number of sequential Gibbs updates needed at each stage.

3.2 Constructing Posterior Likelihood from Score

To utilize the Gibbs corrector for target q_t (where $t \in [\delta, T]$), one needs to have access to $q^i_t(\cdot | x^{-i})$ for all $t \in [\delta, T]$ and $i \in [d]$. Our key observation is that such information is readily available in the score function, which is the density ratio between all Hamming-distance 1 pairs. Notice that

$$q^i_t(x^i | x^{-i}) = \frac{\frac{q_t(x^i | x^{-i})}{q_t(y^i | x^{-i})}}{\sum_{x^i \in [S]} \frac{q_t(x^i | x^{-i})}{q_t(y^i | x^{-i})}} = \frac{\frac{q_t(x)}{q_t(x^{-i} \oplus_i y^i)}}{\sum_{x^i \in [S]} \frac{q_t(x)}{q_t(x^{-i} \oplus_i y^i)}}, \ \forall y^i \in [S].$$

Here the last equality follows from Bayes' rule. Thus, one applicable estimator is

$$\hat{q}^i_t(x^i | x^{-i}) = \frac{1}{S} \sum_{y^i \in [S]} \frac{s_t(x, x^{-i} \oplus_i y^i)}{\sum_{x^i \in [S]} s_t(x, x^{-i} \oplus_i y^i)}, \tag{5}$$

where we have taken an average over y^i. Here, with a slight abuse of notation, we define $s_t(x, x) \equiv 1$ (which does not require any estimation). We have provided more estimator candidates in Appendix B.

To clarify, sampling from the averaged posterior in (5) does *not* require S forward calls: we can first draw an anchor y^i uniformly from $[S]$ and evaluate the score model at $x^{-i} \oplus_i y^i$. Indeed, with the implementation of the well-known Score Entropy Discrete Diffusion (SEDD) model (Lou et al., 2024), a *single* forward output of the score function is of size (B, d, S) (where B is the batch size), which contains *all* the corresponding scores (i.e., density ratios) used to aggregate. Thus, per Gibbs-step, we only need to call the score function *once* to get an estimate of the posterior.

Such a construction enables direct evaluation of the posterior probabilities required by the Gibbs sampler, *which does not require any extra-training and can be directly used for acceleration*. In contrast, previous correctors that employ Metropolis-Hastings algorithms (Huang et al., 2024) require an additional estimator for the likelihood ratio between all pairs of (x, y), which is both computation- and memory-intensive given that $x, y \in [S]^d$ is high dimensional. With the estimator in (5), our Gibbs corrector avoids such auxiliary estimation and remains scalable to high-dimensional domains.

3.3 Convergence Analysis of GADD

We first define a few useful notations. Let $p_{t_k,\ell}$ denote the sampling distribution at outer-loop step $k = N-1, \ldots, 0$ (with target q_{t_k} and at forward time t_k) and at inner-loop step $\ell = 1, \ldots, L_k$, without estimation error. The following assumptions are needed for the score estimate.

Assumption 1. Fix $t_k \in [\delta, T]$ and $\ell \in \{0, \ldots, L_k\}$. Suppose that score estimate satisfies that

$$\mathbb{E}_{x\sim p_{t_k,\ell}} \sum_{y\neq x} R_{t_k}(y,x) \left(s_{t_k}(y,x) - \frac{q_{t_k}(y)}{q_{t_k}(x)} - \frac{q_{t_k}(y)}{q_{t_k}(x)} \log \frac{s_{t_k}(y,x)}{q_{t_k}(y)/q_{t_k}(x)} \right) \leq \varepsilon_{\text{est}}^2,$$

where $p_{t_k,\ell}$ is the sampling distribution on the clean path.

Here we have used the score-entropy error typically employed in the discrete diffusion literature Zhang et al. (2025b); Ren et al. (2025a). The only tweak is that the expectation is to be taken under $p_{t_k,\ell}$. Indeed, $p_{t_k,\ell}$ is a slightly perturbed version of the true target distribution q_{t_k}, which is likely to be satisfied when we use a warmed Gibbs corrector, as is the case for GADD. Later in Section 3.4, we will show that the clean-path estimation error is sufficient for targets having full support.

The following theorem summarizes the main convergence guarantees of GADD.

Theorem 1 (Convergence of GADD)**.** *Suppose that Assumption 1 is satisfied. Let* $T \asymp \log(d/\varepsilon^2)$, $\varepsilon_{est}^2 = \mathcal{O}\left(\frac{\varepsilon^2}{T+\sum_{k=0}^{N-1} L_k}\right)$, $\delta \asymp \frac{\varepsilon}{d}$, *and* $\kappa \asymp \left(\log\log(d/\varepsilon^2)\right)^{-1}$. *Then, choosing* $t_{k+1} - t_k \leq \kappa \min\{1, t_{k+1}\}$, *the total number of steps to achieve* ε *TV error for Algorithm 1 suffices to have*

$$N_{total} \lesssim \frac{d \log^2(dS/\varepsilon)}{\rho_*},$$

where $\rho_* = \inf_{t\in[\delta,T]} \rho_t$ *(see a further characterization in Corollary 1). Note that we have omitted lower-order factors.*

Theorem 1 captures the complexity for the total number of steps, including those of the predictors and the correctors. The proof is in Appendix D. At a high level, the diffusion process and the Gibbs sampler benefit from each other in GADD. First, repeated Gibbs correction geometrically contracts the sampling error under exact scores, yielding the bound established in Theorem 1. On the other hand, the diffusion process transports samples through a sequence of smoothed target distributions, with predictor steps updating all coordinates in parallel. This can substantially reduce the number of sequential Gibbs updates required for sampling. This is the goal of Section 3.5, where we will illustrate the advantage of GADD over vanilla Gibbs samplers.

Beyond GADD, our recursion framework also applies to other correctors, such as CTMC correctors (Campbell et al., 2022), with the analysis provided in Appendix E. We also note that for the sampler in Campbell et al. (2022), although the underlying continuous-time dynamics contracts the sampling error exponentially, their implementations introduce additional approximation error through time discretization, which results in sampling complexity bound retains polynomial dependence on ε^{-1}.

In general, ρ_* may depend on δ and hence on ε. However, when $\mathrm{supp}(q_0)$ is connected, ρ_* admits a positive lower bound independent of ε, which is characterized in the following corollary.

Corollary 1 (GADD under connected support)**.** *Suppose the same conditions of Theorem 1. Then* ρ_* *has a positive* ε*-independent lower bound if and only if* $\mathrm{supp}(q_0)$ *is connected. In such a case, we have* $N_{total} \lesssim \mathrm{polylog}(\varepsilon^{-1})$.

Corollary 1 provides the *first* convergence guarantee for diffusion-based algorithms to achieve $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$ convergence rate when q_0 is connected, which shows the effectiveness of predictor-corrector schemes with Gibbs correctors. In particular, for sampling from well-structured target distributions (e.g., high-temperature Ising models) where the spectral gap satisfies $\rho_* = \Omega(\mathrm{poly}^{-1}(d))$, the total number of sampling steps admits an explicit dependence on the dimension: $N_{\text{total}} = \mathcal{O}\big(\mathrm{poly}(d)\,\mathrm{polylog}(\varepsilon^{-1})\big)$. This is a substantial improvement over prior acceleration methods with theoretical guarantees (Ren et al., 2025b). The proof of Corollary 1 is in Appendix F.

We note that the connectivity of q_0 is a mild structural condition, i.e., any two states in $\mathrm{supp}(q_0)$ must be connected by a sequence of single-coordinate changes that remain within the support. In particular, it allows zero-probability configurations and is weaker than full support. Connectivity, or stronger versions requiring connectivity after conditioning, is a standard assumption in the Gibbs sampling literature (Levin et al., 2017; Chen et al., 2026b). In our analysis, connectivity is necessary and sufficient for $\rho_0 > 0$ (Lemma 7), and hence for a positive lower bound on ρ_* that is uniform in ε (Lemma 8). This is the condition that yields the $\mathrm{polylog}(\varepsilon^{-1})$ dependence in Theorem 1 through our spectral-gap analysis.

3.4 Convergence Under Full Support Target

We now consider targets with full support, which allows us to measure score-estimation error under the true diffusion marginals and to run the sampler to $t = 0$. Specifically, suppose that $0 < l \le q_0(x) \le u$ for all $x \in [S]^d$. We use the following score-estimation assumption.

Assumption 2. Fix $t_k \le T$. Suppose that score estimate satisfies that

$$\mathbb{E}_{x\sim q_{t_k}} \sum_{y\neq x} R_{t_k}(y,x)\left(s_{t_k}(y,x) - \frac{q_{t_k}(y)}{q_{t_k}(x)} - \frac{q_{t_k}(y)}{q_{t_k}(x)} \log \frac{s_{t_k}(y,x)}{q_{t_k}(y)/q_{t_k}(x)}\right) \le \varepsilon_{\text{est}}^2.$$

Under this condition, the exact score $q_0(y)/q_0(x)$ is well-defined and lies in $[l/u, u/l]$ (Zhang et al., 2025b). This also permits evaluating the Gibbs corrector directly at $t = 0$. Moreover, $q_t(x) \in [l, u]$ for all $t \ge 0$ since the uniform-rate CTMC forward process is an averaging process, so the same score bounds hold throughout the diffusion process.

Theorem 2 (Convergence under full support)**.** *Suppose that $0 < l \le q_0(x) \le u$ for all $x \in [S]^d$, such that $u/l = \mathcal{O}(1)$. Suppose that Assumption 2 holds. With the same choice of T and κ, suppose $\varepsilon_{est}^2 = \mathcal{O}\left(\frac{\varepsilon^2}{\left(N+\sum_{k=0}^{N-1} L_k\right)^2}\right)$, and choose $\delta = 0$, $w_i = d^{-1}$, and constant step-sizes: $t_{k+1} - t_k \le \kappa$. Then, the total number of steps to achieve ε TV error for Algorithm 1 suffices to have $N_{total} \lesssim d\log^2(d/\varepsilon)$ (ignoring lower-order factors).*

The proof is in Appendix G. The intuition behind Theorem 2 has two components. First, full support and forward diffusion smoothing allow us to control the error in χ^2 divergence due to the predictor step, effectively bringing a warm-up effect (Lemma 9). A direct change-of-measure argument then converts the score-estimation guarantee under q_t into the TV error bounds along the exact-score reference sampling path. Second, the probability bounds allow comparison with the uniform product distribution, yielding a spectral-gap bound uniform over diffusion time (Lemma 10). Combining these two ingredients yields the stated guarantee.

3.5 Comparison with Vanilla Gibbs

Since Gibbs sampling can directly sample from a target distribution, a natural question is what benefit GADD offers by incorporating Gibbs updates together with diffusion predictors. In this subsection, we show that GADD can yield a factor-of-d improvement in sampling complexity over vanilla random-scan Gibbs, via a simple target whose dynamics can be analyzed explicitly. To this end, we assume perfect score estimation and uniform coordinate-selection weights $w_i = 1/d$. Consider the clean target $q_0 := \boldsymbol{\delta_0}$, which is a singleton at the all-zero point $\mathbf{0} = (0, \ldots, 0) \in [S]^d$ (without loss of generality). To mitigate the effect on initialization, we let both GADD and the vanilla Gibbs be initialized at $\mathrm{Uniform}([S]^d)$. While the Gibbs operator does not exist at $t = 0$, both GADD and the vanilla Gibbs are targeting at q_δ, and we let $\delta \asymp \varepsilon/d$ according to Theorem 1.

Proposition 1 (TV lower bound for vanilla Gibbs)**.** *Let $q_0 = \boldsymbol{\delta_0}$. Fix $\delta > 0$. Write p_B^{Gibbs} as the output distribution of vanilla Gibbs using the Gibbs operator P_δ (targeting at q_δ) at iteration B. Then, in order to achieve $\mathrm{TV}(p_B^{\mathrm{Gibbs}}, q_\delta) \le \varepsilon$, when $\delta \asymp \varepsilon/d$, we must have $B \gtrsim d\log(d/\varepsilon)$.*

The proof of Proposition 1 is in Appendix H. We now consider the performance of GADD, as follows.

Theorem 3 (GADD on singleton target)**.** *Let $q_0 = \boldsymbol{\delta_0}$. Suppose that $T = t_N = 1$ and the step-sizes satisfies that $t_k - t_{k-1} = \frac{1}{4}t_k$ with $\kappa = \frac{1}{4}$ (and, say, $c = \frac{1}{2}$). Then, with any choice of $\{L_k\}_{k\in[N]}$, and assuming perfect estimators, the output of GADD satisfies $\mathrm{TV}(p_{\delta,L_0}, q_\delta) \le d\delta$. Thus, when $\delta \asymp \varepsilon/d$, if we let $L_k = \mathcal{O}(1)$ for all $k \in [N]$, the total number of steps is $N_{\mathrm{total}} \lesssim \log(d/\varepsilon)$.*

Comparing Proposition 1 and Theorem 3 shows that, for the singleton target, GADD reduces the sampling complexity by a factor of $\mathcal{O}(d)$ relative to vanilla Gibbs, even though the conditional Gibbs updates are assumed to be tractable. The underlying reason is that, the diffusion path supplies a sequence of smoothed targets, effectively providing a quantitative *warm-up* effect (see (21)).

The full proof of Theorem 3 is in Appendix I. The proof of Theorem 3 is nontrivial. Unlike previous KL-based analyses (Zhang et al., 2025b; Ren et al., 2025a), we employ a direct TV-based analysis. Different from the Euler upper-bound analysis in Liang et al. (2026) that starts from the rate of change of TV along a continuous-time trajectory, here we directly characterize the discretized predictor mismatch in (20) and relate it to TV distance through (18). Conditional on the selected-coordinate history, Gibbs updates cannot increase this mismatch or the resulting TV bound (see (22)). This approach may extend to broader target classes admitting analogous mismatch recursions.

4 Numerical Experiments for GADD

We include all experiment details in Appendix C.

4.1 Synthetic Experiment

We first numerically compare our GADD algorithm in Algorithm 1 on synthetic data against three baselines: (i) the Euler method, (ii) the Gibbs sampler, both of which are widely used and strong vanilla baselines for sampling from discrete data, and (iii) the θ-Trapezoidal method (Ren et al., 2025b), a recently proposed accelerated sampler for uniform-rate discrete diffusion models. We use an autoregressive target with memory $h = 2$ and a mixed-point target consisting of eight equally weighted random states with $d = S = 8$. Further setup details are in Appendix C.1.

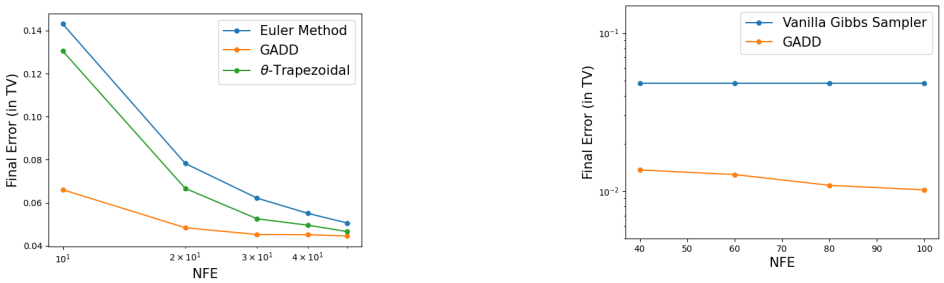

Figure 1: Comparison of GADD on synthetic data against the Euler method, the θ-Trapezoidal algorithm (left), and the Gibbs sampler (right).

We observe that, at the same number of function evaluations (NFEs), our GADD achieves a substantially smaller final error than both the Euler and θ-Trapezoidal methods. This empirical advantage aligns with our theoretical result in Corollary 1: to attain a final total variation error of ε, the required number of steps (and thus the NFE) scales as $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$, which is significantly lower than that of existing samplers, whose complexity scales as $\mathcal{O}(\mathrm{poly}(\varepsilon^{-1}))$.

We further find that our GADD performs particularly well on spiky mixed-point target distributions. In such settings, the vanilla Gibbs sampler is known to suffer from poor mixing behavior (Levin et al., 2017). In contrast, by leveraging the reverse diffusion process, GADD remains effective and achieves strong performance even under these challenging conditions. Although this differs from the singleton setting, the observed gains are consistent with the *warm-start* effect illustrated by Theorem 3.

Ablation Studies. *1. (Dimension and alphabet size)* We conduct additional exact-score experiments on mixed-point targets, fixing the number of mixture points at 16 and averaging over five random seeds. With $L_k = 1$, the TV error remains near 0.01 as we vary $d \in \{4, 6, 8, 10\}$ with $S = 6$, or $S \in \{4, 6, 8, 10\}$ with $d = 6$ (Table 4). These results demonstrate stable performance over the tested ranges. Further details, including an autoregressive-target ablation, are in Appendix C.1.1. *2. (Fresh versus stale scores)* We also vary L_k to compare recomputing scores at each correction step with reusing scores within the inner loop (Table 5). With $d = S = 6$, fresh-score GADD remains stable over the tested correction counts, whereas stale-score error increases at larger L_k. This ablation illustrates the need to tune the correction count when scores are reused.

4.2 Experiment on Text Data

We next consider sampling from text data. We first train a small SEDD uniform model on the WikiText103 dataset using the suggested hyperparameters as in Lou et al. (2024). We then compare the zero-shot sampling performance of GADD against several predictors, including the vanilla Euler method and the θ-Trapezoidal method, as well as the CTMC corrector, which we analyze

Method	NFE=64	NFE=128	NFE=256
Vanilla Euler (Lou et al., 2024)	285.1672 ± 35.1706 (19s)	283.0323 ± 24.3 (39s)	275.3841 ± 41.248 (79s)
θ-Trapezoidal (Ren et al., 2025b)	267.6957 ± 30.701 (23s)	255.1958 ± 6.6874 (48s)	265.9823 ± 18.50979 (97s)
CTMC corrector (Campbell et al., 2022)	272.653 ± 10.6409 (21s)	227.6811 ± 15.499 (44s)	219.8395 ± 24.515 (89s)
GADD (ours)	$\mathbf{192.1529 \pm 35.0539}$ (13s)	$\mathbf{161.5748 \pm 26.1866}$ (26s)	$\mathbf{149.6407 \pm 17.4786}$ (53s)

Table 2: Zero-shot generative perplexity (and time) across different methods. Best result in **bold**.

Method	Hellinger	Perplexity	Next-tokens Pred Err
Vanilla Euler (Lou et al., 2024)	0.1094 ± 0.0023	5.9453 ± 0.2200	0.3643 ± 0.0234
θ-Trapezoidal (Ren et al., 2025b)	0.1155 ± 0.0061	5.5532 ± 0.1552	0.3643 ± 0.0083
CTMC corrector (Campbell et al., 2022)	0.1510 ± 0.0064	11.5623 ± 0.7702	0.4108 ± 0.0096
Euler + GADD ($L = 5$)	0.1050 ± 0.0047	3.5496 ± 0.0999	0.3373 ± 0.0123
Euler + GADD ($L = 50$)	$\mathbf{0.1008 \pm 0.0073}$	$\mathbf{3.4522 \pm 0.0514}$	$\mathbf{0.3350 \pm 0.0087}$
Euler + GADD ($L = 100$)	0.1063 ± 0.0077	3.4543 ± 0.0630	0.3360 ± 0.0076

Table 3: Conditional generation qualities using Lakh pianoroll dataset (Dong et al., 2018).

in Appendix E. We follow the approach in Zhao et al. (2025) and fix the corrector score at each outer-loop, and we do not re-evaluate the score model within the corrector loop. The effect of this score-reuse approximation has been examined in the ablation study (see Appendix C.1.2 for details).

In Table 2, we report the generative perplexity of the zero-shot generated outputs at fixed NFEs.[1] Our proposed GADD method consistently achieves the best performance across all settings. In addition to this metric, we also report the actual wall-clock time. Notably, GADD exhibits lower wall-clock time compared to competing methods. This efficiency gain is likely due to differences in per-step computational cost: each Euler step updates all tokens simultaneously, whereas a random-scan Gibbs correction step updates only a single token per iteration. As a result, our method achieves a favorable trade-off between sample quality and computational efficiency.

4.3 Experiment on Monophonic Music Data

Beyond text data, we also conduct experiments on zero-shot conditional discrete diffusion models (Lou et al., 2024; Liang et al., 2025b) on monophonic music data, obtained from Dong et al. (2018) and following the pre-processing steps in Campbell et al. (2022). Different from unconditional sampling, the goal is to sample from the conditional distribution given part of the observable sequence, but in the zero-shot way without extra-training. In our setting, the goal is to complete a musical sequence given the first 100 notes. To evaluate the effect of correction, we compare the aforementioned methods, along with our GADD method with L correction steps (where $L \equiv L_k,\ \forall k$). We again fix the NFEs across all methods during evaluation. Also, in addition to standard metrics such as the mean Hellinger distance of the histogram (Campbell et al., 2022) and the generative perplexity (for which we trained an autoregressive model to evaluate), we introduce an additional metric: the mean prediction error over the next $F = 4$ tokens, designed to capture the short-term temporal consistency.

From Table 3, incorporating Gibbs correction steps yields significant improvements across all considered metrics. This highlights the generality of our Gibbs corrector, extending effectively even to conditional generation settings. We also observe that increasing the number of correction steps improves generation quality when L is moderate. However, the gains diminish as L grows, becoming marginal for larger values of L. In fact, the sample quality may saturate and even degrade as L becomes large, due to stale correctors and accumulated random noise.

5 Conclusion

In this paper, we have developed a high-accuracy sampling algorithm for uniform-rate discrete diffusion models with provable convergence guarantees. Remarkably, our GADD algorithm achieved $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$ convergence rate for connected targets. We also demonstrated the superiority of GADD on real-data experiments. An interesting direction for future work is to develop a theoretical understanding of general predictor-corrector methods and that for masked diffusion models.

