Efficient Sampling with Discrete Diffusion Models: Sharp and Adaptive Guarantees

Daniil Dmitriev* DANIILD@WHARTON.UPENN.EDU
Zhihan Huang* ZHIHANH@WHARTON.UPENN.EDU
Yuting Wei YTWEI@WHARTON.UPENN.EDU
University of Pennsylvania, Philadelphia, PA 19104, USA.

Editors: Steve Hanneke and Tor Lattimore

Abstract

Diffusion models over discrete spaces have recently shown striking empirical success, yet their theoretical foundations remain incomplete. In this paper, we study the sampling efficiency of score-based discrete diffusion models under a continuous-time Markov chain (CTMC) formulation, with a focus on τ-leaping-based samplers. We establish sharp convergence guarantees for attaining ε accuracy in Kullback-Leibler (KL) divergence for both uniform and masking noising processes. For uniform discrete diffusion, we show that the τ-leaping algorithm achieves an iteration complexity of order $\widetilde{O}(d/\varepsilon)$, with d the ambient dimension of the target distribution, eliminating linear dependence on the vocabulary size S and improving existing bounds by a factor of d; moreover, we establish a matching algorithmic lower bound showing that linear dependence on the ambient dimension is unavoidable in general. For masking discrete diffusion, we introduce a modified τ-leaping sampler whose convergence rate is governed by an intrinsic information-theoretic quantity, termed the *effective total correlation*, which is bounded by $d \log S$ but can be sublinear or even constant for structured data. As a consequence, the sampler provably adapts to low-dimensional structure without prior knowledge or algorithmic modification, yielding sublinear convergence rates for various practical examples (such as hidden Markov models, image data, and random graphs). Our analysis requires no boundedness or smoothness assumptions on the score estimator beyond control of the score entropy loss.

Keywords: Discrete diffusion models, masking diffusion, uniform diffusion, τ-leaping algorithm, low-dimensional adaptation

1. Introduction

Diffusion models have recently emerged as state-of-the-art approaches for high-fidelity image generation and video synthesis (Ho et al. (2020); Dhariwal and Nichol (2021); Song and Ermon (2019); Ho et al. (2022)), and have already led to significant scientific advances in various domains, including climate modeling, protein structure prediction, and materials science (Watson et al. (2023); Zeni et al. (2025); Li et al. (2024)). At their core, diffusion models are built upon two stochastic processes: a forward process that gradually corrupts the data distribution into pure noise, and a reverse process that generates samples by learning the logarithmic gradient of the perturbed marginals, commonly referred to as the score function.

Despite their broad empirical success, diffusion models have been predominantly developed for continuous data. Their extension to discrete domains, such as natural language, graph-structured

* Equal contribution. Authors are listed in alphabetical order.

© 2026 D. Dmitriev, Z. Huang & Y. Wei.

data, and categorical labels, has long remained challenging, although already discussed in Sohl-Dickstein et al. (2015). This perspective began to shift following the seminal work of Austin et al. (2021), which revealed the promise of diffusion-based approaches in discrete settings. Analogous to the continuous case, discrete diffusion models rely on a pair of noisy forward and reverse processes, with sampling achieved by learning appropriate ratios of distributions. Among recent developments (Campbell et al. (2022); Sahoo et al. (2024); Shi et al. (2024); Ou et al. (2025); Bach and Saremi (2025)), score-entropy discrete diffusion (SEDD) has demonstrated striking performance in text generation (Lou et al. (2024)), challenging the long-standing dominance of autoregressive language models. In contrast to autoregressive approaches, diffusion-based language models are not constrained to a fixed generation order (such as left-to-right), and they naturally lend themselves to more flexible forms of controlled generation, including conditional and structured text synthesis.

The promise of discrete diffusion models has spurred growing interest in their theoretical foundations. A particularly influential line of work formulates discrete diffusion through the lens of continuous-time Markov chains (CTMCs) (Campbell et al., 2022), in which the forward dynamics is governed by carefully designed rate matrix, and backward dynamics is approximated via a learned score function. Among the proposed constructions, two choices have emerged as especially prominent: the uniform rate matrix, which induces a uniform stationary distribution for the forward process, and the absorbing rate matrix, which yields a degenerate stationary distribution with an absorbing state. In practice, the performance of the resulting samplers depends sensitively on the choice of the rate matrix (Lou et al. (2024); von Rütte et al. (2025)). Correspondingly, two parallel lines of work have sought to understand the sampling efficiency of discrete diffusion models — specifically, the number of steps required to produce sufficiently accurate samples — under these respective constructions. Representative results include Chen and Ying (2025); Ren et al. (2025); Zhang et al. (2025); Pham et al. (2025); Liang et al. (2025b) for uniform diffusion and Park et al. (2025); Liang et al. (2025a); Conforti et al. (2025) for masking diffusion (also referred to as absorbing diffusion).

Existing theoretical analyses for score-based discrete diffusions suggest that convergence rates typically scale at least linearly with both the size of the vocabulary size S and the ambient dimension d. Such scaling can quickly become prohibitive in applications; for instance, in GPT-2-based tasks, the vocabulary size is $S = 50{,}257$ and the dimension is $d = 10^2 \sim 10^3$ (Lou et al., 2024). These considerations naturally motivate a fundamental question:

How efficient are discrete diffusion models? When is sublinear convergence possible?