[1] For example at NFE=32, the vanilla Euler method takes 32 predictor steps to complete, whereas GADD takes 16 predictor steps and 16 corrector loops.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, for proving part of the mathematical claims, and for implementing some of the baseline methods in the experiment. We have not used generative AI tools for generating synthetic data sets, for designing or providing feedback on research methodology or experiments, for cleaning and reformatting dataset, for supporting qualitative and thematic data analysis, or for interpreting results. The rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. Specifically, we have checked the correctness of the mathematical proofs and have read through the generated texts closely. We have also checked the baseline implementations are correct. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the assumptions underlying our theoretical results alongside their statements, with detailed proofs starting from Appendix D. Algorithm 1 specifies the GADD sampling procedure, with detailed construction of Gibbs correctors from estimated scores and the algorithmic variants. Section 4 and Appendix C document the synthetic, text, and conditional music experiments, including the data and model configurations, training and sampling hyperparameters, baseline implementations, evaluation metrics, and computational resources. Appendix C also explains the practical modifications used for efficient Gibbs updates, facilitating reproduction of the reported experiments. We have also provided the code for GADD and the baseline comparison methods on the text experiment as supplementary material.

REFERENCES

Amira Alakhdar, Barnabas Poczos, and Newell Washburn. Diffusion models in de novo drug design. *Journal of Chemical Information and Modeling*, 2024.

Filippo Ascolani and Giacomo Zanella. Dimension-free mixing times of Gibbs samplers for Bayesian hierarchical models. *The Annals of Statistics*, 52(3):869 – 894, 2024.

Yves F. Atchadé. Approximate spectral gaps for markov chain mixing times in high dimensions. *SIAM Journal on Mathematics of Data Science*, 3(3):854–872, 2021.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Sergey G. Bobkov and Prasad Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19(2):289–336, 2006.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems*, 2022.

Fan Chen, Sinho Chewi, Constantinos Costis Daskalakis, and Alexander Rakhlin. High-accuracy sampling for diffusion models and log-concave distributions. In *Forty-third International Conference on Machine Learning*, 2026a.

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 *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Zongchen Chen, Kuikui Liu, and Eric Vigoda. Optimal mixing of glauber dynamics: Entropy factorization via high-dimensional expansion. *SIAM Journal on Computing*, 2026b.

Wenda Chu, Zihui Wu, Yifan Chen, Yang Song, and Yisong Yue. Split gibbs discrete diffusion posterior sampling. *arXiv preprint arXiv:2503.01161*, 2025.

Giovanni Conforti, Alain Durmus, Le-Tuyet-Nhi Pham, and Gael Raoul. Non-asymptotic convergence of discrete diffusion models: Masked and random walk dynamics. *arXiv preprint arXiv:2512.00580*, 2025.

Justin Deschenaux, Caglar Gulcehre, and Subham Sekhar Sahoo. The diffusion duality, chapter ii: ψ-samplers and efficient curriculum. *arXiv preprint arXiv:2602.21185*, 2026.

Daniil Dmitriev, Zhihan Huang, and Yuting Wei. Efficient sampling with discrete diffusion models: Sharp and adaptive guarantees. *arXiv preprint arXiv:2602.15008*, 2026.

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Khashayar Gatmiry, Sitan Chen, and Adil Salim. High-accuracy and dimension-free sampling with diffusions. In *Forty-third International Conference on Machine Learning*, 2026.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.

Joseph Gonzalez, Yucheng Low, Arthur Gretton, and Carlos Guestrin. Parallel gibbs sampling: From colored fields to thin junction trees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

Samson Gourevitch, Yazid Janati, Dario Shariatian, Umut Simsekli, Eric Moulines, Eric P. Xing, and Alain Durmus. Uniform diffusion models revisited: Leave-one-out denoiser and absorbing state reformulation. *arXiv preprint arXiv:2605.22765*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Xunpeng Huang, Difan Zou, Hanze Dong, Yi Zhang, Yian Ma, and Tong Zhang. Reverse transition kernel: A flexible framework to accelerate diffusion inference. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Kelvin Kan, Xingjian Li, Benjamin J. Zhang, Tuhin Sahai, Stanley Osher, and Markos A. Katsoulakis. Dimension-free convergence of discrete diffusion models: Adjoint equations induce the right space. *arXiv preprint arXiv:2605.17232*, 2026.

Frank P. Kelly. *Reversibility and Stochastic Networks*. Cambridge University Press, 2011.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times (Second Edition)*. American Mathematical Society, 2017.

Jose Lezama, Tim Salimans, Lu Jiang, Huiwen Chang, Jonathan Ho, and Irfan Essa. Discrete predictor-corrector diffusion models for image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Gen Li, Yuchen Zhou, Yuting Wei, and Yuxin Chen. Faster diffusion models via higher-order approximation. *arXiv preprint arXiv:2506.24042*, 2025.

Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Broadening target distributions for accelerated diffusion models via a novel analysis approach. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on score-mismatched diffusion models and zero-shot conditional samplers. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Yuchen Liang, Yingbin Liang, Lifeng Lai, and Ness Shroff. Discrete diffusion models: Novel analysis and new sampler guarantees. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c.

Yuchen Liang, Zhiheng Tan, Ness Shroff, and Yingbin Liang. Sharp convergence rates for masked diffusion models. *arXiv preprint arXiv:2602.22505*, 2026.

Chengyi Liu, Wenqi Fan, Yunqing Liu, Jiatong Li, Hang Li, Hui Liu, Jiliang Tang, and Qing Li. Generative diffusion models on graphs: methods and applications. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Le-Tuyet-Nhi Pham, Dario Shariatian, Antonio Ocello, Giovanni Conforti, and Alain Oliviero Durmus. Discrete Markov probabilistic models: An improved discrete score-based framework with sharp convergence bounds under minimal assumptions. In *Forty-second International Conference on Machine Learning*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019.

Yinuo Ren, Haoxuan Chen, Grant M. Rotskoff, and Lexing Ying. How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. In *The Thirteenth International Conference on Learning Representations*, 2025a.

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. In *Frontiers in Probabilistic Inference: Learning meets Sampling*, 2025b.

Alireza Rezaei and Shayan Oveis Gharan. A polynomial time MCMC method for sampling from continuous determinantal point processes. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Hyunjin Seo, Taewon Kim, Sihyun Yu, and SungSoo Ahn. Learning flexible forward trajectories for masked molecular diffusion. *arXiv preprint arXiv: 2505.16790*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2256–2265, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations*, 2021.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*, 2023.

M. J. Wainwright. Fast score-based sampling via log-concave reductions. *arXiv preprint arXiv:2512.24152*, 2026.

Yuhui Wu, Yuxin Chen, and Yuting Wei. Stochastic Runge-Kutta methods: Provable acceleration of diffusion models. *arXiv preprint arXiv:2410.04760*, 2024.

Xiaochen Zhang, Shuangxi Wang, Ying Fang, and Qiankun Zhang. Mg-diff: A novel molecular graph diffusion model for molecular generation and optimization. *PLOS ONE*, 2025a.

Zikun Zhang, Zixiang Chen, and Quanquan Gu. Convergence of score-based discrete diffusion models: A discrete-time analysis. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Yixiu Zhao, Jiaxin Shi, Feng Chen, Shaul Druckmann, Lester Mackey, and Scott Linderman. Informed correctors for discrete diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *International Conference on Learning Representations (ICLR)*, 2025.

Appendix

A Additional Related Works **15**

B Variants of GADD and Posterior Likelihood Estimate **15**

C Details of Numerical Experiments **16**

C.1 Synthetic Experiment 16

C.2 Text Experiment 17

C.3 Music Experiment 18

C.4 Text Samples 18

D Proof of Theorem 1 **18**

D.1 Step 1: Bounding Grand Initialization Error 19

D.2 Step 2: Bounding Inner-loop Initialization Error 19

D.3 Step 3: Error for Clean Sampling Chain 20

D.4 Step 4: Bounding Errors due to Score Estimation 20

E Analysis of CTMC Corrector **21**

F Proof of Corollary 1 **24**

G Proof of Theorem 2 **24**

G.1 On Clean-path Estimation Error 25

G.2 Explicit Lower Bound of Spectral Gap 26

H Proof of Proposition 1 **26**

I Proof of Theorem 3 **28**

I.1 Step 1: Predictor-only Analysis 28

I.2 Step 2: Incorporating Gibbs correction steps 31

J Proofs of Auxiliary Lemmas **32**

J.1 Proof of Lemma 1 32

J.2 Proof of Lemma 2 33

J.3 Proof of Lemma 3 33

J.4 Proof of Lemma 4 36

J.5 Proof of Lemma 5 37

J.6 Proof of Lemma 6 40

J.7 Proof of Lemma 7 41

J.8 Proof of Lemma 8 42

J.9 Proof of Lemma 9 43

J.10 Proof of Lemma 10 44
J.11 Proof of Lemma 11 45
J.12 Proof of Lemma 12 45

A Additional Related Works

We present the most relevant related works in Section 1.2; additional references are now discussed below.

Theory of Acceleration for Continuous Diffusion Samplers. A substantial body of theory has established accelerated convergence guarantees for continuous diffusion models, including both SDE and ODE solvers. Broadly, two main acceleration strategies have been explored. The first aims to more accurately approximate the continuous-time reverse process, for example by using higher-order numerical solvers such as Runge-Kutta methods or successive refinement (e.g., Wu et al. (2024); Liang et al. (2025a); Li et al. (2025; 2024)). The second acceleration strategy augments the diffusion process with a corrector step that steers particles toward the target distribution. This predictor-corrector paradigm has been widely adopted in empirical studies of both continuous and discrete diffusion models (Campbell et al., 2022; Song et al., 2021). From a theoretical perspective, Huang et al. (2024) proposed the RTK-MALA and RTK-ULD algorithms, which incorporate Metropolis-Adjusted Langevin Algorithm (MALA) and Underdamped Langevin Dynamics (ULD) as correctors, respectively. Notably, RTK-MALA achieves a logarithmic convergence rate in ε^{-1}. More recently, Chen et al. (2026a) introduce first-order rejection sampling to correct reverse-transition proposals using only score evaluations, obtaining polylogarithmic dependence on ε^{-1} under L^2 score-estimation guarantees. Gatmiry et al. (2026) combine a collocation-based ODE solver with a short underdamped Langevin corrector, which converts Wasserstein accuracy into total variation accuracy while preserving polylogarithmic dependence on ε^{-1} under their assumptions. A related approach by Wainwright (2026) reduces score-based sampling to a sequence of strongly log-concave subproblems, allowing efficient inner samplers to yield accelerated sampling guarantees.

Theory on Gibbs Sampling. Since the seminal work of Geman & Geman (1984), Gibbs samplers have been shown to be effective across a wide range of applications, including sampling from graphical models, feasible sets of graph coloring problems (Levin et al., 2017), determinantal point processes (Rezaei & Gharan, 2019), mixtures of densities (Atchadé, 2021), and Bayesian hierarchical models (Ascolani & Zanella, 2024). An orthogonal line of work studies the parallelization of Gibbs samplers under specific graph structures (Gonzalez et al., 2011). Notably, all of these works assume access to the exact posterior likelihood, whereas in diffusion-based settings the posterior must be estimated, introducing additional challenges that are not addressed by existing analyses.

Empirical Studies on the Acceleration of Masked Diffusion Models. Beyond acceleration for uniform-rate models, several empirical studies have focused on masked diffusion models. Lezama et al. (2023) constructed a discrete MCMC transition path for correction and learns a separate corrector kernel. More recently, Zhao et al. (2025) proposed an informed correction strategy. Notably, these approaches are primarily empirical and do not provide theoretical convergence guarantees.

We now discuss Zhao et al. (2025) in more detail. That work employs an informed Gibbs corrector specifically tailored to *masked (i.e., absorbing-rate)* diffusion models, whereas our work focuses on the *uniform-rate* setting. While both of their and our approaches utilize Gibbs-type correctors, uniform-rate diffusion models pose distinct algorithmic challenges. In particular, the score functions in uniform-rate models need to be estimated differently from those in masked diffusion, where the masked score (i.e., a time-scaled clean-data distribution) naturally facilitates Gibbs updates. In contrast, our method relies on posterior likelihood estimators specifically designed for uniform-rate models (e.g., (5)), enabling principled Gibbs corrections in this setting.

B Variants of GADD and Posterior Likelihood Estimate

In this section, we provide some variants of both the GADD algorithm in Algorithm 1 and the posterior estimator in (5). We first provide a parallel variant of Algorithm 1 in Algorithm 2.

Notably, since the score network outputs density-ratio scores for all $i \in [d]$ in a single forward pass, this modification does not incur any additional NFE per corrector step.

Algorithm 2: GADD (parallel variant)

Input: initialization $x_{t_N} \sim p_{\text{init}}$, discretization points $\{t_k\}_{k=0}^N$ (with $t_N = T$ and $t_0 = \delta$), estimated score s, correction steps $\{L_k\}_{k=0}^N$

for $k = N-1$ **to** 0 **do**
  $z_0 =$ the output of the Euler update given $x_{t_{k+1}}$ (using (2)) ;
  for $\ell = 1$ **to** L_k **do**
    Construct posterior $\hat{q}^i_{t_k}(\cdot|z^{-i}_{\ell-1})$ for each $i \in [d]$ (e.g., using (6) or (5));
    Resample $z^i_\ell \sim \hat{q}^i_{t_k}$ for each $i \in [d]$;
  end
  $x_{t_k} = z_{L_k}$;
end
Return: x_{t_0}

Next we provide variants of some posterior likelihood estimators. While (5) is one possible estimator for the posterior likelihood from the trained score, such a construction is not unique. Notice that

$$q^i_t(x^i|x^{-i}) = \left(\sum_{y^i \in [S]} \frac{q^i_t(y^i|x^{-i})}{q^i_t(x^i|x^{-i})}\right)^{-1} = \left(\sum_{y^i \in [S]} \frac{q_t(x^{-i} \oplus_i y^i)}{q_t(x)}\right)^{-1}.$$

Thus, another example is to use the following estimator:

$$\hat{q}^i_t(x^i|x^{-i}) \propto \left(\sum_{y^i \in [S]} s_t(x^{-i} \oplus_i y^i, x)\right)^{-1}. \tag{6}$$

Following the same argument as for (5), another estimator is

$$\hat{q}^i_t(x^i|x^{-i}) = \frac{\sum_{y^i \in [S]} s_t(x, x^{-i} \oplus_i y^i)}{\sum_{y^i \in [S]} \sum_{x^i \in [S]} s_t(x, x^{-i} \oplus_i y^i)}.$$

C Details of Numerical Experiments

In this section, we provide details of the synthetic experiments, ablation studies, and real-data experiments in Section 4. For the real-data experiments, we use a Google H100 GPU during training and L4 GPU during sampling.

C.1 Synthetic Experiment

For the left figure of Figure 1, we consider a synthetic autoregressive-type data model, where the likelihood of each token depends on the preceding $h = 2$ tokens. We choose the hyperparameters (e.g., L_k in GADD) such that the number of function evaluations (NFE) is the same across all methods. The final errors are averaged over 10000 Monte Carlo runs.

For the right figure of Figure 1, we use a mixed-point model with $d = S = 8$: the target is the uniform mixture over $d = 8$ randomly selected states in $[S]^d$. This gives a sparse distribution with equal-weight modes and zero probability at all other states. The final errors are averaged over 10000 Monte Carlo runs. The additional ablations below use 16 mixture points per target.

C.1.1 Scaling with dimension and alphabet size

For the additional mixed-point ablations, we fix the number of mixture points at 16 across all conditions, so that changing d does not also change the number of target modes. We use exact scores and five random seeds. In the dimension sweep, $S = 6$ and $L_k = 1$, while $d \in \{4, 6, 8, 10\}$. In the alphabet-size sweep, $d = 6$ and $L_k = 1$, while $S \in \{4, 6, 8, 10\}$. The results are reported in Table 4.

Across these settings, the mean TV error stays between 0.00917 and 0.01274 in the dimension sweep, and between 0.01064 and 0.01147 in the alphabet-size sweep. We also evaluate the autoregressive target with $S = 4$ and $d \in \{2, 3, 4, 5\}$; its coordinate-marginal TV error ranges from 0.0043 to 0.0055. In the mixed-point experiments, runtime was observed to increase approximately linearly with d and S over the tested ranges. These measurements provide empirical evidence of robustness in the tested regimes; they do not establish the asymptotic parameter dependence of the convergence bounds.

d	TV ($S = 6$)	S	TV ($d = 6$)
4	0.00917 ± 0.00153	4	0.01084 ± 0.00307
6	0.01072 ± 0.00264	6	0.01147 ± 0.00082
8	0.01274 ± 0.00066	8	0.01064 ± 0.00249
10	0.01151 ± 0.00133	10	0.01086 ± 0.00146

Table 4: Exact-score GADD on mixed-point targets: dimension and alphabet-size ablations over five random seeds. The two pairs of columns report separate sweeps, each with 16 mixture points and $L_k = 1$. Lower TV is better.

C.1.2 Fresh versus stale scores

We compare two score-evaluation strategies on a mixed-point target with $d = S = 6$ and 16 mixture points. The fresh-score implementation reevaluates the exact score at the current state before each single-coordinate correction. The stale-score implementation reuses scores within each inner loop. We vary $L_k \in \{0, 1, 2, 4, 8, 16\}$, using 20,000 samples and five random seeds. Table 5 reports the mean TV error and standard deviation across these seeds. Refreshing scores incurs additional score evaluations when $L_k > 1$.

L_k	Fresh-score TV	Stale-score TV
0	0.01274 ± 0.00117	0.01274 ± 0.00117
1	0.01185 ± 0.00052	0.01185 ± 0.00052
2	0.01045 ± 0.00185	0.01185 ± 0.00260
4	0.01231 ± 0.00172	0.01989 ± 0.00314
8	0.01040 ± 0.00225	0.02579 ± 0.00623
16	0.00946 ± 0.00104	0.03720 ± 0.01090

Table 5: Effect of score reuse on mixed-point targets with $d = S = 6$ and 16 mixture points, using 20,000 samples. Results are mean TV error $\pm$ standard deviation over five random seeds. Lower is better.

With fresh scores, the mean TV error remains between 0.00946 and 0.01274 across the tested values of L_k. With stale scores, it decreases from 0.01274 at $L_k = 0$ to 0.01185 at $L_k = 1$, then increases to 0.03720 at $L_k = 16$. The two implementations agree at $L_k = 0$ and $L_k = 1$, when within-loop score reuse has no effect. The results isolate the degradation that can arise from repeated updates with stale scores and motivate selecting L_k using a small geometric sweep on validation data.

C.2 Text Experiment

The sequence length is $d = 128$ and the dictionary size is $S = 50257$. We adopt the small SEDD Uniform model (Lou et al., 2024) with $\sigma_{\min} = 10^{-4}$ and $\sigma_{\max} = 20$.

For optimization, we use Adam with a learning rate of 3×10^{-3}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, and apply a linear warm-up over the first 2500 steps. We train the model for a total of 111K steps. For evaluation, we use the typical GPT-2 model obtained from the HuggingFace library Radford et al. (2019).

For sampling, we use the Euler method as the backbone across all methods. All reported results are averaged over 10 independent runs. For the θ-Trapezoidal method (Ren et al., 2025b), we fix $\theta = 0.5$. For the CTMC correctors (Campbell et al., 2022), we apply one correction step with a step size set to $1.5\times$ that of the Euler discretization, as suggested in their original paper.

For GADD, we employ two techniques similar to those in Zhao et al. (2025). First, we use parallel Gibbs sampling, where the score function is not re-evaluated after each token update. Specifically, we randomly select one token to update at each step, without re-evaluating the score. This significantly reduces NFEs, while incurring only a minor performance degradation when L_k is small. We choose $L_k = 40$ in our experiment. Second, we update only tokens whose estimated probabilities fall below a fixed threshold, which is set to 0.1 in our experiments.

C.3 MUSIC EXPERIMENT

The sequence length is $d = 256$ and the dictionary size is $S = 129$. We use the dataset in Campbell et al. (2022) with the pre-trained SEDD Uniform model from Chu et al. (2025).

For sampling, we condition on the first 100 tokens of each sequence. We fix the number of function evaluations (NFE) to 8 across all methods. During sampling, the conditioned tokens are strictly enforced after each predictor and corrector step. For GADD, instead of randomly sampling from one token location, we use systematic update over all tokens. Again, we only perform correction when the current probability is below the threshold, which in this case is 0.05.

For evaluation, we use the test set that consist of 200 music pieces. All the statistics (mean and standard deviation) are evaluated over five random seeds. As in Campbell et al. (2022), the mean-histogram metric compares the Hellinger distance between the histograms from the true and generated samples. As an additional metric, we report the prediction error rate over the next $F = 4$ tokens to assess short-term temporal consistency.

To compute sequence perplexity, we train a separate autoregressive (AR) model on the dataset in Campbell et al. (2022). The model uses $n_{\text{head}} = 8$, with 6 Transformer layers. The feed-forward dimension is set to 1024, with dropout rate 0.1. The model is trained using a learning rate of 3×10^{-4} for 400 epochs.

C.4 TEXT SAMPLES

We present two samples generated by GADD with $\text{NFE} = 256$ under the text-generation setup in Appendix C.2. The samples are reproduced as reported during the rebuttal, preserving the generated wording, the `<|endoftext|>` tokens, and literal `\n` markers.

Sample 1 (GADD, NFE $= 256$).

```
been killed in battle.  This would have killed those who
killed a group of soldiers on the north side.  <|endoftext|>
Some 8,000 men have destroyed a number of men of the
division commander during the war.  The division lost the
majority of the 1,400 killed and 126 wounded, left in Sir
Robert II. Lawson records that the battle obliterated above
the front line, which increased the losses of the unit,
resulted in the deaths of a survivor.  One of the soldiers
were ill- Nazaried and probably fled; three were killed,
which may have just had broken up.  John Stewart and John
Jacob Morris remain in the 3rd
```

Sample 2 (GADD, NFE $= 256$).

```
10, 1866, Congress convened the actual session of the
meeting of the United States Executive Council, establishing
the Official master statute of the United States, saying
that some past meetings were valid and lesser than the
previous ratification, and that it would address the
issue:\n<|endoftext|> While the judiciary would not take no
action on the court, it would not answer the process of the
case or actions.  .., it is difficult to bestow over the
rights of all the victims.  It is a generalized enforcement
of the promulgation of disease, infections and the vacations
of the treaty and the holding process:\n<|endoftext|> As a
result of the crisis,
```

D PROOF OF THEOREM 1

Overall, the error in Algorithm 1 arises from four distinct sources:

1. the outer-loop (or grand) initialization error at the start of the algorithm;
2. the inner-loop initialization error due to Euler predictors at the beginning of each Gibbs sampling phase; and
3. the inner-loop convergence error incurred by running the Gibbs sampler for a finite number of steps; and

4. all the errors resulting from imperfect score estimation.