1.1. Sampling efficiency and adaptivity

To put our discussion in context, there has been substantial progress in understanding the sample efficiency of continuous diffusion models. Seminal work by Chen et al. (2023b) characterizes the iteration complexity of the DDPM sampler under Lipschitz (or smoothness) assumptions on the score functions across all steps. Subsequent studies significantly relax these conditions and establish convergence guarantees for broader classes of continuous distributions (Benton et al., 2024; Li et al., 2023; Chen et al., 2023a). Nevertheless, it is now well understood that for general distributions, a linear dependence on the ambient dimension d is unavoidable. By contrast, when the target distribution exhibits additional structure — such as Gaussian mixture models or support on low-dimensional manifolds — a growing body of work shows that popular samplers can adaptively

exploit intrinsic low-dimensional geometry, achieving improved efficiency without explicit dimension reduction (see, e.g., Li and Yan (2024); Li et al. (2025); Huang et al. (2024)).

The landscape shifts considerably as we move to discrete diffusion models. Under the CTMC formulation, algorithms such as Gillespie's method and uniformization allow for exact simulation of the reverse process, free of discretization error (Gillespie, 1976; Van Dijk, 1992; Chen and Ying, 2025). However, these methods suffer from high computational costs in high-dimensional settings. Moreover, their convergence guarantees are inherently stochastic, as they depend on a random number of transitions. An alternative and widely adopted approach, particularly in diffusion-based language models, is provided by τ-leaping and its variants, including truncated τ-leaping (Gillespie, 2001; Campbell et al., 2022). Originally developed in chemical kinetics, τ-leaping replaces sequential state transitions with parallel updates across coordinates, offering substantial computational gains in large systems. Yet, our theoretical understanding of τ-leaping remains incomplete. Current state-of-the-art results exhibit at least a linear dependence on vocabulary size S, linear dependence on d for the absorbing case, and quadratic dependence on d for the uniform case (Liang et al. (2025a,b); Conforti et al. (2025)); see Table 1 for more details. It remains an open question whether these dependencies are fundamental information-theoretic barriers or merely analytical artifacts. Furthermore, as in the continuous setting, an ideal sampling algorithm should automatically adjust to the intrinsic difficulty of the target distribution. For example, one would expect substantially faster convergence for Dirac delta measures or uniform target distributions, without prior knowledge of the structure or modifications to the algorithm. Existing analyses of τ-leaping do not illuminate whether such adaptivity is possible. More specifically, we aim to address the question:

Can score-based samplers automatically adapt to structured target distributions?

1.2. Our contributions

The contributions of this work are centered on establishing sharp convergence guarantees for discrete diffusion models, bridging the gap between empirical success and theoretical understanding. Specifically, our results are threefold:

Optimal rates for uniform diffusion: We establish that for the uniform diffusion process, the τ-leaping sampler requires only $\widetilde{O}(d/\varepsilon)$ discretization steps to achieve an ε-error in KL divergence. This result significantly sharpens the previously best-known bound of $\widetilde{O}(d^2S/\varepsilon)$ (Liang et al., 2025b), effectively removing a factor of d and the dependence on the vocabulary size S.

Fundamental lower bounds: We demonstrate that the linear dependence on the dimension d is essentially unimprovable for the τ-leaping algorithm. Specifically, we show that under uniform diffusion, an $o(d)$ complexity bound is unattainable unless the target distribution is already proximal to the uniform measure. This result characterizes a fundamental price of sampling for informative distributions.

Adaptivity for masking diffusion: For the masking diffusion process, we introduce a refined τ-leaping sampler, whose complexity is governed by $\widetilde{O}(\mathcal{D}/\varepsilon)$, where $\mathcal{D}$ is the effective total correlation, an information-theoretic measure of the target distribution's intrinsic complexity. Notably, while $\mathcal{D}$ is always bounded by the classical total correlation and dual total correlation (and thus by $d \log S$), it can be sublinear or even $O(1)$ for highly structured data, allowing our sampler to automatically adapt to low-dimensional target distributions.

In contrast to prior work, our upper bounds do not require boundedness of the score estimator or any auxiliary regularity assumptions beyond a control on the score entropy loss. The key technical

Paper	Noising process	Score Est. Assump.	No Early Stopping	Sampler	Iteration Complexity	Adaptation
Ren et al. (2025)	Uniform	Bounded	✗	τ-leaping	d^2S^2/ε	✗
Liang et al. (2025b)	Uniform	Bounded	✗	τ-leaping	d^2S/ε	✗
Our work, Theorems 1&2	Uniform	No requirement	✓	τ-leaping	d/ε *	✗
Liang et al. (2025a)	Masking	Bounded	✗	τ-leaping	dS/ε	✗
Conforti et al. (2025)	Masking	$\hat{s}_t \approx s_t$	✗	DMPM	dS/ε	✗
Our work, Theorem 3	Masking	No requirement	✓	Algorithm 1	$\mathcal{D}/\varepsilon$	✓

Table 1: Comparison with prior work. Logarithmic factors in the iteration complexity are omitted. Ren et al. (2025) and Liang et al. (2025a) describe bounds without early stopping under more stringent assumptions on the target distribution, the score function, or the score estimator. The bound in Conforti et al. (2025) depends on additional quantities involving the score estimator beyond Assumption 1, which are small whenever $s_t \approx \hat{s}_t$. The quantity $\mathcal{D}$ (defined in Eqn. (10)), is upper bounded by $d\log(S)$ and captures the intrinsic low-dimensional structure of the target distribution. Entry marked with $*$ indicates sharp rates, with matching lower bounds established in Theorem 2.

ingredient includes a Girsanov change-of-measure argument, combined with establishing martingale properties of the sampling dynamics, which effectively separates the approximation error from the discretization error, allowing each to be analyzed independently. For the lower bound, we leverage a log-Sobolev inequality together with a strong data-processing inequality along the uniform noising process. We connect results with interpretable information-theoretic quantities.