Throughout the proof, we use subscript t for forward time. We write $\{p_{t_k,\ell_k}\}_{k\in[N],\ell_k\in[L_k]}$ as the path of *clean* sampling distributions, assuming no estimation error. We also write $\hat{p}_{t,\ell}$ as the *practical* path when estimation error is incorporated. Overall, we will decompose the total error as

$$\begin{aligned}\mathrm{TV}(q_\delta, \hat{p}_{\delta,L_0}) &\le \mathrm{TV}(q_\delta, p_{\delta,L_0}) + \mathrm{TV}(p_{\delta,L_0}, \hat{p}_{\delta,L_0}) \\ &\overset{(i)}{\lesssim} \sqrt{\chi^2(p_{\delta,L_0}\|q_\delta)} + \sqrt{\mathrm{KL}(p_{\delta,L_0}\|\hat{p}_{\delta,L_0})} \\ &\overset{(ii)}{\le} \sqrt{\chi^2(p_{\delta,L_0}\|q_\delta)} + \sqrt{\mathrm{KL}(p_{t_{0:N},0:L}\|\hat{p}_{t_{0:N},0:L})} \end{aligned} \tag{7}$$

where (i) follows from Cauchy-Schwarz and Pinsker's inequality, and (ii) follows from data-processing inequality where $p_{t_{0:N},0:L} := (p_{t_0,0:L_0}, \dots, \hat{p}_{t_{N-1},0:L_{N-1}}, p_{\text{init}})$ denotes the joint probability over the entire sampling path, including all predictor and corrector steps. We will use an induction-based proof for the first term, which arises from inaccurate grand initialization, inner-loop initialization due to predictors, and corrector convergence error. Then, we will separately consider the sampling error due to imperfect score estimation error in the score-entropy loss, which primarily affects the second term.

D.1 Step 1: Bounding Grand Initialization Error

To begin, the following result gives us an upper bound on the grand (i.e., outer-loop) initialization error. We particularly use χ^2 divergence, which is the typical choice to analyze Gibbs convergence.

Lemma 1. *Recall that $p_{init} = \mathrm{Uniform}([S]^d)$ for the uniform-rate CTMC. Then, for all large T's, we have*

$$\left\| \frac{p_{init}}{q_T} - 1 \right\|_{q_T}^2 = \chi^2(p_{init}\|q_T) \lesssim de^{-T}.$$

Proof. See Appendix J.1. □

D.2 Step 2: Bounding Inner-loop Initialization Error

For the first term in (7), our next result concerns the inner-loop initialization error under the Euler predictor, when the scores are perfectly estimated. Specifically, we ask the following: if the error at the end of the previous outer-loop step is sufficiently small, how large can the error be after taking one Euler predictor step and before starting the next Gibbs corrector? The influence of inaccurate scores is left to a later section.

For adjacent steps $s < t$, write $h := t - s$. Recall from (2) the (Euler) predictor update from $p_{t,L}$ to $p_{s,0}$, and let $p_{s|t}$ denote the perfect Euler transition. The following lemma characterizes the error due to the Euler predictor.

Lemma 2. *Fix adjacent steps $s < t$ where $t - s = o(1)$. Suppose that at the end of the previous corrector loop we have $\chi^2(p_{t,L}\|q_t) \le \varepsilon^2$. Then, after one perfect Euler update,*

$$\chi^2(p_{s,0}\|q_s) \lesssim B_s\varepsilon^2 + \chi^2(q_s^{\mathrm{Euler}}\|q_s) \le B_s(1+\varepsilon^2),$$

where

$$B_s := \max_{x_s} \frac{q_s^{\mathrm{Euler}}(x_s)}{q_s(x_s)}, \quad q_s^{\mathrm{Euler}}(x_s) := \sum_{x_t} q_t(x_t)p_{s|t}(x_s|x_t).$$

Here $p_{s,0} = \sum_{x_t} p_{t,L}(x_t)p_{s|t}(x_s|x_t)$ is the inner-loop initialization distribution with a perfect Euler update.

Proof. See Appendix J.2. □

Before we end this section, we note a crude bound for both B_s and $\chi^2(q_s^{\mathrm{Euler}}\|q_s)$, which turns out to be sufficient for our analysis. Indeed, with the uniform-rate CTMC,

$$q_t(x) = \sum_z q_0(z)q_{t|0}(x|z) = \frac{1}{S^d}\sum_z q_0(z)(1+(S-1)e^{-t})^{\sum_i \mathbb{1}\{x^i=z^i\}}(1-e^{-t})^{\sum_i \mathbb{1}\{x^i\neq z^i\}}.$$

Thus, we have $q_t(x) \ge \frac{1}{S^d}(1-e^{-t})^d \sum_z q_0(z) = \frac{1}{S^d}(1-e^{-t})^d$, and thus

$$\log B_s \lesssim d\log(S(1-e^{-s})^{-1}) \lesssim d\log(S\delta^{-1}). \tag{8}$$

D.3 Step 3: Error for Clean Sampling Chain

We are now ready to analyze the first term in (7). Suppose we choose $T \asymp \log(d/\varepsilon^2)$, by (1), the grand initialization error satisfies

$$\chi^2(p_{\text{init}} \| q_T) \lesssim \varepsilon^2.$$

Also, when we choose $\kappa \asymp \big(\log\log(d/\varepsilon^2)\big)^{-1}$, the total number of outer-loop steps satisfies that (cf. Liang et al. (2025c); Chen et al. (2023))

$$N \asymp \frac{T + \log \delta^{-1}}{\kappa} \lesssim \big(\log\log(d/\varepsilon^2)\big)\,(\log(d/\varepsilon^2) + \log\delta^{-1}).$$

We proceed to the inductive step. Suppose that we already have

$$\chi^2(p_{t_{k+1},L} \| q_{t_{k+1}}) \lesssim \varepsilon^2.$$

Then, at step k (corresponding to time t_k), by Lemma 2 we have

$$\chi^2(p_{t_k,0} \| q_{t_k}) \lesssim B_{t_k}(1+\varepsilon^2).$$

Then, the Gibbs corrector gives (cf. (4))

$$\chi^2(p_{t_k,L_k} \| q_{t_k}) \le (1-\rho_{t_k})^{L_k} \chi^2(p_{t_k,0} \| q_{t_k}) \lesssim (1-\rho_{t_k})^{L_k} B_{t_k}(1+\varepsilon^2).$$

Thus, in order for $\chi^2(p_{t_k,L_k} \| q_{t_k}) \le \varepsilon^2$, it suffices to choose

$$L_k \lesssim \frac{\log\big(B_{t_k}(1+\varepsilon^2)/\varepsilon^2\big)}{\rho_{t_k}} \lesssim \frac{d\log(S\delta^{-1}\varepsilon^{-2})}{\rho_{t_k}},$$

where the last inequality follows from (8). In the meantime, with such an L_k, the inductive hypothesis is satisfied. Accumulating these corrector steps, in order for $\chi^2(p_{\delta,L_0} \| q_\delta) \lesssim \varepsilon^2$ (i.e., the first term in (7)), the total number of steps satisfies (with $\delta \asymp \frac{\varepsilon}{d}$):

$$N_{total} = N + \sum_{k=N-1}^{0} L_k \lesssim \sum_{k=N-1}^{0} \frac{d\log(S\delta^{-1}\varepsilon^{-2})}{\rho_{t_k}} \lesssim \frac{d\log^2(dS/\varepsilon)}{\rho_*}.$$

D.4 Step 4: Bounding Errors due to Score Estimation

We now turn to the second term in (7). Since each predictor and corrector step is Markov, we can decompose the total error as

$$\begin{aligned}
&\mathrm{KL}(p_{t_{0:N},0:L} \| \hat{p}_{t_{0:N},0:L}) \\
&= \mathrm{KL}(p_{t_N,L_N} \| \hat{p}_{t_N,L_N}) + \sum_{k=0}^{N-1} \mathbb{E}_{x\sim p_{t_{k+1},L_{k+1}}} \mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \\
&\quad + \sum_{k=0}^{N-1}\sum_{\ell=0}^{L_k-1} \mathbb{E}_{x\sim p_{t_k,\ell}} \mathrm{KL}(p_{t_k,\ell+1|\ell}(\cdot|x) \| \hat{p}_{t_k,\ell+1|\ell}(\cdot|x)) \\
&= \sum_{k=0}^{N-1} \mathbb{E}_{x\sim p_{t_{k+1},L_{k+1}}} \mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \\
&\quad + \sum_{k=0}^{N-1}\sum_{\ell=0}^{L_k-1} \mathbb{E}_{x\sim p_{t_k,\ell}} \mathrm{KL}(p_{t_k,\ell+1|\ell}(\cdot|x) \| \hat{p}_{t_k,\ell+1|\ell}(\cdot|x)) \\
&\le \sum_{k=0}^{N-1} \mathbb{E}_{x\sim p_{t_{k+1},L_{k+1}}} \mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \\
&\quad + \sum_{k=0}^{N-1}\sum_{\ell=0}^{L_k-1} \mathbb{E}_{x\sim p^{-i}_{t_k,\ell}, i\sim w_i} \mathrm{KL}(q^i_{t_k}(\cdot \mid x^{-i}) \| \hat{q}^i_{t_k}(\cdot \mid x^{-i})),
\end{aligned} \tag{9}$$

where the second-to-last line follows because $p_{t_N,L_N} = \hat{p}_{t_N,L_N} = p_{\text{init}}$, and the last line follows by the data-processing inequality and the Gibbs corrector update rule. Here the first term is the error in the inaccurate predictors, and the second term is that due to inaccurate correctors. We first focus on the predictor term below.

Lemma 3. *Fix $k \in \{0, \dots, N-1\}$. Under Assumption 1, we have*

$$\mathbb{E}_{x \sim p_{t_{k+1}, L_{k+1}}} \mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \lesssim (t_{k+1} - t_k)\varepsilon_{est}^2.$$

Proof. See Appendix J.3. □

We next focus on the error due to inaccurate scores in the Gibbs corrector.

Lemma 4. *Fix $t \geq 0$. Under Assumption 1, with the estimated conditional update in* (5)*, we have*

$$\mathbb{E}_{x^{-i} \sim p^{-i}_{t_k, \ell}, i \sim w_i} \mathrm{KL}(q_t^i(\cdot \mid x^{-i}) \| \hat{q}_t^i(\cdot \mid x^{-i})) \lesssim \varepsilon_{est}^2.$$

Proof. See Appendix J.4. □

Therefore, plugging the results of Lemma 3 and Lemma 4 into (9), with the estimator in (5), we have

$$\begin{aligned}
\mathrm{KL}(p_{t_0,L} \| \hat{p}_{t_0,L}) &\leq \mathrm{KL}(p_{t_{0:N},0:L} \| \hat{p}_{t_{0:N},0:L}) \\
&\lesssim \sum_{k=0}^{N-1} (t_{k+1} - t_k)\varepsilon_{\text{est}}^2 + \sum_{k=0}^{N-1} \sum_{\ell=0}^{L_k - 1} \varepsilon_{\text{est}}^2 \\
&\lesssim (T + \sum_{k=0}^{N-1} L_k)\varepsilon_{\text{est}}^2.
\end{aligned}$$

Thus, as long as

$$\varepsilon_{\text{est}}^2 \lesssim \frac{\varepsilon^2}{T + \sum_{k=0}^{N-1} L_k},$$

we get $\mathrm{KL}(p_{t_0,L} \| \hat{p}_{t_0,L}) \lesssim \varepsilon^2$, or, by Pinsker's inequality, we have $\mathrm{TV}(p_{t_0,L}, \hat{p}_{t_0,L}) \lesssim \varepsilon$. Also, we have $\mathrm{TV}(q_0, q_\delta) \leq d\delta \lesssim \varepsilon$. Combining this with the preceding bounds and the triangle inequality gives the desired guarantee relative to q_0. The proof of Theorem 1 is now complete.

E Analysis of CTMC Corrector

In this section, we further apply the recursion framework developed in the proof of Theorem 1 to other existing correction schemes for discrete diffusion models, thereby demonstrating the generality of our analysis. A common class of correctors in the literature is based on CTMC dynamics. Specifically, at outer-loop time t_k, an additional CTMC with an appropriately chosen correction rate is simulated to target the distribution q_{t_k}. One example of such a correction rate is given by (Campbell et al., 2022, Section 4.4). For $x \neq y$, define the exact and estimated correction rates by

$$\begin{aligned}
R_k^c(x, y) &:= R_{t_k}(x, y) + R_{t_k}(y, x)\frac{q_{t_k}(y)}{q_{t_k}(x)}, \\
\hat{R}_k^c(x, y) &:= R_{t_k}(x, y) + R_{t_k}(y, x)s_{t_k}(y, x).
\end{aligned} \tag{10}$$

Their diagonal entries are chosen so that each row sums to zero. Indeed, (Campbell et al., 2022, Proposition 4) shows that the stationary distribution of R_k^c is correct. We again implement the learned corrector using dimension-wise Euler updates in (2), with the total off-diagonal probability capped at $c \in (0, 1)$.

As follows, we analyze the non-asymptotic performance of such a corrector using the framework of Theorem 1.

Theorem 4. *Under the same choices for outer-loop hyperparameters as in Theorem 1, and suppose that Assumption 1 holds, with $\varepsilon_{\text{est}}^2 = \mathcal{O}\left(\frac{\varepsilon^2}{T + \sum_{k=0}^{N-1} T_k}\right)$. Then, ε-TV error can be achieved with the following sufficient total number of steps (ignoring log-dependencies):*

$$N_{total} \lesssim \frac{d^2}{\mu_*^2 \varepsilon^2 (1 - e^{-\delta})^2} \asymp \frac{d^4}{\mu_*^2 \varepsilon^4},$$

where $\mu_* := \min_{0 \leq k < N} \mu_k$ and μ_k is the minimum modified log-Sobolev constant of R_k^c.

This result readily extends to the broader class of CTMC correctors with the correct stationary distribution. In particular, Theorem 4 provides a sufficient complexity bound for the specified Euler-discretized CTMC corrector. The displayed bound contains an explicit ε^{-2} factor, with any additional dependence through the μ_* retained. The proof of Theorem 4 closely parallels that of Theorem 1, relying on the same induction strategy to handle the evolving target distributions and to control the accumulation of corrector errors due to a finite number of corrector steps.

Proof of Theorem 4. Different from Gibbs sampling, the corrector CTMC is a (time-homogeneous) continuous-time process indexed by a continuous variable, denoted as τ (note that this is different from the outer-loop index t). Also, for each k, let the terminal time for the corrector CTMC be T_k, and let the corrector step-size be η_k. (Here we do not need to shrink step-sizes towards the end of the corrector process.) Note that $L_k = T_k/\eta_k$.

Similar to (7), by the triangle inequality and Pinsker's inequality, we have

$$\begin{aligned}\mathrm{TV}(\hat{p}_{\delta,L_0}, q_\delta) &\le \mathrm{TV}(p_{\delta,L_0}, q_\delta) + \mathrm{TV}(p_{\delta,L_0}, \hat{p}_{\delta,L_0}) \\ &\le \mathrm{TV}(p_{\delta,L_0}, q_\delta) + \sqrt{\frac{1}{2}\mathrm{KL}(p_{\delta,L_0} \| \hat{p}_{\delta,L_0})}. \end{aligned} \tag{11}$$

Here the first term is the error on the clean sampling path, whereas the second term is the error due to inaccurate score estimation.

On the error on the clean-sampling path We first focus on the first term. Fix the outer-loop step t_k. Write $\tilde{p}_{t_k,\tau}$ as the (ideal and clean) continuous-time sampling probability at continuous-time τ. This is different from $p_{t_k,\ell}$ that involves numerical discretization. We can decompose the first term in (11) again:

$$\mathrm{TV}(p_{t_k,L_k}, q_{t_k}) \le \mathrm{TV}(\tilde{p}_{t_k,T_k}, q_{t_k}) + \mathrm{TV}(p_{t_k,L_k}, \tilde{p}_{t_k,T_k}). \tag{12}$$

Here the first term is the typical convergence error of a CTMC with time-homogeneous rate matrix. Classical results (Bobkov & Tetali, 2006, Corollary 2.8) have established that, for any initial distribution $p_{t_k,0}$,

$$\mathrm{TV}(\tilde{p}_{t_k,T_k}, q_{t_k})^2 \le 2\log\frac{1}{\min_x q_{t_k}(x)} \cdot e^{-2\mu_k T_k} \le 2d\log\frac{S}{1-e^{-t_k}} e^{-2\mu_k T_k},$$

where μ_k is the modified log-Sobolev constant of the rate matrix R_k^c (defined in (10)). Since the reverse rate $\overleftarrow{R}_{t_k}$ ultimately depends on the underlying target distribution q_{t_k}, μ_k is also target-dependent.

We now consider the second term in (12), which corresponds to the discretization error. The result is included in the following lemma.

Lemma 5. *Suppose that η_k is sufficiently small. With the capped dimension-wise Euler update using exact scores, and noting that $T_k = L_k\eta_k$, we have*

$$\mathrm{TV}(\tilde{p}_{t_k,T_k}, p_{t_k,L_k}) \lesssim \sqrt{\frac{\eta_k T_k d^2}{(1-e^{-\delta})^2}\log\frac{eS}{1-e^{-\delta}}}.$$

Proof. See Appendix J.5. □

Therefore, continuing from (12), with perfect estimation, we have (ignoring logarithmic factors)

$$\mathrm{TV}(p_{t_k,L_k}, q_{t_k}) \lesssim \sqrt{2d\log\frac{S}{1-e^{-t_k}}} e^{-\mu_k T_k} + \sqrt{\frac{\eta_k T_k d^2}{(1-e^{-\delta})^2}}.$$

In order to have $\mathrm{TV}(p_{t_k,L_k}, q_{t_k}) \lesssim \varepsilon$, it suffices to have

$$T_k = \frac{1}{\mu_k}\log\frac{4\sqrt{2d\log(S/(1-e^{-t_k}))}}{\varepsilon}, \qquad \eta_k \asymp \frac{\varepsilon^2(1-e^{-\delta})^2}{8T_k d^2\left(1+\log\frac{S}{1-e^{-\delta}}\right)}$$

which implies that $L_k = \mathcal{O}\left(\frac{d^2}{\mu_k^2 \varepsilon^2 (1-e^{-\delta})^2}\right)$ (omitting log-factors). Thus, when the other hyper-parameters are chosen just as in Theorem 1, if we follow similar analyses, we get (ignoring log-dependencies)

$$N_{total} = N + \sum_{k=N-1}^{0} L_k \lesssim \frac{d^2}{\mu_*^2 \varepsilon^2 (1-e^{-\delta})^2}.$$

On the error due to score estimation We now consider the second error term in (11) related to inaccurate score estimation. Recall that $p_{t_k,\ell}$ and $\hat{p}_{t_k,\ell}$ denote the sampling distributions obtained using exact and estimated scores, respectively, with both under discretization. Similar to (9), this error can be expressed as

$$\begin{aligned}
&\mathrm{KL}(p_{t_{0:N},0:L} \| \hat{p}_{t_{0:N},0:L}) \\
&\le \sum_{k=0}^{N-1} \mathbb{E}_{x\sim p_{t_{k+1},L_{k+1}}} \mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \\
&\quad + \sum_{k=0}^{N-1} \sum_{\ell=0}^{L_k-1} \mathbb{E}_{x\sim p_{t_k,\ell}} \mathrm{KL}(p_{t_k,\ell+1|\ell}(\cdot \mid x) \| \hat{p}_{t_k,\ell+1|\ell}(\cdot \mid x)).
\end{aligned}$$

Here the first term is due to inaccurate predictors, which has already been analyzed back in Lemma 3. The only remaining part is to analyze inaccurate CTMC correctors, which we proceed as follows.

Lemma 6. *Fix $k \in \{0, \dots, N-1\}$ and suppose that η_k is sufficiently small. Under Assumption 1, with the capped dimension-wise Euler corrector defined above, we have*

$$\mathbb{E}_{x\sim p_{t_k,\ell}} \mathrm{KL}(p_{t_k,\ell+1|\ell}(\cdot \mid x) \| \hat{p}_{t_k,\ell+1|\ell}(\cdot \mid x)) \lesssim \eta_k \varepsilon_{est}^2.$$

Proof. See Appendix J.6. □

Thus, the total error due to estimation satisfies

$$\begin{aligned}
&\mathrm{KL}(p_{t_{0:N},0:L} \| \hat{p}_{t_{0:N},0:L}) \\
&\le \sum_{k=0}^{N-1} \mathbb{E}_{x\sim p_{t_{k+1},L_{k+1}}} \mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \\
&\quad + \sum_{k=0}^{N-1} \sum_{\ell=0}^{L_k-1} \mathbb{E}_{x\sim p_{t_k,\ell}} \mathrm{KL}(p_{t_k,\ell+1|\ell}(\cdot \mid x) \| \hat{p}_{t_k,\ell+1|\ell}(\cdot \mid x)) \\
&\overset{(i)}{\lesssim} \sum_{k=0}^{N-1} (t_{k+1} - t_k) \varepsilon_{\mathrm{est}}^2 + \sum_{k=0}^{N-1} L_k \eta_k \varepsilon_{\mathrm{est}}^2 \\
&\lesssim \left(T + \sum_{k=0}^{N-1} T_k\right) \varepsilon_{\mathrm{est}}^2,
\end{aligned}$$

where (i) follows from Lemma 3 for the predictor terms and Lemma 6 for the corrector terms. Thus, as long as

$$\varepsilon_{\mathrm{est}}^2 \lesssim \frac{\varepsilon^2}{T + \sum_{k=0}^{N-1} T_k},$$

we have

$$\mathrm{KL}(p_{\delta,L_0} \| \hat{p}_{\delta,L_0}) \le \mathrm{KL}(p_{t_{0:N},0:L} \| \hat{p}_{t_{0:N},0:L}) \lesssim \varepsilon^2,$$

where we have ignored logarithmic dependencies. The proof is now complete. □

F Proof of Corollary 1

Under the same conditions of Theorem 1, the goal is to show that ρ_* has a lower bound independent of ε when q_0 is connected. Specifically, we say that q_0 is *connected* if, for every $x, y \in \mathrm{supp}(q_0)$, there exist an integer $m \geq 0$ and states $z_0, \ldots, z_m \in \mathrm{supp}(q_0)$ such that

$$z_0 = x, \qquad z_m = y, \qquad \mathrm{Ham}\,(z_{r-1}, z_r) = 1, \quad r = 1, \ldots, m.$$

Equivalently, the subgraph of the Hamming graph induced by $\mathrm{supp}(q_0)$ is connected.