1.3. Notation

For a positive integer n, we denote $[n] := \{1, \dots, n\}$, $\mathbf{1}_n = (1, \dots, 1) \in \mathbb{R}^n$, and $I_n \in \mathbb{R}^{n\times n}$ as the identity matrix. Let $d > 0$ denote the number of dimensions and $S > 0$ denote the vocabulary size. Let MASK denote a special value outside of $[S]$. Let $\mathcal{X} := \mathcal{V}^d$ denote the domain, where, depending on the context, $\mathcal{V} := [S]$ or $\mathcal{V} := [S] \cup \{\text{MASK}\}$. We denote the set of all distributions on $\mathcal{X}$ by $\mathcal{P}(\mathcal{X})$. Let $\mathcal{H}$, KL, and I denote *entropy*, *Kullback-Leibler (KL) divergence*, and *mutual information*, respectively. Let δ_x denote the Dirac measure at point x. We adopt the standard asymptotic notation $O(\cdot), \Omega(\cdot), \Theta(\cdot)$, $\lesssim$, and $\ll$. Additionally, $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}(\cdot)$ are defined analogously except the logarithmic dependency on d, S, and $1/\varepsilon$ is hidden. For a vector $x = (x^1, x^2, \dots, x^d) \in \mathcal{X}$, $i \in [d]$, and $c \in \mathcal{V}$, we define vectors $x^{-i} := (x^1, \dots, x^{i-1}, x^{i+1}, \dots, x^d)$, and $x \oplus_i c, x \odot_i c \in \mathcal{X}$ as follows:

- for all $j \neq i$, $(x \oplus_i c)^j = x^j$, and $(x \oplus_i c)^i = (x^i + c) \bmod |\mathcal{V}|$ [1],
- for all $j \neq i$, $(x \odot_i c)^j = x^j$, and $(x \odot_i c)^i = c$,

For $x, y \in \mathcal{X}$, denote the Hamming distance by $\mathrm{d_H}(x, y) := |\{i \in [d] : x^i \neq y^i\}|$.

1. In this case, we assume that $\mathcal{V}$ has additive structure. We only apply this notation when $\mathcal{V} = [S]$. We use the convention that $0 \bmod S = S$.

2. Preliminaries of discrete diffusion models

2.1. A continuous-time Markov chain formulation

Our goal is to model d-dimensional discrete data $X_0 = (X_0^1, X_0^2, \ldots, X_0^d) \in [S]^d$. Let $q_{\text{data}} = q_0$ denote the probability mass function (p.m.f.) of X_0 that we aim to sample from, and let q_0^i be the marginal p.m.f. of the i-th coordinate. Analogous to continuous diffusion models, the discrete counterparts comprise a forward and a reverse process over the discrete space.

The forward process. We define a forward noising process that progressively transforms the data distribution q_0 to a distribution q_T that is close to an easy-to-sample distribution. This process is modeled using a continuous-time Markov chain (CTMC).

Definition 1 (Continuous-time Markov chain) *A* CTMC *with initial distribution q_0 and rate matrices $(Q_t)_{t\in[0,T]}$ is a right continuous stochastic process $(x_t)_{t\in[0,T]}$ such that*

- *$(x_t)_{t\in[0,T]}$ satisfies the* Markov property*: for any $0 \leqslant s < t \leqslant T$, the conditional distribution of x_t given the history $\{x_u, u \leqslant s\}$ depends only on x_s,*
- *for any $0 \leqslant t < T$, the transition probabilities satisfy, as $\Delta t \to 0^+$:*

$$\Pr(x_{t+\Delta t} = y \mid x_t = x) = \mathbb{I}\{x = y\} + Q_t(x, y)\Delta t + o(\Delta t). \tag{1}$$

Here, the rate matrices satisfy $Q_t(x, y) \geqslant 0$ for all $x \neq y \in \mathcal{X}$ and $Q_t(x, x) = -\sum_{y\neq x} Q_t(x, y)$.

We refer to Feller (1940); Feinberg et al. (2014) for a rigorous treatment. The marginals (q_t) satisfying Eqn. (1) are the solutions to the *Kolmogorov forward equation*: $\mathrm{d}q_t/\mathrm{d}t = Q_t^\top q_t$.

The reverse process. For such CTMC, there exists a time-reversed process with initial distribution q_T, rate matrices $(\overleftarrow{Q}_t)_{t\in[0,T]}$, and marginals $(\overleftarrow{q}_t)_{t\in[0,T]}$, such that $q_t \equiv \overleftarrow{q}_{T-t}$, for $t \in [0, T]$. The forward and reverse rate matrices are explicitly related (Campbell et al., 2022) by

$$\overleftarrow{Q}_t(x, y) = Q_{T-t}(y, x)\frac{q_{T-t}(y)}{q_{T-t}(x)}, \qquad \text{for } x \neq y \in \mathcal{X} \text{ and } 0 \leqslant t \leqslant T. \tag{2}$$

In this paper, we focus on rate matrices that is (1) time-homogeneous, $Q_t \equiv Q$, (2) $Q_t(x, y) = 0$ whenever $\mathrm{d_H}(x, y) \geqslant 2$, and satisfies (3) if $\mathrm{d_H}(x, y) = 1$ and $x^i \neq y^i$, then $Q_t(x, y) = Q^{\text{tok}}(x^i, y^i)$, for some fixed matrix Q^{tok}. In particular, we consider two important instances of CTMCs that are widely-adopted in practice, namely the *uniform noising process* and the *masking* (or absorbing) *noising process*, which are defined through the choice of Q^{tok}.

- **uniform noising process**: A CTMC is a *uniform noising process*, if $Q^{\text{tok}} = \frac{1}{S}\mathbf{1}_S\mathbf{1}_S^\top - I_S$. This CTMC converges to the uniform distribution on the domain $\mathcal{X} := [S]^d$ in the limit.
- **masking noising procces**: A CTMC on the domain $\mathcal{X} := ([S] \cup \{\text{MASK}\})^d$ is a *masking noising process*, if

$$Q^{\text{tok}}(a, b) = \mathbb{I}\{a \neq \text{MASK and } b = \text{MASK}\}, \qquad \text{for } a \neq b \in [S] \cup \{\text{MASK}\}. \tag{3}$$

  The corresponding CTMC converges to the Dirac measure $(\delta_{\text{MASK}})^{\otimes d}$ as $t \to \infty$. Note that we constrain the initial distribution q_0 to be supported on non-masked data, i.e., on $[S]^d$.

2.2. Score estimation

Recall that the reverse process is a CTMC with rate matrices satisfying relation (2), which is similar to the reverse process in the continuous case. The density ratio here generalizes the typical score function $\nabla_x \log q_t(x)$ in the continuous case, and is often referred to as the (concrete) score function for discrete diffusion models (Meng et al., 2022). Formally, for $x \neq y \in \mathcal{X}$, we define *the score function* $s_t(y, x)$ as $s_t(y, x) = \frac{q_t(y)}{q_t(x)}$.

Score entropy loss. For both uniform and masking noising processes, the marginals (q_t), and consequently the score function, are intractable in general. In practice, one therefore resorts to an approximation $\widehat{s}_t(y, x)$ of the true score function $s_t(y, x)$, which is learned from data sampled from the target distribution q_0. To evaluate the quality of the estimated score, a widely used loss function is the *score entropy loss*, originally introduced in Lou et al. (2024), which has since become the de facto standard for training score-based discrete diffusion models. This loss provides a principled objective for matching the approximate score $\widehat{s}_t$ to the true score induced by the forward diffusion process. Specifically, for $t \geqslant 0$ and functions $\widehat{s}, s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the score entropy loss $\mathcal{L}_{\mathrm{SE}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{SE}}(t, \widehat{s}, s) := \mathbb{E}_{x \sim q_t}\Big[\sum_{y \neq x} Q_t(y, x) s(y, x) D(\widehat{s}(y, x), s(y, x))\Big] \geqslant 0.$$

Here, for $a, b \geqslant 0$, $D(a, b) := \frac{a}{b} - 1 - \log \frac{a}{b} \geqslant 0$ is the Bregman divergence for $\phi(a) = -\log a$.

In practice, to implement any sampling algorithm, one has to discretize the continuous dynamics and obtain score estimates at discrete time steps. Suppose score estimates $\widehat{s}_{T-t}$ are obtained at discrete time points $0 \leqslant t_0 < t_1 < \ldots < t_N \leqslant T$. We make the following standard assumption regarding the score estimation errors.

Assumption 1 (Approximation error) *Let $N > 0$ and $0 \leqslant t_0 < t_1 < \ldots < t_N \leqslant T$. We assume*

$$\sum_{k=0}^{N-1} (t_{k+1} - t_k) \mathcal{L}_{\mathrm{SE}}(T - t_k, \widehat{s}_{T-t_k}, s_{T-t_k}) \leqslant \varepsilon_{\mathrm{score}}. \tag{4}$$

This assumption is concerned with the aggregated estimation errors over all N steps. Several works have constructed estimates that satisfy this assumption; examples include Lou et al. (2024); Ou et al. (2025); Benton et al. (2024).

2.3. Score-based sampling algorithms

Armed with the score estimates $(\widehat{s}_{T-t})_{t \in \{t_0, \ldots, t_N\}}$, the objective is to construct a generative model $\widehat{q}_0$ that approximates the data distribution q_0. A natural approach proposed in Campbell et al. (2022) is to define a surrogate CTMC that starts from an easy-to-sample distribution $p_0 \approx q_T$ and approximates the backward dynamics in (2). Concretely, we define the time-inhomogeneous rate matrix

$$\widehat{Q}_t(x, y) = Q_{T-t}(y, x) \widehat{s}_{T-t}(y, x). \tag{5}$$

In practice, score estimates are only available on a fixed discretization $\tau = (t_0, \ldots, t_N)$, and extending these estimates to the full interval $[0, T]$ introduces *discretization error*.

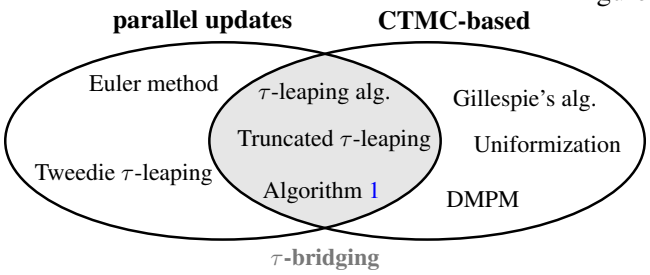

Figure 1: Overview of score-based samplers. The left part comprises score-based samplers which allow parallel updates, defined as τ-leaping strategies in Lou et al. (2024). The right part are the samplers that can be defined through the CTMC framework. At the intersection are τ-bridging strategies, defined in Eqn. (7).