To setup the proof, let us recall what ε controls: δ, T, and κ. Our goal, then, is to show that ρ_* has a lower bound regardless of how these quantities vary, which is thus independent of ε. This is captured in the following lemma.

To this end, let us define

$$\rho_0 := 1 - \lambda_2(P_{0+}), \tag{13}$$

where $P_{0+} := \lim_{t \downarrow 0} P_t$ is the limit of the Gibbs operator as $t \to 0$. Note that ρ_0 is not the necessarily the spectral gap for q_0, particularly when q_0 does not have full support. The following lemma introduces some properties of such a ρ_0.

Lemma 7. *Recall the definition in* (13)*. The following properties hold for P_{0+} and ρ_0:*

1. *Both P_{0+} and ρ_0 exist;*
2. *When q_0 has full support, then ρ_0 is the spectral gap of q_0;*
3. *$\rho_0 > 0$ if and only if q_0 is connected, i.e., $\forall x, y \in \mathrm{supp}(q_0)$, there exists a (Hamming-distance-1) path that connects x and y.*

Proof. See Appendix J.7. □

We now formally show that ρ_* has a lower bound independent of ε, in the following lemma.

Lemma 8. *Recall the definition of ρ_0 in* (13)*, and suppose that $\rho_0 > 0$. Then, ρ_* has a lower bound independent of the target accuracy ε.*

Proof. See Appendix J.8. □

Conversely, if $\mathrm{supp}(q_0)$ is disconnected, Lemma 7 gives $\rho_0 = 0$. Since $\delta \to 0$ as $\varepsilon \to 0$, we have

$$0 \leq \rho_* \leq \rho_\delta \to \rho_0 = 0,$$

and thus no positive lower bound independent of ε exists.

This concludes the proof of Corollary 1. Also, combined with Theorem 1, we have finally established the $\mathrm{polylog}(\varepsilon^{-1})$ dependence of GADD for such targets having connected support.

G Proof of Theorem 2

From the proof of Theorem 1, all the results are still applicable until the induction (of the clean sampling chain). Now, with constant noise schedule: $t_{k+1} - t_k = \kappa$, we have $N \asymp T/\kappa$.

We again choose $T \asymp \log(d/\varepsilon^2)$, so that from Lemma 1, the grand initialization still satisfies

$$\chi^2(p_{\mathrm{init}} \| q_T) \leq \varepsilon^2.$$

Suppose that $\chi^2(p_{t_{k+1}, L_{k+1}} \| q_{t_{k+1}}) \leq \varepsilon^2$. The following lemma controls the chi-square divergence after each exact Euler predictor.

Lemma 9. *Suppose that $0 < l \leq q_0(x) \leq u$ for all x and $\kappa(u/l) \leq c < 1$. For each exact Euler predictor step with $0 \leq t_k < t_{k+1}$ and $t_{k+1} - t_k \leq \kappa$,*

$$1 + \chi^2(p_{t_k,0} \| q_{t_k}) \leq \frac{u}{l} \exp\left(\frac{(u/l-1)^2}{4(1-c)}\right) \left(1 + \chi^2(p_{t_{k+1}, L_{k+1}} \| q_{t_{k+1}})\right).$$

Proof. See Appendix J.9. □

Choose the implicit constant in κ sufficiently small so that $\kappa(u/l) \le c < 1$. By Lemma 9, when $u/l = \mathcal{O}(1)$, since c is constant, we have

$$1 + \chi^2(p_{t_k,0} \| q_{t_k}) \lesssim 1.$$

Thus, the Gibbs contraction in (4) gives

$$\chi^2(p_{t_k,L_k} \| q_{t_k}) \lesssim e^{-\rho_{t_k} L_k}.$$

Thus, when $T \asymp \log(d/\varepsilon^2)$ and $\kappa \asymp \big(\log\log(d/\varepsilon^2)\big)^{-1}$, a sufficient choice is that

$$L_k \asymp \frac{\log(1/\varepsilon^2)}{\rho_{t_k}}, \qquad N_{\text{total}} \lesssim \frac{T \log(1/\varepsilon^2)}{\kappa \rho_*}. \tag{14}$$

With a sufficiently large implicit constant in L_k, this ensures $\chi^2(p_{t_k,L_k} \| q_{t_k}) \le \varepsilon^2$, completing the induction.

In the rest of the proof, we address two issues. First, we show that a score estimation error measured along the clean diffusion path is sufficient to control the estimation error incurred by GADD. Second, we derive an explicit lower bound on the path-wise spectral gap ρ_*.

G.1 On Clean-path Estimation Error

We now consider the estimation error guarantee under Assumption 2. Again, by the inductive hypothesis and Lemma 9, and since $u/l = O(1)$ and c is constant, we have, by the Gibbs contraction in (4),

$$1 + \chi^2(p_{t_k,\ell} \| q_{t_k}) \lesssim 1, \qquad \ell = 0, \dots, L_k. \tag{15}$$

Note that the same bound holds at grand initialization. Now,

$$\begin{aligned}
&\mathrm{TV}(p_{t_0,L_0}, \hat{p}_{t_0,L_0}) \\
&\overset{(i)}{\le} \sum_{k=0}^{N-1} \mathbb{E}_{x \sim p_{t_{k+1},L_{k+1}}} \mathrm{TV}(p_{t_k|t_{k+1}}(\cdot|x), \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \\
&\quad + \sum_{k=0}^{N-1} \sum_{\ell=0}^{L_k-1} \mathbb{E}_{\substack{i \sim w_i \\ x^{-i} \sim p^{-i}_{t_k,\ell}}} \mathrm{TV}(q^i_{t_k}(\cdot|x^{-i}), \hat{q}^i_{t_k}(\cdot|x^{-i})) \\
&\overset{(ii)}{\le} \sum_{k=0}^{N-1} \left(\frac{1 + \chi^2(p_{t_{k+1},L_{k+1}} \| q_{t_{k+1}})}{2} \mathbb{E}_{x \sim q_{t_{k+1}}} \mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \right)^{1/2} \\
&\quad + \sum_{k=0}^{N-1} \sum_{\ell=0}^{L_k-1} \left(\frac{1 + \chi^2(p_{t_k,\ell} \| q_{t_k})}{2} \mathbb{E}_{\substack{i \sim w_i \\ x^{-i} \sim q^{-i}_{t_k}}} \mathrm{KL}(q^i_{t_k}(\cdot|x^{-i}) \| \hat{q}^i_{t_k}(\cdot|x^{-i})) \right)^{1/2} \\
&\overset{(iii)}{\lesssim} \left(\sum_{k=0}^{N-1} \sqrt{t_{k+1} - t_k} + \sum_{k=0}^{N-1} L_k \right) \varepsilon_{\text{est}} \\
&\overset{(iv)}{\le} \left(N + \sum_{k=0}^{N-1} L_k \right) \varepsilon_{\text{est}},
\end{aligned}$$

where (i) follows similarly as in (9) (albeit the TV version), (iii) follows from (15) and the local KL estimates from the proofs of Lemmas 3 and 4, and (iv) follows from $t_{k+1} - t_k \le \kappa \le 1$. Here (ii) follows because

$$\begin{aligned}
&\Big(\mathbb{E}_{x \sim p_{t_{k+1},L_{k+1}}} \mathrm{TV}(p_{t_k|t_{k+1}}(\cdot|x), \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \Big)^2 \\
&= \left(\mathbb{E}_{x \sim q_{t_{k+1}}} \left[\frac{p_{t_{k+1},L_{k+1}}(x)}{q_{t_{k+1}}(x)} \mathrm{TV}(p_{t_k|t_{k+1}}(\cdot|x), \hat{p}_{t_k|t_{k+1}}(\cdot|x)) \right] \right)^2 \\
&\overset{(v)}{\le} \mathbb{E}_{x \sim q_{t_{k+1}}} \left(\frac{p_{t_{k+1},L_{k+1}}(x)}{q_{t_{k+1}}(x)} \right)^2 \cdot \mathbb{E}_{x \sim q_{t_{k+1}}} \mathrm{TV}(p_{t_k|t_{k+1}}(\cdot|x), \hat{p}_{t_k|t_{k+1}}(\cdot|x))^2 \\
&= \big(1 + \chi^2(p_{t_{k+1},L_{k+1}} \| q_{t_{k+1}})\big) \cdot \mathbb{E}_{x \sim q_{t_{k+1}}} \mathrm{TV}(p_{t_k|t_{k+1}}(\cdot|x), \hat{p}_{t_k|t_{k+1}}(\cdot|x))^2
\end{aligned}$$

$$\overset{(vi)}{\leq} \frac{1+\chi^2(p_{t_{k+1},L_{k+1}}\|q_{t_{k+1}})}{2} \cdot \mathbb{E}_{x\sim q_{t_{k+1}}} \mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x)\|\hat{p}_{t_k|t_{k+1}}(\cdot|x)),$$

where (v) follows from Cauchy-Schwarz inequality, and (vi) follows from Pinsker's inequality. The corrector term follows analogously, using $\chi^2(p^{-i}_{t_k,\ell}\|q^{-i}_{t_k}) \leq \chi^2(p_{t_k,\ell}\|q_{t_k})$. This establishes (ii).

Therefore, upon choosing

$$\varepsilon^2_{\mathrm{est}} = \mathcal{O}\left(\frac{\varepsilon^2}{\left(N+\sum_{k=0}^{N-1} L_k\right)^2}\right),$$

we have $\mathrm{TV}(p_{t_0,L_0}, \hat{p}_{t_0,L_0}) \lesssim \varepsilon$. Combining this with the clean-chain error bound gives the desired TV guarantee.

G.2 Explicit Lower Bound of Spectral Gap

Full-support target also yields an explicit lower bound on the spectral gap ρ_t. The following lemma provides such a uniform lower bound for all $t \geq 0$.

Lemma 10. *Suppose that $q_0(x) \in [l,u]$ for all x with $u/l = \mathcal{O}(1)$. Also suppose that uniform random-scan Gibbs updates is used, i.e., $w_i = 1/d$. Then, the spectral gap satisfies that*

$$\rho_t \geq \frac{1}{d(u/l)} \gtrsim d^{-1}, \quad \forall t \geq 0.$$

Proof. See Appendix J.10. □

Finally, we can plug in the lower bound in Lemma 10 into (14) and get, in order to achieve ε-TV error, it suffices to have

$$N_{\mathrm{total}} \lesssim \frac{dT}{\kappa}\log(1/\varepsilon^2).$$

The proof of Theorem 2 is now complete.

H Proof of Proposition 1

We consider $d, S \geq 2$ and sufficiently small ε. Fix time $t > 0$ and coordinate $i \in [d]$. The idea is that, as long as the coordinate has been chosen at least once, since the target is a product distribution, the marginal at this coordinate $(p_t^{\mathrm{Gibbs}})^i$ will be correct (i.e., q^i_δ). Otherwise, $(p_t^{\mathrm{Gibbs}})^i$ remains a uniform distribution over $[S]$. Indeed, after t random-scan Gibbs updates, the probability that i has never been selected is

$$\mathbb{P}\left\{i \text{ has not been selected at time } t\right\} = \left(1-\frac{1}{d}\right)^t.$$

For each $i \in [d]$, define

$$Z^i_t := \mathbb{1}\left\{i \text{ has not been selected during the first } t \text{ updates}\right\}.$$

We also define the total number of unselected coordinates as $J_t := \sum_{i=1}^d Z^i_t$ and

$$m_t := \mathbb{E}[J_t] = d\left(1-\frac{1}{d}\right)^t.$$

Now consider the terminal time B. We first lower-bound the probability that at least one coordinate has never been selected. For distinct $i, j \in [d]$,

$$\begin{aligned}\mathbb{E}[Z^i_B Z^j_B] &= \mathbb{P}\left\{\text{neither } i,j \text{ has been selected at time } B\right\} \\ &= \left(1-\frac{2}{d}\right)^B \leq \left(1-\frac{1}{d}\right)^{2B} = \mathbb{E}[Z^i_B]\mathbb{E}[Z^j_B].\end{aligned}$$

Thus, the indicators $\{Z_B^i\}_{i=1}^d$ are pairwise negatively correlated. Since $(Z_B^i)^2 = Z_B^i$ (both are indicators), we have

$$\begin{aligned}\mathbb{E}[J_B^2] &= \sum_{i=1}^d \mathbb{E}[Z_B^i] + 2 \sum_{1\le i<j\le d} \mathbb{E}[Z_B^i Z_B^j] \\ &\le \sum_{i=1}^d \mathbb{E}[Z_B^i] + 2 \sum_{1\le i<j\le d} \mathbb{E}[Z_B^i]\mathbb{E}[Z_B^j] \\ &= \sum_{i=1}^d \mathbb{E}[Z_B^i] + \left(\sum_{i=1}^d \mathbb{E}[Z_B^i]\right)^2 - \sum_{i=1}^d \mathbb{E}[Z_B^i]^2 \\ &\le m_B + m_B^2.\end{aligned}$$

Thus, since $J_B > 0$, by the second-moment inequality we have

$$\mathbb{P}\{J_B > 0\} \ge \frac{\mathbb{E}[J_B]^2}{\mathbb{E}[J_B^2]} \ge \frac{m_B}{1+m_B}. \tag{16}$$

We now translate this event into a total-variation lower bound. With the singleton target, since the coordinates evolve independently along the uniform-rate forward process, we have $q_\delta(\mathbf{0}) = \left(\frac{1+(S-1)e^{-\delta}}{S}\right)^d$. Also, conditioned on J_B, q_δ is a product distribution. Moreover, given coordinate i, if i has been selected at least once, the corresponding marginal would equal to q_δ^i. Otherwise, it stays uniform over $[S]$. In other words,

$$p_B^{\mathrm{Gibbs}}(\mathbf{0}|J_B) = \left(\frac{1+(S-1)e^{-\delta}}{S}\right)^{d-J_B} \left(\frac{1}{S}\right)^{J_B}.$$

Taking average over J_B yields

$$p_B^{\mathrm{Gibbs}}(\mathbf{0}) = \left(\frac{1+(S-1)e^{-\delta}}{S}\right)^{d} \mathbb{E}_{J_B}\left(S\left(\frac{1+(S-1)e^{-\delta}}{S}\right)\right)^{-J_B}.$$

Thus,

$$\begin{aligned}\mathrm{TV}(p_B^{\mathrm{Gibbs}}, q_\delta) &\ge q_\delta(\mathbf{0}) - p_B^{\mathrm{Gibbs}}(\mathbf{0}) \\ &= \left(\frac{1+(S-1)e^{-\delta}}{S}\right)^{d} \mathbb{E}_{J_B}\left[1 - \left(1+(S-1)e^{-\delta}\right)^{-J_B}\right] \\ &\ge \left(\frac{1+(S-1)e^{-\delta}}{S}\right)^{d} \left(1 - \frac{1}{1+(S-1)e^{-\delta}}\right) \mathbb{P}\{J_B > 0\} \\ &= \left(\frac{1+(S-1)e^{-\delta}}{S}\right)^{d-1} \frac{S-1}{S} e^{-\delta} \mathbb{P}\{J_B > 0\} \\ &\ge \left(\frac{1+(S-1)e^{-\delta}}{S}\right)^{d-1} \frac{S-1}{S} e^{-\delta} \frac{m_B}{1+m_B},\end{aligned}$$

where the last line follows from (16). Now, as $\delta \asymp \varepsilon/d$, we have $\frac{1+(S-1)e^{-\delta}}{S}, e^{-\delta} \gtrsim 1$, and thus

$$\mathrm{TV}(p_B^{\mathrm{Gibbs}}, q_\delta) \gtrsim \frac{m_B}{1+m_B}.$$

Therefore, in order to have $\mathrm{TV}(p_B^{\mathrm{Gibbs}}, q_\delta) \le \varepsilon$, it is necessary that

$$m_B = d\left(1 - \frac{1}{d}\right)^B \lesssim \varepsilon,$$

and thus

$$B = \Omega\left(d \log \frac{d}{\varepsilon}\right).$$

The proof of Proposition 1 is now complete.

I PROOF OF THEOREM 3

With the singleton target, the marginal distribution at each time t is

$$q_t^i(0) = \frac{1+(S-1)e^{-t}}{S}, \quad q_t^i(a) = \frac{1-e^{-t}}{S}, \ \forall a \neq 0.$$

Thus, the score function is, $\forall (x,y) : \mathrm{Ham}\,(x,y) = 1$,

$$\frac{q_t(y)}{q_t(x)} = \begin{cases} \dfrac{1-e^{-t}}{1+(S-1)e^{-t}}, & x^i = 0, \quad y^i \neq 0, \\ \dfrac{1+(S-1)e^{-t}}{1-e^{-t}}, & x^i \neq 0, \quad y^i = 0, \\ 1, & \text{otherwise.} \end{cases}$$

Thus, the Euler predictor as in (2) can be written as, $\forall (x,y) : \mathrm{Ham}\,(x,y) = 1$,

$$\mathbb{P}\left\{x^i_{t_{k-1}} = y^i \mid x^i_{t_k} = x^i\right\} = \begin{cases} \dfrac{1}{S} \cdot \dfrac{1-e^{-t_k}}{1+(S-1)e^{-t_k}}(t_k - t_{k-1}), & x^i = 0, \quad y^i \neq 0, \\ \dfrac{1}{S} \cdot \dfrac{1+(S-1)e^{-t_k}}{1-e^{-t_k}}(t_k - t_{k-1}), & x^i \neq 0, \quad y^i = 0, \\ \dfrac{1}{S}(t_k - t_{k-1}), & \text{otherwise.} \end{cases}$$

Two observations follow. First, note that the score function is the same regardless of coordinates. Since both the initialization and the Euler update are independent across the coordinates, so is the marginal at each step. Second, the predictor does not differentiate between non-zero entries, neither does the initialization. Thus, by symmetry, we would have $q^{\mathrm{Euler}}(a) = q^{\mathrm{Euler}}(b)$ for any non-zero pair (a,b) where $a \neq b$.

I.1 STEP 1: PREDICTOR-ONLY ANALYSIS

Fix coordinate $i \in [d]$. Let us define the aggregated states:

$$\mathrm{in} := \{x^i_{t_k} = 0\}, \qquad \mathrm{out} := \{x^i_{t_k} \neq 0\},$$

with corresponding state probability π_k. At the initialization, we have $\pi_N = (1/S, (S-1)/S)^\intercal$. Then, the Euler update at each coordinate can be viewed as $\pi_k = P^{\mathrm{Euler}}_{k+1}\pi_{k+1}$, where

$$\begin{aligned} P^{\mathrm{Euler}}_k &:= \begin{pmatrix} \mathbb{P}\left\{x^i_{t_{k-1}} = 0 | x^i_{t_k} = 0\right\} & \mathbb{P}\left\{x^i_{t_{k-1}} = 0 | x^i_{t_k} \neq 0\right\} \\ \mathbb{P}\left\{x^i_{t_{k-1}} \neq 0 | x^i_{t_k} = 0\right\} & \mathbb{P}\left\{x^i_{t_{k-1}} \neq 0 | x^i_{t_k} \neq 0\right\} \end{pmatrix} \\ &= \begin{pmatrix} 1 - \frac{S-1}{S} \cdot \frac{1-e^{-t_k}}{1+(S-1)e^{-t_k}}(t_k - t_{k-1}) & \frac{1}{S} \cdot \frac{1+(S-1)e^{-t_k}}{1-e^{-t_k}}(t_k - t_{k-1}) \\ \frac{S-1}{S} \cdot \frac{1-e^{-t_k}}{1+(S-1)e^{-t_k}}(t_k - t_{k-1}) & 1 - \frac{1}{S} \cdot \frac{1+(S-1)e^{-t_k}}{1-e^{-t_k}}(t_k - t_{k-1}) \end{pmatrix} \\ &=: \begin{pmatrix} 1 - A_k\eta_k & B_k\eta_k \\ A_k\eta_k & 1 - B_k\eta_k \end{pmatrix}, \end{aligned}$$

where in the last line we have defined

$$A_k := \frac{S-1}{S} \cdot \frac{1-e^{-t_k}}{1+(S-1)e^{-t_k}}, \qquad B_k := \frac{1}{S} \cdot \frac{1+(S-1)e^{-t_k}}{1-e^{-t_k}}, \qquad \eta_k := t_k - t_{k-1}.$$

Note that we do not have the $(S-1)$ pre-factor in B_k because with any $a \neq 0$, we have $\mathbb{P}\left\{x^i_{t_{k-1}} = 0 | x^i_{t_k} \neq 0\right\} = \mathbb{P}\left\{x^i_{t_{k-1}} = 0 | x^i_{t_k} = a\right\}$.