τ-leaping algorithm. As mentioned above, a widely-adopted sampler is the τ-leaping algorithm (Campbell et al., 2022) which approximates Eqn. (5) with multiple possible transitions within each discretization interval. Formally, for $k \in \{0, \dots, N-1\}$ and $t \in [t_k, t_{k+1})$, given x_{t_k} and $\widehat{s}_{T-t_k}$, τ-leaping obtains $x_{t_{k+1}}$ as a random vector, whose coordinates are sampled independently via d one-dimensional CTMCs: for each $i \in [d]$, the initial distribution is $\delta_{x^i_{t_k}}$ and rate matrices are [2]:

$$\widehat{Q}^i_t(a,b) = \widehat{Q}_{T-t_k}(x_{t_k}, x_{t_k} \oplus_i (b-a)), \qquad \text{for } a \neq b \in \mathcal{V}. \tag{6}$$

The formulation in Eqn. (6), however, requires either an additive structure on the state space or the restriction that each coordinate undergoes at most one transition between discretization points. Existing analyses for uniform and masking diffusions (Campbell et al., 2022; Liang et al., 2025b) adopt the latter assumption. In Section 3.1, we explore the necessity of this requirement for the uniform noising process. Lou et al. (2024) generalizes τ-leaping by introducing a class of samplers termed *τ-leaping strategies*, which allow arbitrary transformations $x^i_{t_{k+1}} = f^i_k(\widehat{s}_{T-t_k}, x_{t_k})$. Both the Euler method and Tweedie τ-leaping fall into this class. They however remain challenging for direct theoretical analysis due to the absence of CTMC structure.

This paper: τ-bridging strategies. We introduce a structured class of samplers that generalizes the τ-leaping algorithm while keeping it theoretically tractable. We name this class of algorithms the *τ-bridging strategies*, which retains the parallel updating structure, while remaining analytically tractable. A τ-bridging strategy generates $x_{t_{k+1}}$ from x_{t_k} by evolving d independent one-dimensional CTMCs on $[t_k, t_{k+1})$. For each coordinate $i \in [d]$, the chain is initialized at $\delta_{x^i_{t_k}}$ and has a rate matrix

$$\widehat{Q}^i_t = G^i_t(\widehat{s}_{T-t_k}, x_{t_k}), \tag{7}$$

for some mapping $G^i_t : \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}} \times \mathcal{X} \to \mathbb{R}^{\mathcal{V}\times\mathcal{V}}$. Compared to general τ-leaping strategies, τ-bridging strategies restrict updates to CTMC-based transitions. This restriction preserves parallel coordinate updates while facilitating theoretical analysis. Figure 1 summarizes the relationships among these classes of sampling algorithms.

An representative instance of a τ-bridging sampler is the *truncated τ-leaping* sampler of Liang et al. (2025b). For $k \in [N]$ and $t \in [t_k, t_{k+1})$, the corresponding rate matrices take the form

$$G^i_t(\widehat{s}_{T-t_k}, x_{t_k})(a,b) = Q_{T-t_k}(x_{t_k} \odot_i b, x_{t_k})\widehat{s}_{T-t_k}(x_{t_k}, x_{t_k} \odot_i b)\mathbb{I}\{x^i_{t_k} = a\} \quad \text{for } a \neq b \in \mathcal{V}. \tag{8}$$

2. The algorithm admits an equivalent Poisson formulation, in which dS Poisson random variables corresponding to coordinate-value transitions are sampled and applied in parallel.

The indicator $\mathbb{I}\{x^i_{t_k} = a\}$ enforces at most one transition per coordinate $i \in [d]$ within each discretization interval $[t_k, t_{k+1})$. In Section 3.2, we show that an instance of this scheme achieves sublinear complexity for the masking noising process under mild distributional assumptions. To the best of our knowledge, this is the first result establishing such a guarantee.

3. Main results

In this section, we characterize the sampling efficiency of samplers in the class of τ-bridging strategies, for both uniform and masking noising processes. We develop sharp convergence guarantees and point out cases where adaptivity is automatically achieved. Proof sketches for all results appear in Appendix B, with full proofs deferred to the corresponding sections of the Appendix.

3.1. Uniform discrete diffusion

3.1.1. A SHARP CONVERGENCE CHARACTERIZATION

We begin with the uniform discrete diffusion models, whose forward dynamics are given by the uniform noising proecss. We establish explicit sampling guarantees for the τ-leaping algorithm, measured in KL divergence. The proof is given in Appendix D.1.

Theorem 1 *Let $q_{\text{data}} = q_0$ be the data distribution on $\mathcal{X} := [S]^d$. For $0 = t_0 < t_1 < \ldots < t_N = T$, let $\Delta := \max_k\{t_{k+1} - t_k\} = O(1)$. Set $p_0 = \text{Unif}(\mathcal{X})$. Under Assumption 1, the τ-leaping algorithm initialized at p_0 generates a sample from $p_{\text{output}} = p_T$, such that*

$$\mathsf{KL}(q_{\text{data}} \,\|\, p_{\text{output}}) \lesssim \varepsilon_{\text{score}} + e^{-T} d \log(S) + \Delta d \log(S/\Delta). \tag{9}$$

As expected, the KL divergence bound in Theorem 1 decomposes into three terms. The first term $\varepsilon_{\text{score}}$ quantifies the quality of score estimation and captures the accumulation of estimation errors over the N discretization steps. The second term corresponds to the initialization error, arising from initializing the sampler with the uniform distribution p_0 instead of the true terminal distribution q_T; this term decays exponentially in the diffusion horizon T. Finally, the third term accounts for the discretization error incurred by approximating the continuous-time reverse process with a discrete-time τ-leaping scheme.

To further interpret Theorem 1 and place it in context with existing results, we highlight several of its salient features. First, the discretization error scales linearly with the dimension d and only logarithmically with the vocabulary size S. This matches the result obtained for the random walk model (Conforti et al., 2025) and reveals that the discretization error is insensitive to the distribution scale, as has been shown for continuous diffusion models (e.g., Huang et al. (2024)). Second, the theorem permits a flexible choice of step size schedules and does not require early stopping. In contrast to prior analyses that rely on carefully selected step sizes and introduce an early stopping time δ (where the algorithm outputs $p_{T-\delta}$ in place of p_T), the bound in Theorem 1 depends only on the maximum step size. Moreover, the same bound applies uniformly to early stopping variants: the right-hand side of (9) remains unchanged for any $\delta \ll 1$. The only requirement we have on score estimation is Assumption 1, with no additional boundedness or regularity conditions (typically assumed in existing literature). As a result, the theorem applies to a broad class of score estimation procedures commonly used in practice.

Next, we specialize Theorem 1 to a concrete choice of discretization schedule to derive the iteration complexity required to obtain an ε-accurate sampler in KL divergence. For a simple step

size schedule, it turns out that d/ε steps (up to logarithmic factors) suffice for convergence, significantly improving upon the state-of-the-art complexity of d^2S/ε from Liang et al. (2025b). Refer to Appendix D.2 for the proof.

Corollary 1 *For the setting in Theorem 1 and $\varepsilon > 0$, the output of the τ-leaping algorithm with constant step size schedule $t_{k+1} - t_k = T/N$ for $k \in [N-1]$, achieves $\mathsf{KL}(q_{\text{data}} \,\|\, p_{\text{output}}) \lesssim \varepsilon_{\text{score}} + \varepsilon$, provided that time horizon $T = \log(d\log(S)/\varepsilon)$ and iteration number $N = \widetilde{O}\left(d/\varepsilon\right)$.*

Other step size schedules commonly adopted in practice, such as exponential-then-constant and log-linear (Lou et al., 2024), satisfy the same iteration complexity as discussed in Appendix D.2.

3.1.2. A MATCHING LOWER BOUND FOR τ-LEAPING

While Theorem 1 establishes an upper bound for τ-leaping algorithm scaling nearly linearly with the dimension d, and logarithmically with the vocabulary size S, the fundamental question remains: is this dependence an intrinsic limit or merely a technical artifact? We show that the former is indeed the case by establishing a matching lower bound.

We note that for target distributions sufficiently close to uniform, sampling can be achieved with very few steps, as the forward CTMC converges efficiently to its limit. To avoid these pathological instances, we restrict our focus to the class of distributions that remain sufficiently well-separated from the uniform distribution. Specifically, for any $\gamma \in [0,1]$, define subset $\mathcal{P}^\gamma(\mathcal{X}) \subseteq \mathcal{P}(\mathcal{X})$ as

$$\mathcal{P}^\gamma(\mathcal{X}) = \left\{q_0 \in \mathcal{P}(\mathcal{X}) : \mathcal{H}(q_1) \leqslant (1-\gamma)\cdot \mathcal{H}(\text{Unif}(\mathcal{X})) = (1-\gamma)d\log(S)\right\},$$

where q_1 is the marginal distribution at $t = 1$ of the uniform noising process initialized at q_0, $\text{Unif}(\mathcal{X})$ is the uniform distribution on $\mathcal{X}$, and $\mathcal{H}(\cdot)$ denotes the entropy function of a distribution. Intuitively, for $\gamma \in (0,1)$, the class $\mathcal{P}^\gamma(\mathcal{X})$ imposes a structural constraint on the convergence of the forward process; it describes distributions that do not mix rapidly. In this sense, for $\gamma = O(1)$, $\mathcal{P}^\gamma(\mathcal{X})$ contains distributions that remain informative enough in the forward process when $t = 1$, which covers most of the interesting distributions in practice, since they carry non-trivial information characterized by relatively low entropy.

When sampling from a distribution in $\mathcal{P}^\gamma(\mathcal{X})$ with τ-leaping algorithm, it turns out that the iteration complexity bound in Corollary 1 can not be improved up to logarithmic factors. We formalize this statement with the following lower bound. The proof is given in Appendix D.3.

Theorem 2 *For any target distribution $q_0 \in \mathcal{P}^\gamma(\mathcal{X})$ and early stopping time $0 \leqslant \delta \ll 1$, denote the path measure of the backward process by $Q \overset{d}{=} \{\overleftarrow{q}_t\}_{t\in[0,T-\delta]}$ and the sampling process by $P \overset{d}{=} \{p_t\}_{t\in[0,T-\delta]}$. Let $\gamma = \Omega(1)$. Then, for any step size schedule $0 = t_0 < t_1 < \ldots < t_N = T-\delta$ with $\max_k\{t_{k+1} - t_k\} \leqslant \frac{1}{2}$ it takes τ-leaping algorithm at least*

$$N = \Omega(d\log(S))$$

number of iterations to achieve

$$\mathsf{KL}(Q \,\|\, P) \leqslant \varepsilon_{\text{score}} + O(1).$$

We make several remarks concerning the nature and implications of our lower bound.

Theorem 2 reveals that for informative target distributions in $\mathcal{P}^{\gamma}(\mathcal{X})$, ensuring that the KL divergence between the sampling process and the reverse process is small requires the number of steps to scale at least linearly with the dimension d, which cannot be avoided for general distributions. In addition, the lower bound is uniform over both early stopping schedules ($0 < \delta \ll 1$) and non-early stopping schemes ($\delta = 0$).