In order to upper-bound TV, we are especially interested in the error probability $(q^{\mathrm{Euler}}_{t_k})^i(\mathrm{out})$, which is the second element in π_k. Write this as π^o_k (where we have omitted i dependence here for brevity). Then, we have $\pi^o_N = \frac{S-1}{S}$, and

$$\begin{aligned} \pi^o_{k-1} &= A_k\eta_k(1-\pi^o_k) + (1 - B_k\eta_k)\pi^o_k \\ &= A_k\eta_k + (1 - (A_k + B_k)\eta_k)\,\pi^o_k. \end{aligned} \tag{17}$$

Also note that $q^i_{t_k}(\text{out}) = \frac{S-1}{S}(1-e^{-t_k})$. Now, the TV error at t_{k-1} can be expressed as

$$\begin{aligned}
\text{TV}((q^{\text{Euler}}_{t_k})^i, q^i_{t_k}) &= \frac{1}{2}\sum_{a\in[S]} \left|(q^{\text{Euler}}_{t_k})^i(a) - q^i_{t_k}(a)\right| \\
&\overset{(i)}{=} \frac{1}{2}\left|\pi^o_k - \frac{S-1}{S}(1-e^{-t_k})\right| + \frac{1}{2}\sum_{a\neq 0}\left|\frac{\pi^o_k}{S-1} - \frac{1-e^{-t_k}}{S}\right| \\
&= \left|\pi^o_k - \frac{S-1}{S}(1-e^{-t_k})\right|, \qquad (18)
\end{aligned}$$

where (i) follows because each non-zero entry has equal marginal probability with the Euler predictor. From the recursion in (17), we can also write out another recursion for the TV error (without the absolute value). Then, the predictor step at t_k yields

$$\begin{aligned}
&\pi^o_{k-1} - \frac{S-1}{S}(1-e^{-t_{k-1}}) \\
&= A_k\eta_k + (1-(A_k+B_k)\eta_k)\left(\pi^o_k - \frac{S-1}{S}(1-e^{-t_k})\right) \\
&\quad + (1-(A_k+B_k)\eta_k)\frac{S-1}{S}(1-e^{-t_k}) - \frac{S-1}{S}(1-e^{-t_{k-1}}) \\
&= (1-(A_k+B_k)\eta_k)\left(\pi^o_k - \frac{S-1}{S}(1-e^{-t_k})\right) \\
&\quad + \left(A_k - (A_k+B_k)\frac{S-1}{S}(1-e^{-t_k})\right)\eta_k + \left(\frac{S-1}{S}(1-e^{-t_k}) - \frac{S-1}{S}(1-e^{-t_{k-1}})\right) \\
&= (1-(A_k+B_k)\eta_k)\left(\pi^o_k - \frac{S-1}{S}(1-e^{-t_k})\right) \\
&\quad + \left(A_k\left(1-\frac{S-1}{S}(1-e^{-t_k})\right) - B_k\frac{S-1}{S}(1-e^{-t_k})\right)\eta_k + \frac{S-1}{S}e^{-t_k}(e^{\eta_k}-1) \\
&= (1-(A_k+B_k)\eta_k)\left(\pi^o_k - \frac{S-1}{S}(1-e^{-t_k})\right) \\
&\quad + \left(\frac{S-1}{S}\frac{1-e^{-t_k}}{1+(S-1)e^{-t_k}}\frac{1+(S-1)e^{-t_k}}{S} - \frac{1}{S}\frac{1+(S-1)e^{-t_k}}{1-e^{-t_k}}\frac{S-1}{S}(1-e^{-t_k})\right)\eta_k \\
&\quad + \frac{S-1}{S}e^{-t_k}(e^{\eta_k}-1) \\
&= (1-(A_k+B_k)\eta_k)\left(\pi^o_k - \frac{S-1}{S}(1-e^{-t_k})\right) \\
&\quad + \frac{S-1}{S}\left(\frac{1-e^{-t_k}}{S} - \frac{1+(S-1)e^{-t_k}}{S}\right)\eta_k + \frac{S-1}{S}e^{-t_k}(e^{\eta_k}-1) \\
&= (1-(A_k+B_k)\eta_k)\left(\pi^o_k - \frac{S-1}{S}(1-e^{-t_k})\right) + \frac{S-1}{S}e^{-t_k}(e^{\eta_k}-1-\eta_k). \qquad (19)
\end{aligned}$$

We also note that

$$\begin{aligned}
A_k + B_k &= \frac{S-1}{S}\frac{1-e^{-t_k}}{1+(S-1)e^{-t_k}} + \frac{1}{S}\frac{1+(S-1)e^{-t_k}}{1-e^{-t_k}} \\
&= \frac{\frac{S-1}{S}(1-e^{-t_k})^2 + \frac{1}{S}(1+(S-1)e^{-t_k})^2}{(1+(S-1)e^{-t_k})(1-e^{-t_k})} \\
&= \frac{\frac{S-1}{S}(1-2e^{-t_k}+e^{-2t_k}) + \frac{1}{S}(1+2(S-1)e^{-t_k}+(S-1)^2e^{-2t_k})}{(1+(S-1)e^{-t_k})(1-e^{-t_k})} \\
&= \frac{1+(S-1)e^{-2t_k}}{(1+(S-1)e^{-t_k})(1-e^{-t_k})} \\
&= \frac{1}{1+(S-1)e^{-t_k}} + \frac{e^{-t_k}}{1-e^{-t_k}},
\end{aligned}$$

which implies that, for $t_k \in (0, 1]$,

$$A_k + B_k \in \left(\frac{1}{t_k} - \frac{1}{2}, 1 + \frac{1}{t_k}\right).$$

Now, under the step-size with $t_N = 1$, $t_0 = \delta$, and $\kappa = 1/4$, we have $\eta_k = t_k/4$, and thus

$$e^{-t_k}(e^{\eta_k} - 1 - \eta_k) \le \frac{\eta_k^2}{2}.$$

We also need the sign of the excess mismatch. The preceding bounds and $\eta_k = t_k/4$ imply

$$\frac{1}{2} \le 1 - (A_k + B_k)\eta_k \le \frac{3}{4} + \frac{t_k}{8} \le \frac{7}{8}.$$

The local residual $e^{-t_k}(e^{\eta_k} - 1 - \eta_k)$ is nonnegative. Since the initial excess is $(S-1)e^{-1}/S > 0$, the recursion in (19) preserves nonnegativity. It is also monotone in the incoming excess, a property we will use when incorporating Gibbs corrections. Thus, from (19), we get

$$\begin{aligned}
&\pi^o_{k-1} - \frac{S-1}{S}(1 - e^{-t_{k-1}}) \\
&= (1 - (A_k + B_k)\eta_k)\left(\pi^o_k - \frac{S-1}{S}(1 - e^{-t_k})\right) + \frac{S-1}{S}e^{-t_k}(e^{\eta_k} - 1 - \eta_k).
\end{aligned}$$

Now, for each coordinate $i \in [d]$, define the normalized excess mismatch

$$\Delta_{k,i} := \frac{S}{S-1} \cdot \frac{1}{t_k}\left(\pi^o_k - \frac{S-1}{S}(1 - e^{-t_k})\right). \tag{20}$$

Thus,

$$\begin{aligned}
\Delta_{k-1,i} &= \frac{S}{S-1}\frac{1}{t_{k-1}}(1 - (A_k + B_k)\eta_k) \cdot \frac{\Delta_{k,i}}{\frac{S}{S-1}\frac{1}{t_k}} + \frac{e^{-t_k}(e^{\eta_k} - 1 - \eta_k)}{t_{k-1}} \qquad (21)\\
&\le \frac{t_k}{t_{k-1}}\left(1 - \left(\frac{1}{t_k} - \frac{1}{2}\right)\eta_k\right) \cdot \Delta_{k,i} + \frac{1}{2t_{k-1}}\eta_k^2 \\
&= \frac{t_k}{t_{k-1}}\left(1 - \frac{\eta_k}{t_k} + \frac{\eta_k}{2}\right)\Delta_{k,i} + \frac{1}{2t_{k-1}}\eta_k^2 \\
&= \frac{t_k}{t_{k-1}}\left(\frac{t_{k-1}}{t_k} + \frac{\eta_k}{2}\right)\Delta_{k,i} + \frac{t_k - t_{k-1}}{2t_{k-1}}\eta_k \\
&= \left(1 + \frac{t_k}{t_{k-1}}\frac{\eta_k}{2}\right)\Delta_{k,i} + \frac{1}{2}\left(\frac{t_k}{t_{k-1}} - 1\right)\eta_k \\
&\overset{(ii)}{=} \left(1 + \frac{2}{3}\eta_k\right)\Delta_{k,i} + \frac{1}{6}\eta_k
\end{aligned}$$

where (ii) follows because $t_{k-1} = \frac{3}{4}t_k$ by the chosen step-sizes. Equivalently,

$$\Delta_{k-1,i} + \frac{1}{4} \le \left(1 + \frac{2}{3}\eta_k\right)\left(\Delta_{k,i} + \frac{1}{4}\right) + \frac{1}{6}\eta_k - \frac{1}{4}\left(1 + \frac{2}{3}\eta_k\right) + \frac{1}{4} \le \left(1 + \frac{2}{3}\eta_k\right)\left(\Delta_{k,i} + \frac{1}{4}\right).$$

Recursively applying this inequality yields

$$\begin{aligned}
\Delta_{0,i} + \frac{1}{4} &\le \left(\Delta_{N,i} + \frac{1}{4}\right)\prod_{k=N}^{1}\left(1 + \frac{2}{3}\eta_k\right) \\
&\le \left(\Delta_{N,i} + \frac{1}{4}\right)\exp\left(\frac{2}{3}\sum_{k=N}^{1}\eta_k\right) \\
&\le \left(\Delta_{N,i} + \frac{1}{4}\right)e^{2/3}.
\end{aligned}$$

Further, at initialization, we have

$$\Delta_{N,i} = \frac{S}{S-1} \cdot \left(\frac{S-1}{S} - \frac{S-1}{S}(1 - e^{-t_k}) \right) = e^{-1}.$$

Therefore,

$$\pi_0^o - \frac{S-1}{S}(1 - e^{-t_0}) = \frac{S-1}{S} t_0 \Delta_{0,i} \le \frac{S-1}{S} \delta C_0,$$

where $C_0 := (e^{-1} + \frac{1}{4})e^{2/3} - \frac{1}{4} \in (0,1)$ is a universal constant. Hence, from (18), the predictor-only dynamics satisfies

$$\mathrm{TV}(q_\delta^{\mathrm{Euler}}, q_\delta) \le \sum_{i=1}^{d} \mathrm{TV}((q_\delta^{\mathrm{Euler}})^i, q_\delta^i) \le d \frac{S-1}{S} \delta \le d\delta.$$

I.2 Step 2: Incorporating Gibbs correction steps

We next show that inserting exact random-scan Gibbs corrections after each predictor step does not increase the preceding predictor-only error bound. The idea is that, conditioned on the corrected indices, the excess mismatch defined in (20) cannot increase.

Let $I_{t_k} \in [d]^{L_k}$ denote the sequence of coordinates selected by the Gibbs corrector at time t_k, and let

$$\mathcal{H} := (I_{t_{N-1}}, \dots, I_\delta)$$

denote the entire history of selected coordinates. We will first condition on $\mathcal{H}$. Indeed, given $\mathcal{H}$, the sampling distribution remains a product distribution along the sampling process. This is because, given the chosen coordinates, the Gibbs corrector update at corrector step $\ell \in [L_k]$ yields a product distribution. Indeed, since the Gibbs update kernel $p_{t_k,\ell}^{I_{t_k,\ell}}(\cdot | x_{t_k,\ell}^{-I_{t_k,\ell}})$ is independent of $x_{t_k,\ell}^{-I_{t_k,\ell}}$, such an independence structure will remain after the update.

For each coordinate $i \in [d]$, let

$$\pi_{k,\ell,i}^{o,\mathcal{H}} := \mathbb{P}\left\{ x_{t_k,\ell}^i \neq 0 \mid \mathcal{H} \right\}$$

denote its "out" probability after ℓ Gibbs correction steps at time t_k, and define the corresponding normalized excess mismatch by (cf. (20))

$$\Delta_{k,\ell,i}^{\mathcal{H}} := \frac{S}{S-1} \frac{1}{t_k} \left(\pi_{k,\ell,i}^{o,\mathcal{H}} - \frac{S-1}{S}(1 - e^{-t_k}) \right).$$

Our goal below is to show that, $\forall k \in [N]$,

$$\Delta_{k,L_k,i}^{\mathcal{H}} \le \Delta_{k,i}, \ \forall i \in [d]. \tag{22}$$

We will prove (22) by induction. First, at initialization,

$$\Delta_{N,L_N,i}^{\mathcal{H}} = \Delta_{N,i} = e^{-1}, \qquad \forall i \in [d],$$

where we note that $L_N = 0$. This establish the base case. Then, for the inductive step, suppose that $\Delta_{k,L_k,i}^{\mathcal{H}} \le \Delta_{k,i}$ for all $i \in [d]$ at time t_k. Then, from (21), the predictor step yields

$$\Delta_{k-1,0,i}^{\mathcal{H}} = \frac{t_k}{t_{k-1}} \left(1 - (A_k + B_k)\eta_k\right) \Delta_{k,L_k,i}^{\mathcal{H}} + \frac{e^{-t_k}(e^{\eta_k} - 1 - \eta_k)}{t_{k-1}}. \tag{23}$$

This has the same recursive structure as (21). Its multiplier and additive term are both nonnegative, so

$$0 \le \Delta_{k,L_k,i}^{\mathcal{H}} \le \Delta_{k,i} \quad \Longrightarrow \quad 0 \le \Delta_{k-1,0,i}^{\mathcal{H}} \le \Delta_{k-1,i}.$$

Thus, the predictor preserves the nonnegative comparison needed for the induction. Also, for the Gibbs correction at time t_{k-1}, since the target $q_{t_{k-1}}$ is a product distribution, if $I_{t_k,\ell} = i$, then coordinate i can be resampled exactly from its target marginal $q_{t_{k-1}}^i$. This way, the excess mismatch is obviously zero. Otherwise, if $I_{t_k,\ell} \neq i$, $\Delta_{k-1,\ell,i}$ is kept unchanged. Thus, $\forall i \in [d]$,

$$\Delta_{k-1,L_{k-1},i}^{\mathcal{H}} \le \dots \Delta_{k-1,0,i}^{\mathcal{H}} \le \Delta_{k-1,i},$$

which completes the induction.

Finally, conditioned on $\mathcal{H}$, the same TV bounding technique can be applied, yielding

$$\mathrm{TV}(p^{\mathcal{H}}_{\delta,L_0}, q_\delta) \le \sum_{i=1}^{d} \mathrm{TV}((p^{\mathcal{H}}_{\delta,L_0})^i, q^i_\delta) \le \sum_{i=1}^{d} \left| \pi^{o,\mathcal{H}}_{0,i} - \frac{S-1}{S}(1-e^{-\delta}) \right| \le d\delta.$$

Since TV is convex, we have

$$\begin{aligned}
\mathrm{TV}(p_{\delta,L_0}, q_\delta) &= \frac{1}{2}\sum_x |p_{\delta,L_0}(x) - q_\delta(x)| \\
&= \frac{1}{2}\sum_x \left|\mathbb{E}_{\mathcal{H}}\left[p^{\mathcal{H}}_{\delta,L_0}(x)\right] - q_\delta(x)\right| \\
&\le \mathbb{E}_{\mathcal{H}}\left[\frac{1}{2}\sum_x \left|p^{\mathcal{H}}_{\delta,L_0}(x) - q_\delta(x)\right|\right] \\
&= \mathbb{E}_{\mathcal{H}}\mathrm{TV}(p^{\mathcal{H}}_{\delta,L_0}, q_\delta) \\
&\le d\delta.
\end{aligned}$$

When $\delta \asymp \varepsilon/d$, the total number of predictor steps is

$$N \lesssim \frac{T + \log \delta^{-1}}{\kappa} \lesssim \log(d/\varepsilon).$$

Since the corrector steps are constant, the same rate applies to the total number of steps. The proof is now complete.

J Proofs of Auxiliary Lemmas

We provide proofs for all helping lemmas in this section.

J.1 Proof of Lemma 1

We have

$$\begin{aligned}
&\chi^2(p_{\text{init}} \| q_T) \\
&= \mathbb{E}_{x_T \sim q_T}\left(\frac{p_{\text{init}}(x_T)}{q_T(x_T)} - 1\right)^2 \\
&= \mathbb{E}_{x_T \sim p_{\text{init}}}\frac{p_{\text{init}}(x_T)}{q_T(x_T)} - 1 \\
&= \mathbb{E}_{x_T \sim p_{\text{init}}}\frac{p_{\text{init}}(x_T)}{\mathbb{E}_{x_0 \sim q_0}[q_{T|0}(x_T|x_0)]} - 1 \\
&\overset{(i)}{\le} \mathbb{E}_{x_T \sim p_{\text{init}}}\mathbb{E}_{x_0 \sim q_0}\frac{p_{\text{init}}(x_T)}{q_{T|0}(x_T|x_0)} - 1 \\
&= \mathbb{E}_{x_T \sim p_{\text{init}}}\mathbb{E}_{x_0 \sim q_0}\prod_{i=1}^{d}\left(\frac{1}{1-e^{-T}}\right)^{\mathbb{1}\{x^i_T \neq x^i_0\}}\left(\frac{1}{1+(S-1)e^{-T}}\right)^{\mathbb{1}\{x^i_T = x^i_0\}} - 1,
\end{aligned} \tag{24}$$

where (i) follows because the function $f(u) = \frac{1}{u}$ is convex on $u > 0$. Now, for each x_0 and x_T,

$$\prod_{i=1}^{d}\left(\frac{1}{1-e^{-T}}\right)^{\mathbb{1}\{x^i_T \neq x^i_0\}}\left(\frac{1}{1+(S-1)e^{-T}}\right)^{\mathbb{1}\{x^i_T = x^i_0\}} \le \left(\frac{1}{1-e^{-T}}\right)^{d}.$$

Thus,

$$\chi^2(p_{\text{init}} \| q_T) \le \left(\frac{1}{1-e^{-T}}\right)^{d} - 1 \overset{(ii)}{\asymp} de^{-T},$$

where (ii) follows when T diverges. The proof is now complete.

J.2 Proof of Lemma 2

We first decompose the density ratio as

$$\frac{p_{s,0}}{q_s} - 1 = \frac{q_s^{\mathrm{Euler}}}{q_s}\left(\frac{p_{s,0}}{q_s^{\mathrm{Euler}}} - 1\right) + \left(\frac{q_s^{\mathrm{Euler}}}{q_s} - 1\right).$$

Taking the $L^2(q_s)$ norm and using the triangle inequality gives

$$\left\|\frac{p_{s,0}}{q_s} - 1\right\|_{q_s} \le \left\|\frac{q_s^{\mathrm{Euler}}}{q_s}\left(\frac{p_{s,0}}{q_s^{\mathrm{Euler}}} - 1\right)\right\|_{q_s} + \left\|\frac{q_s^{\mathrm{Euler}}}{q_s} - 1\right\|_{q_s}.$$

With Young's inequality, we get

$$\begin{aligned}
\chi^2(p_{s,0}\|q_s) &= \left\|\frac{p_{s,0}}{q_s} - 1\right\|_{q_s}^2 \\
&\lesssim \left\|\frac{q_s^{\mathrm{Euler}}}{q_s}\left(\frac{p_{s,0}}{q_s^{\mathrm{Euler}}} - 1\right)\right\|_{q_s}^2 + \left\|\frac{q_s^{\mathrm{Euler}}}{q_s} - 1\right\|_{q_s}^2 \\
&= \left\|\frac{q_s^{\mathrm{Euler}}}{q_s}\left(\frac{p_{s,0}}{q_s^{\mathrm{Euler}}} - 1\right)\right\|_{q_s}^2 + \chi^2(q_s^{\mathrm{Euler}}\|q_s) \\
&= \sum_x q_s(x)\left(\frac{q_s^{\mathrm{Euler}}(x)}{q_s(x)}\right)^2\left(\frac{p_{s,0}(x)}{q_s^{\mathrm{Euler}}(x)} - 1\right)^2 + \chi^2(q_s^{\mathrm{Euler}}\|q_s) \\
&= \sum_x q_s^{\mathrm{Euler}}(x)\frac{q_s^{\mathrm{Euler}}(x)}{q_s(x)}\left(\frac{p_{s,0}(x)}{q_s^{\mathrm{Euler}}(x)} - 1\right)^2 + \chi^2(q_s^{\mathrm{Euler}}\|q_s) \\
&\le B_s\chi^2(p_{s,0}\|q_s^{\mathrm{Euler}}) + \chi^2(q_s^{\mathrm{Euler}}\|q_s) \\
&\overset{(i)}{\le} B_s\chi^2(p_{t,L}\|q_t) + \chi^2(q_s^{\mathrm{Euler}}\|q_s) \\
&\le B_s\varepsilon^2 + \chi^2(q_s^{\mathrm{Euler}}\|q_s),
\end{aligned}$$

where (i) follows from the data processing inequality for χ^2-divergence. Finally, note that

$$\chi^2(q_s^{\mathrm{Euler}}\|q_s) = \sum_x q_s(x)\left(\frac{q_s^{\mathrm{Euler}}(x)}{q_s(x)} - 1\right)^2 = \sum_x q_s^{\mathrm{Euler}}(x)\left(\frac{q_s^{\mathrm{Euler}}(x)}{q_s(x)} - 1\right) \le B_s.$$

The proof is now complete.