This lower bound is algorithm-dependent: it relies on structural properties of the τ-leaping algorithm and therefore differs fundamentally from information-theoretic or minimax lower bounds. In principle, alternative sampling schemes may circumvent the linear dependence on d. Indeed, in Section 3.2, we show that a modified τ-leaping procedure achieves sublinear dependence on d for structured target distributions under the masking noising process. Whether analogous improvements are possible for uniform discrete diffusion through modified algorithms remains an open question.

When the target distribution has high entropy, the lower bound need not apply. Indeed, when q_{data} satisfies $\mathsf{KL}(q_{\mathrm{data}} \| \mathrm{Unif}(\mathcal{X})) = o(d)$, one can show that $\mathcal{H}(q_1) = \Theta(d \log S)$, and that a sample from a distribution with the KL error at most $\varepsilon_{\mathrm{score}} + \varepsilon$ can be obtained using $N = o(d)$ steps. A precise formulation of this claim is given in Appendix D.4.

We remark that the quantity controlled in Theorem 2 is the KL divergence between two path measures, rather than the divergence between the terminal output distributions, which may appear weaker than the upper bound in Corollary 1. However, to the best of our knowledge, all existing upper-bound analyses for the KL divergence, including ours, proceed by first bounding the KL divergence between path measures and then invoking the data-processing inequality. Consequently, the lower bound applies to all current analysis techniques. In this sense, Theorem 2 establishes the optimality of the iteration complexity in Corollary 1 within the scope of the existing analysis techniques.

3.2. Masking discrete diffusion

We now turn our attention to the masking noising process. Our main result in this setting is an upper bound that intrinsically depends on the structural properties of the target distribution q_{data}, rather than scaling with the ambient dimension d. This aligns with the intuition that for highly structured distributions — such as a sparse mixture of Dirac measures — a sensible sampler should converge at a sublinear scale, or perhaps even logarithmically, with d.

3.2.1. PRELIMINARIES

We begin by recalling two fundamental quantities in information theory: total correlation and dual total correlation: for a distribution q over $[S]^d$ and $x \sim q$, The **Total correlation** $\mathcal{C}(q)$ and **Dual total correlation** $\mathcal{B}(q)$, are defined as

$$\mathcal{C}(q) := \sum_{i=1}^{d} \mathcal{H}(x^i) - \mathcal{H}(x) \qquad \text{and} \qquad \mathcal{B}(q) := \mathcal{H}(x) - \sum_{i=1}^{d} \mathcal{H}(x^i \mid x^{-i}).$$

We now introduce a time-dependent quantity associated with the masking noising process. Consider a masking noising process defined by Eqn. (3) with marginals $(q_t)_{t \geqslant 0}$. For $x \in ([S] \cup \{\mathrm{MASK}\})^d$ and $i \neq j \in [d]$, let $x^{-(i,j)}$ denote the collection of all unmasked elements of x, excluding i-th and j-th coordinates. For $x_t \sim q_t$, define

$$\mathcal{I}(t) := \sum_{i \neq j \in [d]} \mathrm{I}(x_t^i; x_t^j \mid x_t^{-(i,j)}) \geqslant 0 \quad \text{and} \quad \mathcal{D}(q_0) := \int_0^\infty \min(1, t)\mathcal{I}(t)\mathrm{d}t, \tag{10}$$

where $\mathrm{I}(A; B \mid C)$ denotes the conditional mutual information. We refer to $\mathcal{D}(q_0)$ as the *effective total correlation* of the target distribution. Lemma 16 shows that total correlation and dual total correlation can be expressed through $\mathcal{I}(t)$ by

$$\mathcal{B}(q_0) = \int_0^\infty \mathcal{I}(t)\mathrm{d}t \quad \text{and} \quad \mathcal{C}(q_0) = \int_0^\infty (e^t - 1)\mathcal{I}(t)\mathrm{d}t.$$

Consequently, $\mathcal{D}(q_0) \leqslant \min(\mathcal{B}(q_0), \mathcal{C}(q_0))$. The statement and its proof are given in Appendix F.1.

Note that both $\mathcal{B}(q_0), \mathcal{C}(q_0)$, and hence $\mathcal{D}(q_0)$ are upper bounded by $d \log(S)$. Moreover, there exist distributions q_0 with $\mathcal{B}(q_0) = O(1)$ while $\mathcal{C}(q_0) = \Omega(d \log(S))$, and vice versa. We refer to Austin (2020) for a detailed study of total correlation and dual total correlation. Importantly, there also exist natural distributions, for which both $\mathcal{B}(q_0)$ and $\mathcal{C}(q_0)$ are of order d, while $\mathcal{D}(q_0)$ remains small. See Proposition 5 for an example of such distribution.

3.2.2. AN ADAPTIVE CHARACTERIZATION

Equipped with the above notation, we present our main result on masking diffusion models.

Theorem 3 *Let $q_{\mathrm{data}} = q_0$ be a distribution on $[S]^d$. For $0 = t_0 < t_1 < \ldots < t_N = T$, let $h_k := t_{k+1} - t_k$ be the step size and assume that $\Delta := \max_k h_k = O(1)$. Let*

$$p_0 := \left((1 - e^{-T})\delta_{\mathrm{MASK}} + S^{-1} e^{-T} \sum_{k=1}^{S} \delta_k \right)^{\otimes d}.$$

Under Assumption 1, Algorithm 1 initialized at p_0 produces a sample from $p_{\mathrm{output}} = p_T$, such that

$$\mathsf{KL}(q_{\mathrm{data}} \,\|\, p_{\mathrm{output}}) \lesssim \varepsilon_{\mathrm{score}} + e^{-T} d \log(S) + \sum_{k=0}^{N-1} h_k \int_{T - t_{k+1}}^{T - t_k} \mathcal{I}(t)\mathrm{d}t. \tag{11}$$

The proof is given in Appendix E.1, and a few remarks on the consequences and implications of Theorem 3 are in order.