J.3 Proof of Lemma 3

Fix the state at t_{k+1} as $x \in [S]^d$ and fix $i \in [d]$. Define

$$\phi(u, v) := u\log\frac{u}{v} - u + v. \tag{25}$$

We first recall the (non-capped) exact Euler update in (2). For each $a \neq x^i$,

$$p^i_{t_k|t_{k+1}}(a|x) = (t_{k+1} - t_k)R_{t_{k+1}}(x^{-i} \oplus_i a, x)\frac{q_{t_{k+1}}(x^{-i} \oplus_i a)}{q_{t_{k+1}}(x)}.$$

We verify that the exact update is unaffected by capping. Indeed, with the uniform-rate forward process, when the step-sizes are chosen as in Theorem 1, we have

$$\begin{aligned}
\sum_{a\neq x^i} p^i_{t_k|t_{k+1}}(a|x) &= \frac{t_{k+1} - t_k}{S}\cdot\frac{\sum_{a\neq x^i} q_{t_{k+1}}(x^{-i} \oplus_i a)}{q_{t_{k+1}}(x)} \\
&= \frac{t_{k+1} - t_k}{S}\cdot\frac{\sum_{a\neq x^i} q_{t_{k+1}}(a|x^{-i})}{q_{t_{k+1}}(x^i|x^{-i})} \\
&= \frac{t_{k+1} - t_k}{S}\left(\frac{1}{q^i_{t_{k+1}}(x^i \mid x^{-i})} - 1\right)
\end{aligned}$$

$$\leq \frac{t_{k+1}-t_k}{S} \cdot \frac{1}{q^i_{t_{k+1}}(x^i \mid x^{-i})}$$
$$\overset{(i)}{\leq} \frac{t_{k+1}-t_k}{1-e^{-t_{k+1}}}$$
$$\lesssim \kappa$$
$$\leq c,$$

when κ vanishes. Here (i) follows from the fact that

$$q^i_{t_{k+1}}(x^i \mid x^{-i}) = \sum_{x^i_0} q^i_{t_{k+1}|0}(x^i|x^i_0) q^i_{0|t_{k+1}}(x^i_0 \mid x^{-i})$$
$$\geq \frac{1-e^{-t_{k+1}}}{S} \sum_{x^i_0} q^i_{0|t_{k+1}}(x^i_0 \mid x^{-i}) = \frac{1-e^{-t_{k+1}}}{S}.$$

Also, under the target and step-sizes of Theorem 2, we already have

$$\sum_{a \neq x^i} p^i_{t_k|t_{k+1}}(a|x) \leq (t_{k+1}-t_k)\frac{u}{l} \lesssim \kappa \frac{u}{l} \leq c,$$

again when κ is vanishing. Thus, for sufficiently small κ, capping does not affect the clean Euler transitions.

Now, for the estimated transition, define the rescaling factor

$$\alpha^i_k(x) := \min\left\{1,\ \frac{c}{(t_{k+1}-t_k)\sum_{a \neq x^i} R_{t_{k+1}}(x^{-i} \oplus_i a, x) s_{t_{k+1}}(x^{-i} \oplus_i a, x)}\right\},$$

with $\alpha^i_k(x) = 1$ when the denominator is zero. Thus, the capped estimated transition of (2) is thus

$$\hat{p}^i_{t_k|t_{k+1}}(a|x) = \alpha^i_k(x)(t_{k+1}-t_k) R_{t_{k+1}}(x^{-i} \oplus_i a, x) s_{t_{k+1}}(x^{-i} \oplus_i a, x), \qquad a \neq x^i,$$

with $\hat{p}^i_{t_k|t_{k+1}}(x^i|x) = 1 - \sum_{a \neq x^i} \hat{p}^i_{t_k|t_{k+1}}(a|x)$. With this, note that $\sum_{a \neq x^i} \hat{p}^i_{t_k|t_{k+1}}(a|x) \leq c$, and (both exact and) the estimated diagonal probabilities are at least $1-c$.

To start the KL analysis, note that each coordinate is sampled independently conditional on x. Because of this, we can first decompose the KL divergence as a sum of dimension-wise KL divergence.

$$\mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}_{t_k|t_{k+1}}(\cdot|x)) = \sum_{i=1}^d \mathrm{KL}(p^i_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}^i_{t_k|t_{k+1}}(\cdot|x)). \tag{26}$$

Also, on each dimension $i \in [d]$, since both coordinate transition kernels are normalized,

$$\mathrm{KL}(p^i_{t_k|t_{k+1}}(\cdot|x) \| \hat{p}^i_{t_k|t_{k+1}}(\cdot|x))$$
$$= \sum_{a \in [S]} p^i_{t_k|t_{k+1}}(a|x) \log \frac{p^i_{t_k|t_{k+1}}(a|x)}{\hat{p}^i_{t_k|t_{k+1}}(a|x)}$$
$$= \sum_{a \neq x^i} \phi\left(p^i_{t_k|t_{k+1}}(a|x), \hat{p}^i_{t_k|t_{k+1}}(a|x)\right) + \phi\left(p^i_{t_k|t_{k+1}}(x^i|x), \hat{p}^i_{t_k|t_{k+1}}(x^i|x)\right)$$
$$\overset{(ii)}{\leq} (t_{k+1}-t_k) \sum_{a \neq x^i} R_{t_{k+1}}(x^{-i} \oplus_i a, x) \phi\left(\frac{q_{t_{k+1}}(x^{-i} \oplus_i a)}{q_{t_{k+1}}(x)}, s_{t_{k+1}}(x^{-i} \oplus_i a, x)\right)$$
$$+ \phi\left(p^i_{t_k|t_{k+1}}(x^i|x), \hat{p}^i_{t_k|t_{k+1}}(x^i|x)\right)$$
$$= (t_{k+1}-t_k) \sum_{a \neq x^i} R_{t_{k+1}}(x^{-i} \oplus_i a, x) \phi\left(\frac{q_{t_{k+1}}(x^{-i} \oplus_i a)}{q_{t_{k+1}}(x)}, s_{t_{k+1}}(x^{-i} \oplus_i a, x)\right)$$

$$+\phi\left(1-\sum_{a\neq x^i} p^i_{t_k|t_{k+1}}(a|x), 1-\sum_{a\neq x^i}\hat{p}^i_{t_k|t_{k+1}}(a|x)\right), \tag{27}$$

where (ii) is because of the following. When the cap is not active (i.e., when $\alpha^i_k(x)=1$ for the given x), the inequality is obvious (which is actually an equality). Otherwise, when the cap is active, we first note a useful relationship:

$$\begin{aligned}\phi(u,\alpha v)-\phi(u,v) &= u\log\frac{u}{\alpha v}-u+\alpha v-\left(u\log\frac{u}{v}-u+v\right)\\ &= -u\log\alpha+(\alpha-1)v.\end{aligned}$$

Taking $u=p^i_{t_k|t_{k+1}}(a|x)$, $v=(t_{k+1}-t_k)R_{t_{k+1}}(x^{-i}\oplus_i a,x)s_{t_{k+1}}(x^{-i}\oplus_i a,x)$, and $\alpha=\alpha^i_k(x)<1$ and summing over $a\neq x^i$, we have

$$\begin{aligned}&\sum_{a\neq x^i}\phi\left(p^i_{t_k|t_{k+1}}(a|x),\hat{p}^i_{t_k|t_{k+1}}(a|x)\right)\\ &\quad-(t_{k+1}-t_k)\sum_{a\neq x^i}R_{t_{k+1}}(x^{-i}\oplus_i a,x)\phi\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)},s_{t_{k+1}}(x^{-i}\oplus_i a,x)\right)\\ &=-\left(\sum_{a\neq x^i}p^i_{t_k|t_{k+1}}(a|x)\right)\log\alpha^i_k(x)\\ &\quad+(\alpha^i_k(x)-1)(t_{k+1}-t_k)\sum_{a\neq x^i}R_{t_{k+1}}(x^{-i}\oplus_i a,x)s_{t_{k+1}}(x^{-i}\oplus_i a,x)\\ &\overset{(iii)}{=}-\left(\sum_{a\neq x^i}p^i_{t_k|t_{k+1}}(a|x)\right)\log\alpha^i_k(x)+(\alpha^i_k(x)-1)\frac{c}{\alpha^i_k(x)}\\ &=-\left(\sum_{a\neq x^i}p^i_{t_k|t_{k+1}}(a|x)\right)\log\alpha^i_k(x)+c-\frac{c}{\alpha^i_k(x)}\\ &\overset{(iv)}{\leq}-c\log\alpha^i_k(x)+c-\frac{c}{\alpha^i_k(x)}\\ &\overset{(v)}{\leq}0,\end{aligned} \tag{28}$$

where (iii) follows when the cap is active, (iv) follows because $\sum_{a\neq x^i}p^i_{t_k|t_{k+1}}(a|x)\leq c$ (with $-\log\alpha^i_k(x)>0$), and (v) follows because $\log z\leq z-1$ when $z>0$. Combining the two cases, (27) is now established.

To tackle the second term above, we now use a useful property for the ϕ function, as follows.

Lemma 11. *Recall from* (25) *that* $\phi(u,v):=u\log\frac{u}{v}-u+v$. *For* $u\geq 0$ *and* $v>0$,

$$\frac{(u-v)^2}{2\max\{u,v\}}\leq\phi(u,v)\leq\frac{(u-v)^2}{v}.$$

Proof. See Appendix J.11. □

Lemma 11 further implies that, $\forall u,v\in(0,1)$,

$$\phi(1-u,1-v)\leq\frac{(u-v)^2}{1-v}\leq\frac{2\max\{u,v\}}{1-v}\phi(u,v).$$

Now, since both the sum of off-diagonal probabilities (for both clean and estimated Euler transitions) are at most $c<1$, Lemma 11 and the log-sum inequality give

$$\phi\left(1-\sum_{a\neq x^i}p^i_{t_k|t_{k+1}}(a|x),1-\sum_{a\neq x^i}\hat{p}^i_{t_k|t_{k+1}}(a|x)\right)$$

$$\leq \frac{2c}{1-c}\phi\left(\sum_{a\neq x^i} p^i_{t_k|t_{k+1}}(a|x), \sum_{a\neq x^i} \hat{p}^i_{t_k|t_{k+1}}(a|x)\right)$$
$$\overset{(v)}{\leq} \frac{2c}{1-c}\sum_{a\neq x^i}\phi\left(p^i_{t_k|t_{k+1}}(a|x), \hat{p}^i_{t_k|t_{k+1}}(a|x)\right)$$
$$\leq \frac{2c}{1-c}(t_{k+1}-t_k)\sum_{a\neq x^i} R_{t_{k+1}}(x^{-i}\oplus_i a, x)\phi\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)}, s_{t_{k+1}}(x^{-i}\oplus_i a, x)\right), \quad (29)$$

where (v) follows from log-sum inequality, and the last inequality follows again from (28) (whether or not the transition is capped). Combining (26), (27) and (29), since c is constant, we finally obtain

$$\mathbb{E}_{x\sim p_{t_{k+1},L_{k+1}}}\mathrm{KL}(p_{t_k|t_{k+1}}(\cdot|x)\|\hat{p}_{t_k|t_{k+1}}(\cdot|x))$$
$$\lesssim (t_{k+1}-t_k)\mathbb{E}_{x\sim p_{t_{k+1},L_{k+1}}}\sum_{i=1}^{d}\sum_{a\neq x^i} R_{t_{k+1}}(x^{-i}\oplus_i a, x)\phi\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)}, s_{t_{k+1}}(x^{-i}\oplus_i a, x)\right)$$
$$\leq (t_{k+1}-t_k)\varepsilon^2_{\text{est}}.$$

The proof is now complete.

J.4 Proof of Lemma 4

We first recall the estimator in (5):

$$\hat{q}^i_t(a|x^{-i}) = \frac{1}{S}\sum_{b\in[S]}\frac{s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)}{\sum_{a\in[S]} s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)},$$
$$q^i_t(a|x^{-i}) = \frac{1}{S}\sum_{b\in[S]}\frac{q_t(x^{-i}\oplus_i a)/q_t(x^{-i}\oplus_i b)}{\sum_{a\in[S]} q_t(x^{-i}\oplus_i a)/q_t(x^{-i}\oplus_i b)}.$$

We first fix $i\in[d]$ and $x^{-i}\in[S]^{d-1}$. With some abuse of notation, write

$$\hat{q}^b(a) := \frac{s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)}{\sum_{a\in[S]} s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)},$$

with $q^b(a)$ its clean version. Note that $q^b(a)$ is the same for each $b\in[S]$ since the $q_t(x^{-i}\oplus_i b)$ term is canceled out in the clean version. Thus,

$$\mathrm{KL}(q^i_t(\cdot\mid x^{-i})\|\hat{q}^i_t(\cdot\mid x^{-i})) = \mathrm{KL}\left(\frac{1}{S}\sum_{b\in[S]} q^b\middle\|\frac{1}{S}\sum_{b\in[S]}\hat{q}^b\right)$$
$$= \mathrm{KL}\left(q^1\middle\|\frac{1}{S}\sum_{b\in[S]}\hat{q}^b\right)$$
$$\overset{(i)}{\leq} \frac{1}{S}\sum_{b\in[S]}\mathrm{KL}(q^1\|\hat{q}^b)$$
$$= \frac{1}{S}\sum_{b\in[S]}\mathrm{KL}(q^b\|\hat{q}^b), \quad (30)$$

where (i) follows from Jensen's inequality since KL divergence is convex. Also, writing

$$\hat{Z}_b := \sum_{a\in[S]} s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b), \quad Z_b := \sum_{a\in[S]}\frac{q_t(x^{-i}\oplus_i a)}{q_t(x^{-i}\oplus_i b)} = \frac{1}{q_t(b|x^{-i})},$$

we have

$$\mathrm{KL}(q^b\|\hat{q}^b) = \sum_{a\in[S]} q^b(a)\log\frac{q^b(a)}{\hat{q}^b(a)}$$

$$
\begin{aligned}
&= \sum_{a\in[S]} \frac{q_t(x^{-i}\oplus_i a)/q_t(x^{-i}\oplus_i b)}{Z_b} \log \frac{\frac{q_t(x^{-i}\oplus_i a)/q_t(x^{-i}\oplus_i b)}{Z_b}}{\frac{s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)}{\hat{Z}_b}} \\
&= \frac{1}{Z_b}\sum_{a\in[S]} \frac{q_t(x^{-i}\oplus_i a)}{q_t(x^{-i}\oplus_i b)} \left(\log \frac{q_t(x^{-i}\oplus_i a)/q_t(x^{-i}\oplus_i b)}{s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)} + \log\frac{\hat{Z}_b}{Z_b}\right) \\
&= \frac{1}{Z_b}\sum_{a\in[S]} \frac{q_t(x^{-i}\oplus_i a)}{q_t(x^{-i}\oplus_i b)} \log \frac{q_t(x^{-i}\oplus_i a)/q_t(x^{-i}\oplus_i b)}{s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)} + \log\frac{\hat{Z}_b}{Z_b} \\
&\overset{(ii)}{\le} \frac{1}{Z_b}\sum_{a\in[S]} \frac{q_t(x^{-i}\oplus_i a)}{q_t(x^{-i}\oplus_i b)} \log \frac{q_t(x^{-i}\oplus_i a)/q_t(x^{-i}\oplus_i b)}{s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)} + \frac{\hat{Z}_b}{Z_b} - 1 \\
&\overset{(iii)}{=} \frac{1}{Z_b}\sum_{a\in[S]} \phi\left(\frac{q_t(x^{-i}\oplus_i a)}{q_t(x^{-i}\oplus_i b)}, s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)\right) \\
&\overset{(iv)}{=} q_t^i(b|x^{-i}) \sum_{a\in[S]} \phi\left(\frac{q_t(x^{-i}\oplus_i a)}{q_t(x^{-i}\oplus_i b)}, s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)\right)
\end{aligned}
$$

where (ii) follows from $\log x \le x-1$, for (iii) we recall from (25) that $\phi(u,v) := u\log\frac{u}{v} - u + v$, and for (iv) we note that $Z_b^{-1} = q_t^i(b|x^{-i})$. Plugging into (30), we get

$$
\begin{aligned}
&\mathbb{E}_{x^{-i}\sim p^{-i}_{t_k,\ell}, i\sim w_i} \mathrm{KL}(q_t^i(\cdot\mid x^{-i}) \| \hat{q}_t^i(\cdot\mid x^{-i})) \\
&\le \mathbb{E}_{x^{-i}\sim p^{-i}_{t_k,\ell}, i\sim w_i} \frac{1}{S}\sum_{b\in[S]} q_t^i(b|x^{-i}) \sum_{a\in[S]} \phi\left(\frac{q_t(x^{-i}\oplus_i a)}{q_t(x^{-i}\oplus_i b)}, s_t(x^{-i}\oplus_i a, x^{-i}\oplus_i b)\right) \\
&= \mathbb{E}_{x^{-i}\sim p^{-i}_{t_k,\ell}, i\sim w_i, x^i\sim q_t^i(\cdot|x^{-i})} \sum_{a\in[S]} R_t(x^{-i}\oplus_i a, x)\phi\left(\frac{q_t(x^{-i}\oplus_i a)}{q_t(x)}, s_t(x^{-i}\oplus_i a, x)\right) \\
&\le \mathbb{E}_{x^{-i}\sim p^{-i}_{t_k,\ell}, i\sim w_i, x^i\sim q_t^i(\cdot|x^{-i})} \sum_{j=1}^{d}\sum_{a\neq x^j} R_t(x^{-j}\oplus_j a, x)\phi\left(\frac{q_t(x^{-j}\oplus_j a)}{q_t(x)}, s_t(x^{-j}\oplus_j a, x)\right) \\
&= \mathbb{E}_{x\sim p_{t_k,\ell+1}} \sum_{j=1}^{d}\sum_{a\neq x^j} R_t(x^{-j}\oplus_j a, x)\phi\left(\frac{q_t(x^{-j}\oplus_j a)}{q_t(x)}, s_t(x^{-j}\oplus_j a, x)\right) \\
&\le \varepsilon_{\text{est}}^2.
\end{aligned}
$$

The proof is now complete.

J.5 Proof of Lemma 5

The proof follows similar to (Liang et al., 2025c, Theorem 2), but with several key differences. In particular, (Liang et al., 2025c, Theorem 2) is specialized to the predictor error in the outer-loop CTMC. In such a case, the analyzed reverse CTMC has a corresponding forward process that provides nice geometric properties (e.g., dimension-wise factorization used in (Liang et al., 2025c, Lemma 2)). This is different from our case with the corrector CTMC, which does not have such nice properties. On the other hand, different from the time-inhomogeneous reverse CTMC, the corrector CTMC has a time-homogeneous rate that helps the analysis.

Fix an interval $[\eta_k\ell, \eta_k(\ell+1)]$ and an initial state z. Recall the definition of R_k^c in (10), and consider the auxiliary CTMC initialized at z, whose off-diagonal rates are constant on this interval:

$$
\bar{R}_\tau^{c,z}(x, x^{-i}\oplus_i a) := \frac{1}{S}\left(1 + \frac{q_{t_k}(z^{-i}\oplus_i a)}{q_{t_k}(z)}\right), \qquad a \neq x^i.
$$

(We kept τ here not because of inhomogeneity but for purpose of identification.) All other off-diagonal rates are zero, and diagonal entries are chosen to make each row sum to zero. In particular,

$$
\bar{R}_\tau^{c,z}(z,\cdot) = R_k^c(z,\cdot), \qquad \tau\in[\eta_k\ell, \eta_k(\ell+1)],
$$

although the two generators need not agree at intermediate states $x \neq z$. On each successive interval, we would reset z to the auxiliary chain's state at the left endpoint and use the corresponding generator throughout that interval. Let $\bar{p}_{t_k,\ell}$ denote the resulting distribution at corrector time $\eta_k \ell$, with $\bar{p}_{t_k,0} = p_{t_k,0}$.

Using the conditional-probability lower bound from the proof of Lemma 3, for every $\tau \in [\eta_k \ell, \eta_k(\ell+1)]$, every reference state z, and neighboring states x, y,

$$\frac{1}{S} \leq R_k^c(x,y),\ \bar{R}_\tau^{c,z}(x,y) \leq \frac{2}{1-e^{-\delta}}.$$

Moreover,

$$\sum_{y:y\neq x} R_k^c(x,y) \leq \frac{2d}{1-e^{-\delta}}, \qquad \sum_{y:y\neq x} \bar{R}_\tau^{c,z}(x,y) \leq \frac{2d}{1-e^{-\delta}}. \tag{31}$$

Then, suppose that $\eta_k \leq \frac{c(1-e^{-\delta})}{2d}$, by (31), the exact off-diagonal probability in each coordinate is at most c. Thus, we again see that capping does not change the exact Euler kernel.

To begin, we first decompose the TV error as

$$\mathrm{TV}(\tilde{p}_{t_k,T_k}, p_{t_k,L_k}) \leq \mathrm{TV}(\tilde{p}_{t_k,T_k}, \bar{p}_{t_k,L_k}) + \mathrm{TV}(\bar{p}_{t_k,L_k}, p_{t_k,L_k}), \tag{32}$$

where we recall that $\tilde{p}_{t_k,T_k}$ is the distribution on the continuous corrector path at $\tau = T_k$, p_{t_k,L_k} is the desired distribution after L_k Euler corrector update, and $\bar{p}_{t_k,L_k}$ is that from the auxiliary CTMC with score ratios frozen at each interval's initial state. We note that $\tilde{p}_{t_k,0} = \bar{p}_{t_k,0} = p_{t_k,0}$.

First, we focus on the first term of (32). Now, applying (Liang et al., 2025c, Theorem 1) on each interval conditional on its initial state, followed by the KL chain rule over grid times and data processing, gives the following bound. We omit the dependence of g_τ on $x_{\eta_k \ell}$ when they are obvious.

$$\begin{aligned}
&\mathrm{KL}(\tilde{p}_{t_k,T_k} \| \bar{p}_{t_k,L_k}) \\
&\leq \sum_{\ell=0}^{L_k-1} \int_{\eta_k\ell}^{\eta_k(\ell+1)} \\
&\quad \mathbb{E}_{\substack{x_\tau \sim \tilde{p}_{t_k,\tau} \\ x_{\eta_k\ell} \sim \tilde{p}_{t_k,\eta_k\ell}}} \underbrace{\left[\sum_{y:\mathrm{Ham}(x_\tau,y)=1} \bar{R}_\tau^{c,x_{\eta_k\ell}}(x_\tau,y) - R_k^c(x_\tau,y) + R_k^c(x_\tau,y) \log \frac{R_k^c(x_\tau,y)}{\bar{R}_\tau^{c,x_{\eta_k\ell}}(x_\tau,y)} \right]}_{=:g_\tau(x_\tau)} \mathrm{d}\tau \\
&\overset{(i)}{=} \sum_{\ell=0}^{L_k-1} \int_{\eta_k\ell}^{\eta_k(\ell+1)} \mathbb{E}_{\substack{x_\tau \sim \tilde{p}_{t_k,\tau} \\ x_{\eta_k\ell} \sim \tilde{p}_{t_k,\eta_k\ell}}} [g_\tau(x_\tau) - g_\tau(x_{\eta_k\ell})]\, \mathrm{d}\tau,
\end{aligned}$$

where (i) follows because $\bar{R}_\tau^{c,x_{\eta_k\ell}}(x_{\eta_k\ell}, \cdot) = R_k^c(x_{\eta_k\ell}, \cdot)$, and hence $g_\tau(x_{\eta_k\ell}) = 0$. Here, note that $g_\tau(x) \geq 0$ and

$$\begin{aligned}
&\mathbb{E}_{\substack{x_\tau \sim \tilde{p}_{t_k,\tau} \\ x_{\eta_k\ell} \sim \tilde{p}_{t_k,\eta_k\ell}}} [g_\tau(x_\tau) - g_\tau(x_{\eta_k\ell})] \\
&= \mathbb{E}_{x_{\eta_k\ell} \sim \tilde{p}_{t_k,\eta_k\ell}} \left[\sum_{x_\tau} \tilde{p}_{t_k,\tau|\eta_k\ell}(x_\tau | x_{\eta_k\ell}) g_\tau(x_\tau) - g_\tau(x_{\eta_k\ell}) \right] \\
&\overset{(ii)}{\leq} \mathbb{E}_{x_{\eta_k\ell} \sim \tilde{p}_{t_k,\eta_k\ell}} \left[\sup_x g_\tau(x)\, \mathbb{P}\left\{ x_\tau \neq x_{\eta_k\ell} \mid x_{\eta_k\ell} \right\} \right] \\
&\overset{(iii)}{\leq} (\tau - \eta_k\ell) \mathbb{E}_{x_{\eta_k\ell} \sim \tilde{p}_{t_k,\eta_k\ell}} \left[\sup_x g_\tau(x) \sum_{y:y\neq x_{\eta_k\ell}} R_k^c(x_{\eta_k\ell}, y) \right] \\
&\overset{(iv)}{\leq} \frac{8d^2(\tau - \eta_k\ell)}{(1-e^{-\delta})^2} \log \frac{eS}{1-e^{-\delta}},
\end{aligned}$$

where (ii) follows because $g_\tau(x_{\eta_k\ell}) = 0$ (and, with some abuse of notation, $\mathbb{P}$ is under $\tilde{p}_{t_k,\tau|\eta_k\ell}$), (iii) follows because

$$\mathbb{P}\left\{ x_\tau \neq z \mid x_{\eta_k\ell} = z \right\}$$

$$
\begin{aligned}
&\le \mathbb{P}\left\{\text{at least one jump during } (\eta_k\ell, \tau] \mid x_{\eta_k\ell} = z\right\} \\
&= 1 - \exp\left(-(\tau - \eta_k\ell) \sum_{y: y\neq z} R_k^c(z, y)\right) \\
&\le (\tau - \eta_k\ell) \sum_{y: y\neq z} R_k^c(z, y),
\end{aligned}
$$

and (iv) is because of the following. First, the upper bound for the total out-going rate is in (31). Besides, we can obtain an upper bound uniformly on $g_\tau(x)$:

$$
\begin{aligned}
g_\tau(x) &= \sum_{y:\mathrm{Ham}(x,y)=1} \left(\bar{R}_\tau^{c,x_{\eta_k\ell}}(x, y) - R_k^c(x, y) + R_k^c(x, y) \log \frac{R_k^c(x, y)}{\bar{R}_\tau^{c,x_{\eta_k\ell}}(x, y)}\right) \\
&\le \sum_{y:\mathrm{Ham}(x,y)=1} \bar{R}_\tau^{c,x_{\eta_k\ell}}(x, y) + \sum_{y:\mathrm{Ham}(x,y)=1} R_k^c(x, y) \log \frac{2S}{1 - e^{-\delta}} \\
&\le \frac{2d}{1 - e^{-\delta}} \left(1 + \log \frac{2S}{1 - e^{-\delta}}\right) \\
&\le \frac{4d}{1 - e^{-\delta}} \log \frac{eS}{1 - e^{-\delta}}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathrm{KL}(\tilde{p}_{t_k,T_k} \| \bar{p}_{t_k,L_k}) \\
&\le \frac{8d^2}{(1 - e^{-\delta})^2} \log \frac{eS}{1 - e^{-\delta}} \sum_{\ell=0}^{L_k-1} \int_{\eta_k\ell}^{\eta_k(\ell+1)} (\tau - \eta_k\ell)\, \mathrm{d}\tau \\
&= \frac{4d^2}{(1 - e^{-\delta})^2} \log \frac{eS}{1 - e^{-\delta}} \sum_{\ell=0}^{L_k-1} \eta_k^2 \\
&= \frac{4\eta_k T_k d^2}{(1 - e^{-\delta})^2} \log \frac{eS}{1 - e^{-\delta}},
\end{aligned}
$$

and, by Pinsker's inequality, we further have

$$
\mathrm{TV}(\tilde{p}_{t_k,T_k}, \bar{p}_{t_k,L_k}) \le \sqrt{\frac{2\eta_k T_k d^2}{(1 - e^{-\delta})^2} \log \frac{eS}{1 - e^{-\delta}}}. \tag{33}
$$

We now turn to the second term in (32), which compares the update in the auxiliary process to the Euler ones. Recall from (2) that the exact Euler transition is

$$
p^i_{t_k,\ell+1|\ell}(a|z) = \eta_k R_k^c(z, z^{-i} \oplus_i a), \qquad \forall a \neq z^i.
$$

Since $\bar{R}_\tau^{c,z}$ is fixed on $\tau \in [\eta_k\ell, \eta_k(\ell + 1)]$, we can also explicitly solve the Kolmogorov forward equation for its transition probability:

$$
\begin{aligned}
\bar{p}^i_{t_k,\ell+1|\ell}(a|z) &= R_k^c(z, z^{-i} \oplus_i a) \frac{1 - \exp\left(-\eta_k \left(\frac{1}{S} \sum_{b\in[S]} \left(1 + \frac{q_{t_k}(z^{-i}\oplus_i b)}{q_{t_k}(z)}\right)\right)\right)}{\frac{1}{S} \sum_{b\in[S]} \left(1 + \frac{q_{t_k}(z^{-i}\oplus_i b)}{q_{t_k}(z)}\right)} \\
&= R_k^c(z, z^{-i} \oplus_i a) \frac{1 - \exp\left(-\eta_k \left(1 + \frac{1}{Sq^i_{t_k}(z^i|z^{-i})}\right)\right)}{1 + \frac{1}{Sq^i_{t_k}(z^i|z^{-i})}}, \qquad a \neq z^i.
\end{aligned}
$$

where we note that

$$
\frac{1}{S} \sum_{b\in[S]} \left(1 + \frac{q_{t_k}(z^{-i} \oplus_i b)}{q_{t_k}(z)}\right) = 1 + \frac{q^{-i}_{t_k}(z^{-i})}{Sq_{t_k}(z)} = 1 + \frac{1}{Sq^i_{t_k}(z^i \mid z^{-i})}.
$$

Now, using $1 - e^{-u} \le u$ when $u > 0$, we have $\bar{p}^i_{t_k,\ell+1|\ell}(a|z) \le p^i_{t_k,\ell+1|\ell}(a|z)$ for each $a \neq z^i$. On the other hand, using $u - (1 - e^{-u}) \le u^2/2$, we have

$$\begin{aligned}
&\bar{p}^i_{t_k,\ell+1|\ell}(z^i|z) - p^i_{t_k,\ell+1|\ell}(z^i|z) \\
&= \sum_{a \neq z^i} p^i_{t_k,\ell+1|\ell}(a|z) - \bar{p}^i_{t_k,\ell+1|\ell}(a|z) \\
&= \sum_{a \neq z^i} R^c_k(z, z^{-i} \oplus_i a) \cdot \frac{\eta_k \left(1 + \frac{1}{Sq^i_{t_k}(z^i|z^{-i})}\right) - \left(1 - \exp\left(-\eta_k \left(1 + \frac{1}{Sq^i_{t_k}(z^i|z^{-i})}\right)\right)\right)}{1 + \frac{1}{Sq^i_{t_k}(z^i|z^{-i})}} \\
&\le \frac{1}{2} \sum_{a \neq z^i} R^c_k(z, z^{-i} \oplus_i a) \eta_k^2 \left(1 + \frac{1}{Sq^i_{t_k}(z^i|z^{-i})}\right) \\
&\le \frac{2\eta_k^2}{(1 - e^{-\delta})^2},
\end{aligned}$$

where the last line follows because

$$\begin{aligned}
\sum_{a \neq z^i} R^c_k(z, z^{-i} \oplus_i a) &= \frac{1}{S} \sum_{a \neq z^i} \left(1 + \frac{q_{t_k}(z^{-i} \oplus_i a)}{q_{t_k}(z)}\right) \\
&\le 1 + \frac{1}{Sq^i_{t_k}(z^i|z^{-i})} \\
&\le 1 + \frac{1}{1 - e^{-\delta}} \\
&\le \frac{2}{1 - e^{-\delta}}.
\end{aligned}$$

Thus, since each dimension is independent, with data-processing inequality, the overall TV distance can be bounded as

$$\begin{aligned}
&\mathrm{TV}(\bar{p}_{t_k,L_k}, p_{t_k,L_k}) \\
&\le \sum_{\ell=0}^{L_k-1} \sup_z \mathrm{TV}(\bar{p}_{t_k,\ell+1|\ell}(\cdot|z), p_{t_k,\ell+1|\ell}(\cdot|z)) \\
&\le \sum_{\ell=0}^{L_k-1} \sum_{i=1}^{d} \sup_z \mathrm{TV}(\bar{p}^i_{t_k,\ell+1|\ell}(\cdot|z), p^i_{t_k,\ell+1|\ell}(\cdot|z)) \\
&= \sum_{\ell=0}^{L_k-1} \sum_{i=1}^{d} \sup_z \left(\bar{p}^i_{t_k,\ell+1|\ell}(z^i|z) - p^i_{t_k,\ell+1|\ell}(z^i|z)\right) \\
&\le \frac{2d\eta_k T_k}{(1 - e^{-\delta})^2}. \qquad (34)
\end{aligned}$$

Finally, combining (32), (33), and (34), we obtain that

$$\mathrm{TV}(\tilde{p}_{t_k,T_k}, p_{t_k,L_k}) \le \sqrt{\frac{2\eta_k T_k d^2}{(1 - e^{-\delta})^2} \log \frac{eS}{1 - e^{-\delta}}} + \frac{2d\eta_k T_k}{(1 - e^{-\delta})^2}.$$

Since TV is no more than 1, the right-hand-side is dominated by the square-root term. The proof is now complete.

J.6 Proof of Lemma 6

The proof will be similar to that of Lemma 3, as the CTMC corrector reserves the parallel structure of the Euler predictor.

As in the proof of Lemma 3, due to the independence structure, we still get

$$\mathrm{KL}(p_{t_k,\ell+1|\ell}(\cdot \mid x) \| \hat{p}_{t_k,\ell+1|\ell}(\cdot \mid x)) = \sum_{i=1}^{d} \mathrm{KL}(p^i_{t_k,\ell+1|\ell}(\cdot \mid x) \| \hat{p}^i_{t_k,\ell+1|\ell}(\cdot \mid x)).$$

The difference is in the transition probabilities. For $a \neq x^i$, the uncapped $p^i_{t_k,\ell+1|\ell}(a|x)$ and $\hat{p}^i_{t_k,\ell+1|\ell}(a|x)$ are, respectively,

$$\frac{\eta_k}{S}\left(1+\frac{q_{t_k}(x^{-i}\oplus_i a)}{q_{t_k}(x)}\right), \qquad \frac{\eta_k}{S}\left(1+s_{t_k}(x^{-i}\oplus_i a, x)\right).$$

As in Lemma 3, we cap the total estimated off-diagonal mass at c, with $\hat{p}^i_{t_k,\ell+1|\ell}$ being the resulting kernel. The step-size condition and (31) ensure that the exact off-diagonal mass is also at most c, so capping again does not change the exact kernel. Recall $\phi(u,v) := u\log\frac{u}{v} - u + v$ from (25). Thus, for each $i \in [d]$, we can similarly apply (27) and (29) established in the proof of Lemma 3 and obtain

$$\mathrm{KL}(p^i_{t_k,\ell+1|\ell}(\cdot|x)\|\hat{p}^i_{t_k,\ell+1|\ell}(\cdot|x)) \lesssim \frac{\eta_k}{S}\sum_{a\neq x^i}\phi\left(1+\frac{q_{t_k}(x^{-i}\oplus_i a)}{q_{t_k}(x)}, 1+s_{t_k}(x^{-i}\oplus_i a, x)\right).$$

Now, by the log-sum inequality, we have

$$\phi(1+r,1+s) \leq \phi(1,1)+\phi(r,s) = \phi(r,s).$$

Thus,

$$\mathrm{KL}(p^i_{t_k,\ell+1|\ell}(\cdot|x)\|\hat{p}^i_{t_k,\ell+1|\ell}(\cdot|x)) \lesssim \frac{\eta_k}{S}\sum_{a\neq x^i}\phi\left(\frac{q_{t_k}(x^{-i}\oplus_i a)}{q_{t_k}(x)}, s_{t_k}(x^{-i}\oplus_i a, x)\right).$$

Finally, adding over the coordinates and taking the corresponding expectation, we get

$$\begin{aligned}
&\mathbb{E}_{x\sim p_{t_k,\ell}}\mathrm{KL}(p_{t_k,\ell+1|\ell}(\cdot\mid x)\|\hat{p}_{t_k,\ell+1|\ell}(\cdot\mid x))\\
&\lesssim \eta_k\mathbb{E}_{x\sim p_{t_k,\ell}}\sum_{y:\mathrm{Ham}(x,y)=1}\frac{1}{S}\phi\left(\frac{q_{t_k}(y)}{q_{t_k}(x)}, s_{t_k}(y,x)\right)\\
&\lesssim \eta_k\varepsilon^2_{\text{est}},
\end{aligned}$$

where the last line follows from Assumption 1. The proof is now complete.

J.7 Proof of Lemma 7

Recall from (3) that the random-scan Gibbs kernel (at forward time t) is defined as

$$P_t(x,y) = \sum_{i=1}^d w_i \mathbf{1}\{x^{-i}=y^{-i}\}q^i_t(y^i|x^{-i}).$$

Thus, as long as $q^i_t(y^i|x^{-i})$ has a well-defined limit for each pair of (x,y) as $t\downarrow 0$, the entrywise limit $P_{0+}(x,y)$ is also well-defined.

To this end, we recall that with the uniform-rate transition, q_t is expressed with

$$\begin{aligned}
q_t(x) &= \sum_z q_0(z)q_{t|0}(x|z)\\
&= \frac{1}{S^d}\sum_z q_0(z)(1+(S-1)e^{-t})^{\sum_i \mathbb{1}\{x^i=z^i\}}(1-e^{-t})^{\sum_i \mathbb{1}\{x^i\neq z^i\}}\\
&\overset{(i)}{\asymp} \frac{1}{S^{r(x)}}\sum_{z\in\mathcal{R}(x)} q_0(z)t^{r(x)}\\
&=: C(x)t^{r(x)}
\end{aligned}$$

where we define

$$r(x) := \min_{z\in\mathrm{supp}(q_0)}\mathrm{Ham}(x,z), \quad \mathcal{R}(x) := \{z\in\mathrm{supp}(q_0) : \mathrm{Ham}(x,z) = r(x)\},$$

and $C(x) := S^{-r(x)}q_0(\mathcal{R}(x)) > 0$. Here (i) follows because as $t\downarrow 0$, $(1+(S-1)e^{-t}) \asymp S$ and $1-e^{-t}\asymp t$. Thus,

$$q^i_t(a|x^{-i}) = \frac{q_t(x^{-i}\oplus_i a)}{\sum_{a\in[S]} q_t(x^{-i}\oplus_i a)}$$

$$\asymp \frac{C(x^{-i} \oplus_i a) t^{r(x^{-i} \oplus_i a)}}{\sum_{a \in [S]} C(x^{-i} \oplus_i a) t^{r(x^{-i} \oplus_i a)}}$$
$$\asymp \frac{C(x^{-i} \oplus_i a) t^{r(x^{-i} \oplus_i a)}}{t^{r_*} \sum_{a \in [S]: r(x^{-i} \oplus_i a) = r_*} C(x^{-i} \oplus_i a)},$$

where $r_* := \min_{a \in [S]} r(x^{-i} \oplus_i a)$. Therefore, we have

$$\lim_{t \downarrow 0} q_t^i(a \mid x^{-i}) = \begin{cases} \frac{C(x^{-i} \oplus_i a)}{\sum_{a: r(x^{-i} \oplus_i a) = r_*} C(x^{-i} \oplus_i a)}, & \text{if } r(x^{-i} \oplus_i a) = r_*, \\ 0, & \text{if } r(x^{-i} \oplus_i a) > r_*. \end{cases}$$

Since this limit is well-defined, so is P_{0+}. Since ρ_0 is a function of the limiting operator (to be sure, not the spectral gap for q_0 necessarily), ρ_0 is also well-defined.

We now turn to the second claim. When q_0 has full support, $r(x) = 0,\ \forall x \in [S]^d$ and thus $r_* = 0$. Also, $C(x) = q_0(x)$. With this,

$$q_t^i(a|x^{-i}) \asymp \frac{q_0(x^{-i} \oplus_i a)}{\sum_{a \in [S]} q_0(x^{-i} \oplus_i a)} = q_0^i(a|x^{-i}), \quad \forall x^{-i} \in [S]^{d-1},$$

and thus $P_{0+} = P_0$ and ρ_0 is the spectral gap for q_0.

We now move on to the last claim. First for the "if" direction, note that $\forall x \in \mathrm{supp}(q_0)$, we have $r(x) = 0$. Thus, as long as $x^{-i} \oplus_i a \in \mathrm{supp}(q_0)$, we have $r_* = 0$. Thus,

$$q_t^i(a|x^{-i}) \asymp \frac{q_0(x^{-i} \oplus_i a)}{\sum_{a \in [S]} q_0(x^{-i} \oplus_i a)} = q_0^i(a|x^{-i}), \quad \forall x^{-i} : x^{-i} \oplus_i a \in \mathrm{supp}(q_0),$$

and otherwise the limit is 0. Thus, $\mathrm{supp}(q_0)$ is closed under P_{0+}. To handle states outside the support, fix $x \notin \mathrm{supp}(q_0)$ and choose $z \in \mathrm{supp}(q_0)$ such that $\mathrm{Ham}\,(x, z) = r(x)$. Choose a coordinate i with $x^i \neq z^i$ and set $y = x^{-i} \oplus_i z^i$. Since changing one coordinate changes the distance to the support by at most one,

$$r(y) = \min_{a \in [S]} r(x^{-i} \oplus_i a) = r(x) - 1.$$

The limiting conditional-probability formula above, together with $w_i > 0$, therefore implies $P_{0+}(x, y) > 0$. Repeating this construction gives a positive-probability path from x into $\mathrm{supp}(q_0)$. As a result, *all states outside the support are transient*. When the support is connected, the restriction of P_{0+} to the support is irreducible and aperiodic, since its diagonal entries are positive. Hence, P_{0+} has a unique closed irreducible aperiodic class, so $\lambda_2(P_{0+}) < 1$ and therefore $\rho_0 > 0$.

For the "only if" direction, suppose that $\mathrm{supp}(q_0)$ is disconnected, and let

$$\mathrm{supp}(q_0) = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_m, \qquad m \geq 2,$$

where $\mathcal{C}_i$ corresponds to the i-th connected components in Hamming distance. Indeed, in the limit $t \downarrow 0$, if we start from $x \in \mathrm{supp}(q_0)$, the only possibly candidate to select are those $y \in \mathrm{supp}(q_0)$ such that $\mathrm{Ham}\,(x, y) \leq 1$. Thus, only with these y's, we can have $P_{0+}(x, y) > 0$. Thus, if $x \in \mathcal{C}_i$, we must also have $y \in \mathcal{C}_i$ with P_{0+}, yielding each $\mathcal{C}_i$ a closed class under such an operator. Because the underlying graph is reducible, the eigenvalue 1 has multiplicity at least 2, i.e., $\lambda_2(P_{0+}) = 1$. As a result, $\rho_0 = 1 - \lambda_2(P_{0+}) = 0$.

The proof is now complete.

J.8 Proof of Lemma 8

The goal is to establish a lower bound on ρ_t with no dependence on δ, T, or κ. Thus, we will consider the entire continuous-time process $(q_t)_{t \geq 0}$.

We consider three regimes. First, on the high-noise regime, since $q_\infty = \mathrm{Unif}([S]^d)$, we have $\rho_t \to \rho_\infty = \min_i w_i > 0$ as $t \to \infty$. Thus, by definition of the limit, there exists $\bar{t} > 0$ such that $\forall t \geq \bar{t}$ we have $\rho_t > \rho_\infty/2$. Similarly, on the low-noise regime, since $\rho_0 > 0$, there exists $\underline{t} > 0$ such that $\forall t < \underline{t}$ we have $\rho_t > \rho_0/2$. Note that both $\bar{t}$ and $\underline{t}$ only depend on the values of ρ_∞ and ρ_0 but are independent of δ, T, and κ. Finally, in the middle regime, we have $\min_{t \in [\underline{t}, \bar{t}]} \rho_t > 0$. The positivity is guaranteed because $q_t(x) > 0$ for all $x \in [S]^d$ when $t > 0$. Also, the lower bound only depends on the values of $\underline{t}$ and $\bar{t}$, and thus it is again independent of ε. the proof is now complete.

J.9 Proof of Lemma 9

Fix k. Recall the (exact) Euler update from (2):

$$p^i_{t_k|t_{k+1}}(a|x) = \frac{1}{S}\frac{q_{t_{k+1}}(x^{-i} \oplus_i a)}{q_{t_{k+1}}(x)}(t_{k+1} - t_k), \quad \forall a \neq x^i,$$

with $p^i_{t_k|t_{k+1}}(x^i|x) = 1 - \sum_{a \neq x^i} p^i_{t_k|t_{k+1}}(a|x)$. The key idea is to construct some reference Euler update, as follows. Let K be such an update, still independent over the dimensions. On each dimension $i \in [d]$, define

$$K^i_{t_k|t_{k+1}}(a|x^i) := (1 - (t_{k+1} - t_k))\,\mathbb{1}\left\{a = x^i\right\} + \frac{t_{k+1} - t_k}{S}. \tag{35}$$

Essentially, such a reference Euler kernel sets every score ratio to 1. It is useful because it preserves the $[l, u]$ bound for full support $q_{t_{k+1}}$ by its doubly stochasticity:

$$(q_{t_{k+1}} K_{t_k|t_{k+1}})(y) \leq u \leq (u/l) q_{t_k}(y). \tag{36}$$

Now, we have

$$\begin{aligned}
&1 + \chi^2(p_{t_k,0} \| q_{t_k}) \\
&= \sum_y \frac{p_{t_k,0}(y)^2}{q_{t_k}(y)} \\
&\overset{(i)}{\leq} \frac{u}{l} \sum_y \frac{\left(\sum_x p_{t_{k+1},L_{k+1}}(x) p_{t_k|t_{k+1}}(y|x)\right)^2}{\sum_x q_{t_{k+1}}(x) K_{t_k|t_{k+1}}(y|x)} \\
&= \frac{u}{l} \sum_y \frac{1}{\sum_x q_{t_{k+1}}(x) K_{t_k|t_{k+1}}(y|x)} \left(\sum_x \frac{p_{t_{k+1},L_{k+1}}(x) p_{t_k|t_{k+1}}(y|x)}{\sqrt{q_{t_{k+1}}(x) K_{t_k|t_{k+1}}(y|x)}} \sqrt{q_{t_{k+1}}(x) K_{t_k|t_{k+1}}(y|x)} \right)^2 \\
&\overset{(ii)}{\leq} \frac{u}{l} \sum_y \sum_x \frac{p_{t_{k+1},L_{k+1}}(x)^2 p_{t_k|t_{k+1}}(y|x)^2}{q_{t_{k+1}}(x) K_{t_k|t_{k+1}}(y|x)} \\
&= \frac{u}{l} \sum_x \frac{p_{t_{k+1},L_{k+1}}(x)^2}{q_{t_{k+1}}(x)} \sum_y \frac{p_{t_k|t_{k+1}}(y|x)^2}{K_{t_k|t_{k+1}}(y|x)},
\end{aligned} \tag{37}$$

where (i) follows from (36), and (ii) follows from Cauchy-Schwarz inequality.

Note that the first term in (37) is essentially the χ^2 divergence before the predictor step. Thus, it remains to upper bound the second term, which is itself a χ^2 divergence between the Euler transition kernels, for each x. By the product structure of the Euler transition kernels,

$$\begin{aligned}
\sum_y \frac{p_{t_k|t_{k+1}}(y|x)^2}{K_{t_k|t_{k+1}}(y|x)} &= 1 + \chi^2(p_{t_k|t_{k+1}}(\cdot|x) \| K_{t_k|t_{k+1}}(\cdot|x)) \\
&= \prod_{i=1}^d \left(1 + \chi^2(p^i_{t_k|t_{k+1}}(\cdot|x) \| K^i_{t_k|t_{k+1}}(\cdot|x^i))\right) \\
&\leq \exp\left(\sum_{i=1}^d \chi^2(p^i_{t_k|t_{k+1}}(\cdot|x) \| K^i_{t_k|t_{k+1}}(\cdot|x^i)) \right),
\end{aligned} \tag{38}$$

where the last line follows because $\prod_i (1 + u_i) \leq \prod_i e^{u_i} = e^{\sum_i u_i}$. Now, for each dimension i,

$$\begin{aligned}
&\chi^2(p^i_{t_k|t_{k+1}}(\cdot|x) \| K^i_{t_k|t_{k+1}}(\cdot|x^i)) \\
&= \sum_{a \in [S]} K^i_{t_k|t_{k+1}}(a|x^i) \left(\frac{p^i_{t_k|t_{k+1}}(a|x)}{K^i_{t_k|t_{k+1}}(a|x^i)} - 1 \right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{t_{k+1}-t_k}{S}\sum_{a\neq x^i}\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)}-1\right)^2 + \frac{\left(p^i_{t_k|t_{k+1}}(x^i|x)-K^i_{t_k|t_{k+1}}(x^i|x^i)\right)^2}{K^i_{t_k|t_{k+1}}(x^i|x^i)}\\
&= \frac{t_{k+1}-t_k}{S}\sum_{a\neq x^i}\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)}-1\right)^2\\
&\quad + \frac{\frac{1}{S^2}(t_{k+1}-t_k)^2}{1-\frac{S-1}{S}(t_{k+1}-t_k)}\left(\sum_{a\neq x^i}\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)}-1\right)\right)^2\\
&\overset{(iii)}{\leq} \frac{t_{k+1}-t_k}{S\left(1-\frac{S-1}{S}(t_{k+1}-t_k)\right)}\sum_{a\neq x^i}\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)}-1\right)^2\\
&\overset{(iv)}{\leq} \frac{t_{k+1}-t_k}{S(1-c)}\sum_{a\neq x^i}\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)}-1\right)^2,
\end{aligned}
$$

where (iii) follows from Cauchy-Schwarz inequality and the fact that $\kappa < 1$, and (iv) follows because $\kappa < c$, which is the capping threshold.

Thus, it remains to provide an upper bound for the summation term. This is summarized in the following lemma.

Lemma 12. *Suppose that $0 < l \leq q_0(x) \leq u$ for all $x \in [S]^d$. Under the uniform-rate forward diffusion, for every $t_{k+1} > 0$ and $x \in [S]^d$,*

$$
\frac{1}{S}\sum_{i=1}^{d}\sum_{a\neq x^i}\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)}-1\right)^2 \leq \frac{(u/l-1)^2}{4(e^{t_{k+1}}-1)}.
$$

Proof. See Appendix J.12. □

Thus, with Lemma 12, and combining (37) and (38), we get

$$
\begin{aligned}
1+\chi^2(p_{t_k,0}\|q_{t_k}) &\leq \frac{u}{l}\sum_x \frac{p_{t_{k+1},L_{k+1}}(x)^2}{q_{t_{k+1}}(x)}\sum_y \frac{p_{t_k|t_{k+1}}(y|x)^2}{K_{t_k|t_{k+1}}(y|x)}\\
&\leq \frac{u}{l}\sum_x \frac{p_{t_{k+1},L_{k+1}}(x)^2}{q_{t_{k+1}}(x)}\exp\left(\frac{(t_{k+1}-t_k)(u/l-1)^2}{4(1-c)(e^{t_{k+1}}-1)}\right)\\
&\overset{(v)}{\leq} \frac{u}{l}\sum_x \frac{p_{t_{k+1},L_{k+1}}(x)^2}{q_{t_{k+1}}(x)}\exp\left(\frac{(u/l-1)^2}{4(1-c)}\right)\\
&= \frac{u}{l}\left(1+\chi^2(p_{t_{k+1},L_{k+1}}\|q_{t_{k+1}})\right)\exp\left(\frac{(u/l-1)^2}{4(1-c)}\right),
\end{aligned}
$$

where (v) follows because $t_{k+1}-t_k \leq t_{k+1} \leq e^{t_{k+1}}-1$. The proof is now complete.

J.10 Proof of Lemma 10

Fix $t \geq 0$. The goal is to obtain a lower bound for ρ_t via the variational characterization of the spectral gap:

$$
\rho_t = \inf_{\substack{f:[S]^d\to\mathbb{R}\\ \mathrm{Var}_{q_t}(f)>0}} \frac{\mathcal{D}_{q_t}(f)}{\mathrm{Var}_{q_t}(f)}, \tag{39}
$$

where, for any function f, we define

$$
\mathcal{D}_\pi(f) := \frac{1}{2}\sum_{x,y\in[S]^d}\pi(x)P(x,y)(f(x)-f(y))^2
$$

as the Dirichlet form of the uniform random-scan Gibbs kernel (Levin et al., 2017). For brevity, in the following we will write $\pi = q_t$ and $\nu = \mathrm{Uniform}([S]^d)$.

Note that by properties of the uniform-rate diffusion process, we have $\pi(x) \in [l, u]$. Standard analysis yields

$$\mathcal{D}_\pi(f) = \langle f, (I - P_t)f\rangle_\pi = \frac{1}{d}\sum_{i=1}^{d} \mathcal{D}_\pi^i(f) = \frac{1}{d}\sum_{i=1}^{d} \mathbb{E}_\pi\left[\mathrm{Var}_{\pi^i}(f \mid x^{-i})\right],$$

where $\mathcal{D}_\pi^i$ is the Dirichlet form corresponding to the i-th component. Here, for each i, since conditional expectation is the best squared-error approximation by a function of x^{-i}, we have

$$\begin{aligned}
\mathbb{E}_\pi[\mathrm{Var}_{\pi^i}(f \mid x^{-i})] &= \inf_g \mathbb{E}_\pi[(f - g(x^{-i}))^2] \\
&\geq \inf_x \left(\frac{\pi(x)}{\nu(x)}\right) \cdot \inf_g \mathbb{E}_\nu[(f - g(x^{-i}))^2] \\
&\geq (S^d l) \cdot \inf_g \mathbb{E}_\nu[(f - g(x^{-i}))^2] \\
&= (S^d l) \cdot \mathbb{E}_\nu[\mathrm{Var}_{\nu^i}(f \mid x^{-i})].
\end{aligned}$$

Averaging over $i \in [d]$, we thus have

$$\frac{1}{d}\sum_{i=1}^{d} \mathbb{E}_\pi\left[\mathrm{Var}_{\pi^i}(f \mid x^{-i})\right] \geq \frac{S^d l}{d}\sum_{i=1}^{d} \mathbb{E}_\nu[\mathrm{Var}_{\nu^i}(f \mid x^{-i})] \geq \frac{S^d l}{d}\mathrm{Var}_\nu(f),$$

where the last inequality follows from the tensorization of variance, and we note that $\nu = \mathrm{Uniform}([S]^d)$ is a product distribution.

It remains to lower bound $\mathrm{Var}_\nu(\cdot)$ with $\mathrm{Var}_\pi(\cdot)$. Indeed, with a similar technique as the above, we can obtain

$$\mathrm{Var}_\pi(f) = \inf_c \mathbb{E}_\pi[(f - c)^2] \leq \mathbb{E}_\pi[(f - \mathbb{E}_\nu[f])^2] \leq \sup_x \left(\frac{\pi(x)}{\nu(x)}\right) \mathrm{Var}_\nu(f) \leq (S^d u) \cdot \mathrm{Var}_\nu(f).$$

Therefore, invoking (39), we get

$$\rho_t \geq \frac{1}{d(u/l)} \gtrsim d^{-1}.$$

The proof is now complete.

J.11 Proof of Lemma 11

For the upper bound, using $\log x \leq x - 1$ for $x > 0$, we have

$$\phi(u, v) = u \log \frac{u}{v} - u + v \leq u\left(\frac{u}{v} - 1\right) - u + v = \frac{u^2 - uv - uv + v^2}{v} = \frac{(u - v)^2}{v}.$$

For the lower bound, let $f(x) := x \log x$, and we note that $f'(x) = \log x + 1$ and $f''(x) = 1/x$. Thus,

$$\phi(u, v) = u \log u - v \log v - (1 + \log v)(u - v) = f(u) - f(v) - f'(v)(u - v).$$

By Taylor's theorem with integral remainder,

$$\begin{aligned}
\phi(u, v) &= (u - v)^2 \int_0^1 \frac{1 - \theta}{v + \theta(u - v)} \mathrm{d}\theta \\
&\geq \frac{(u - v)^2}{\max\{u, v\}} \int_0^1 (1 - \theta) \mathrm{d}\theta \\
&= \frac{(u - v)^2}{2 \max\{u, v\}}.
\end{aligned}$$

J.12 Proof of Lemma 12

Fix $x \in [S]^d$. Let $Z_{t_{k+1}}$ have independent coordinates with

$$q^i_{t_{k+1}|0}(Z^i_{t_{k+1}} = a|x) = e^{-t_{k+1}} \mathbb{1}\left\{a = x^i\right\} + \frac{1 - e^{-t_{k+1}}}{S}.$$

Thus, $Z_{t_{k+1}}$ is obtained by applying the forward diffusion to the fixed state x. For brevity, throughout this proof, we write $\mathbb{P}$, $\mathbb{E}$, Cov, and Var as the probability, expectation, and variance under the measure $q_{t_{k+1}|0}$ (where we also omit dependence on x for brevity).

To derive the upper bound, we need to express the desired marginal probabilities in terms of $Z_{t_{k+1}}$. Notably, by symmetry of the forward transition kernel,

$$\begin{aligned} q_{t_{k+1}}(x) &= \sum_z q_0(z)\, q_{t_{k+1}|0}(x|z) \\ &\overset{(i)}{=} \sum_z q_0(z)\, q_{t_{k+1}|0}(z|x) \\ &= \mathbb{E}[q_0(Z_{t_{k+1}})], \end{aligned}$$

where (i) follows from symmetry of the uniform forward transition kernel. Also, if we define $m_i(a) := \mathbb{E}[q_0(Z_{t_{k+1}}) \mid Z^i_{t_{k+1}} = a]$, we have that $\mathbb{E}[m_i(Z^i_{t_{k+1}})] = \mathbb{E}[q_0(Z_{t_{k+1}})] = q_{t_{k+1}}(x)$ by tower property, and

$$\begin{aligned} &q_{t_{k+1}}(x^{-i} \oplus_i a) \\ &= \sum_{z\in[S]^d} q_0(z)\, q_{t_{k+1}|0}(x^{-i} \oplus_i a \mid z) \\ &= \sum_{b\in[S]} \sum_{z^{-i}\in[S]^{d-1}} q_0(z^{-i} \oplus_i b)\, q^i_{t_{k+1}|0}(a|b) \prod_{j\neq i} q^j_{t_{k+1}|0}(x^j|z^j) \\ &= \sum_{b\in[S]} \sum_{z^{-i}\in[S]^{d-1}} q_0(z^{-i} \oplus_i b)\, q^i_{t_{k+1}|0}(b|a) \prod_{j\neq i} q^j_{t_{k+1}|0}(x^j|z^j) \\ &= \sum_{b\in[S]} q^i_{t_{k+1}|0}(b|a) \sum_{z^{-i}\in[S]^{d-1}} q_0(z^{-i} \oplus_i b) \prod_{j\neq i} q^j_{t_{k+1}|0}(z^j|x^j) \\ &= \sum_{b\in[S]} q^i_{t_{k+1}|0}(b|a) m_i(b), \end{aligned}$$

where the last line follows because

$$\begin{aligned} m_i(a) &= \sum_{z^{-i}} q_0(z^{-i} \oplus_i a)\, \mathbb{P}(Z^{-i}_{t_{k+1}} = z^{-i} \mid Z^i_{t_{k+1}} = a) \\ &= \sum_{z^{-i}} q_0(z^{-i} \oplus_i a) \prod_{j\neq i} q^j_{t_{k+1}|0}(z^j|x^j). \end{aligned}$$

An intuition for this is that, because of the independence structure in the forward process, changing the initial state from x to $x^{-i} \oplus_i a$ changes only the distribution of $Z^i_{t_{k+1}}$. Thus, for any such $i \in [d]$ and $a \in [S]$,

$$\begin{aligned} &q_{t_{k+1}}(x^{-i} \oplus_i a) - q_{t_{k+1}}(x) \\ &= \sum_{b\in[S]} \left(q^i_{t_{k+1}|0}(b|a) - q^i_{t_{k+1}|0}(b|x^i) \right) m_i(b) \\ &= e^{-t_{k+1}} \sum_{b\in[S]} \left(\mathbb{1}\{b = a\} - \mathbb{1}\left\{b = x^i\right\} \right) m_i(b) \\ &= e^{-t_{k+1}} \left(m_i(a) - m_i(x^i) \right). \end{aligned}$$

Thus, in order to provide an upper bound, we essentially need to control $\left(m_i(a) - m_i(x^i)\right)^2$ for the summation over a. We try to control this term using variance, as follows. Note that $\mathbb{E}[m_i(Z^i_{t_{k+1}})] = \mathbb{E}[q_0(Z_{t_{k+1}})] = q_{t_{k+1}}(x)$ and

$$\mathrm{Cov}\left(m_i(Z^i_{t_{k+1}}), m_j(Z^j_{t_{k+1}}) \right) = 0, \quad \forall i \neq j,$$

since each dimension is independent under $q_{t_{k+1}|0}$. Moreover, by the definition of m_i and conditional expectation,

$$\mathrm{Cov}\left(q_0(Z_{t_{k+1}}), m_i(Z^i_{t_{k+1}}) \right)$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(q_0(Z_{t_{k+1}}) - q_{t_{k+1}}(x)\right)\left(m_i(Z^i_{t_{k+1}}) - q_{t_{k+1}}(x)\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[q_0(Z_{t_{k+1}}) - q_{t_{k+1}}(x) \mid Z^i_{t_{k+1}}\right]\left(m_i(Z^i_{t_{k+1}}) - q_{t_{k+1}}(x)\right)\right] \\
&= \mathbb{E}\left[\left(m_i(Z^i_{t_{k+1}}) - q_{t_{k+1}}(x)\right)^2\right] \\
&= \mathrm{Var}(m_i(Z^i_{t_{k+1}})).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
0 \le \mathbb{E}&\left[\left(q_0(Z_{t_{k+1}}) - q_{t_{k+1}}(x) - \sum_{i=1}^{d}\left(m_i(Z^i_{t_{k+1}}) - q_{t_{k+1}}(x)\right)\right)^2\right] \\
&= \mathrm{Var}(q_0(Z_{t_{k+1}})) + \sum_{i=1}^{d}\mathrm{Var}(m_i(Z^i_{t_{k+1}})) \\
&\quad - 2\sum_{i=1}^{d}\mathrm{Cov}\left(q_0(Z_{t_{k+1}}), m_i(Z^i_{t_{k+1}})\right) \\
&\quad + 2\sum_{1\le i<j\le d}\mathrm{Cov}\left(m_i(Z^i_{t_{k+1}}), m_j(Z^j_{t_{k+1}})\right) \\
&= \mathrm{Var}(q_0(Z_{t_{k+1}})) - \sum_{i=1}^{d}\mathrm{Var}(m_i(Z^i_{t_{k+1}})).
\end{aligned}
$$

Here, since $q_0(Z_{t_{k+1}})$ is bounded by $[l, u]$, we have

$$
\sum_{i=1}^{d}\mathrm{Var}(m_i(Z^i_{t_{k+1}})) \le \mathrm{Var}(q_0(Z_{t_{k+1}})) \le \frac{(u-l)^2}{4}.
$$

On the other hand, the pairwise representation of variance implies

$$
\begin{aligned}
\mathrm{Var}(m_i(Z^i_{t_{k+1}})) &= \frac{1}{2}\sum_{a,b\in[S]}\mathbb{P}(Z^i_{t_{k+1}} = a)\mathbb{P}(Z^i_{t_{k+1}} = b)\left(m_i(a) - m_i(b)\right)^2 \\
&\ge \frac{1}{2}\sum_{a\neq x^i}\mathbb{P}(Z^i_{t_{k+1}} = a)\mathbb{P}(Z^i_{t_{k+1}} = x^i)\left(m_i(a) - m_i(x^i)\right)^2 \\
&\quad + \frac{1}{2}\sum_{b\neq x^i}\mathbb{P}(Z^i_{t_{k+1}} = x^i)\mathbb{P}(Z^i_{t_{k+1}} = b)\left(m_i(x^i) - m_i(b)\right)^2 \\
&= \mathbb{P}(Z^i_{t_{k+1}} = x^i)\sum_{a\neq x^i}\mathbb{P}(Z^i_{t_{k+1}} = a)\left(m_i(a) - m_i(x^i)\right)^2 \\
&\ge \frac{e^{-t_{k+1}}(1 - e^{-t_{k+1}})}{S}\sum_{a\neq x^i}\left(m_i(a) - m_i(x^i)\right)^2.
\end{aligned}
$$

Combining all the preceding bounds yields

$$
\begin{aligned}
&\frac{1}{S}\sum_{i=1}^{d}\sum_{a\neq x^i}\left(\frac{q_{t_{k+1}}(x^{-i}\oplus_i a)}{q_{t_{k+1}}(x)} - 1\right)^2 \\
&= \frac{1}{S}\sum_{i=1}^{d}\sum_{a\neq x^i}\frac{(q_{t_{k+1}}(x^{-i}\oplus_i a) - q_{t_{k+1}}(x))^2}{q_{t_{k+1}}(x)^2} \\
&= \frac{1}{S}e^{-2t_{k+1}}\sum_{i=1}^{d}\sum_{a\neq x^i}\frac{(m_i(a) - m_i(x^i))^2}{q_{t_{k+1}}(x)^2}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{q_{t_{k+1}}(x)^2(e^{t_{k+1}}-1)} \sum_{i=1}^{d} \mathrm{Var}(m_i(Z^i_{t_{k+1}})) \\
&\leq \frac{1}{q_{t_{k+1}}(x)^2(e^{t_{k+1}}-1)} \cdot \frac{(u-l)^2}{4} \\
&\leq \frac{(u/l-1)^2}{4(e^{t_{k+1}}-1)}.
\end{aligned}
$$

The proof is now complete.