As in Theorem 1, the last term in the upper bound corresponds to the discretization error measured using integrated mutual information defined in Eqn. (10). While the first two terms are generic, the third term governs the dependence on the dimension d and reflects the information-theoretic properties of the target distribution. For structured distributions, our algorithm implicitly adapts to the underlying structure of the target distribution without requiring any prior knowledge of that structure or any modification to the algorithm itself.

We analyze the performance of truncated τ-leaping as an alternative to Algorithm 1 in Appendix E.3, which has an additional d/N^2 term in the upper bound Eqn. (11), ignoring lower-order contributions. Although for structured target distributions the resulting iteration complexity already scales as $\sqrt{d}$ rather than d (as in the existing literature), it does not fully adapt to the geometry of the target distribution. To provide some intuition, the standard (or truncated) τ-leaping algorithms informally satisfies for $t \in [t_k, t_{k+1})$ (see Eqn. (8))

$$G_t^i(s_{T-t_k}, x_{t_k}) \approx G_{t_k}^i(s_{T-t_k}, x_{t_k}), \quad \text{and thus} \quad \widehat{Q}_t \approx \overleftarrow{Q}_{t_k}, \tag{12}$$

Algorithm 1: Modified truncated τ-leaping

Input:
Initial distribution: p_0,
Discretization steps: $0 = t_0 < t_1 < \ldots < t_N = T$,
Score estimate function: $\widehat{s}_{T-t}$ for $t \in \{t_0, \ldots, t_{N-1}\}$.
Output: Sample $\widehat{x} \in [S]^d$.
Sample x_0 from p_0
for $k = 0, \ldots, N-1$ **do**
  for $i \in m(x_{t_k}) := \{i,\ \textit{such that}\ x^i_{t_k} = \mathrm{MASK}\}$ **do**
    $\widehat{Q}^i_k(a) \leftarrow \widehat{s}_{T-t_k}(x_{t_k} \odot_i a, x_{t_k}), \quad \text{for } a \in [S]$
    $\widehat{Q}^i_k(\mathrm{MASK}) \leftarrow -\sum_{a \in [S]} \widehat{Q}^i_k(a)$
    if $k < N-1$ **then**
      $\Delta_k \leftarrow (e^{T-t_k} - 1) \log\left(\frac{e^T - e^{t_k}}{e^T - e^{t_{k+1}}}\right)$
      $\mathcal{P}_k \leftarrow \exp(\widehat{Q}^i_k(\mathrm{MASK})\Delta_k)$
    end
    else
      $\mathcal{P}_k \leftarrow 0$
    end
    $$x^i_{t_{k+1}} \leftarrow \begin{cases} \mathrm{MASK}, & \text{with probability } \mathcal{P}_k, \\ a, & \text{with probability } \frac{\widehat{Q}^i_k(a)}{\sum_{b \in [S]} \widehat{Q}^i_k(b)}(1 - \mathcal{P}_k), \ \text{for } a \in [S]. \end{cases}$$
  end
end
return x_{t_N}

where we recall the mapping G^i_t from Eqn. (7). That is, even when the score estimation is exact, $\widehat{s}_{T-t_k} \equiv s_{T-t_k}$, the τ-leaping algorithm introduces a mismatch between the surrogate and true rate matrices as $s_{T-t_k} \not\equiv s_{T-t}$. Algorithm 1 corrects this discrepancy by enforcing

$$G^i_t(s_{T-t_k}, x_{t_k}) \approx G^i_t(s_{T-t}, x_{t_k}), \quad \text{and thus} \quad \widehat{Q}_t \approx \overleftarrow{Q}_t, \tag{13}$$

through the rescaling of the score estimate function: $\widehat{s}_{T-t} = \frac{e^{T-t_k}-1}{e^{T-t}-1}\widehat{s}_{T-t_k}$. As it is a linear transformation of the score estimate function, we can simulate its dynamics only at discrete points $T - t_0, \ldots, T - t_N$, see Algorithm 1 and Lemma 13. This leads to a sharper upper bound in Theorem 3 relative to the analogous bound for truncated τ-leaping (Theorem 5; see also Remark 3). Empirically, the benefit of rescaling the score function in masking discrete diffusion models has also been observed in prior work; see, for example, Lou et al. (2024); Ou et al. (2025).

Notably, our results are closely connected to an intriguing parallel line of work on masking diffusion models (Li and Cai (2025); Chen et al. (2025)), which focuses on the design of unmasking schedules without adopting a CTMC perspective. In particular, Chen et al. (2025) derives optimal unmasking schedules and discusses two representative instances, in which the number of steps scales linearly with $\mathcal{B}(q_{\mathrm{data}})$ and $\mathcal{C}(q_{\mathrm{data}})$, respectively. Their algorithms require an a priori estimate of $\mathcal{B}(q_{\mathrm{data}})$ and $\mathcal{C}(q_{\mathrm{data}})$ or a doubling search procedure to calibrate the unmasking schedule, and

rely on a different sampling mechanism. The fact that our score-based samplers automatically exploit similar information-theoretic quantities without additional hyperparameters underscores both the fundamental nature of these quantities and the robustness of the CTMC framework.

Next, we derive iteration complexity guarantees for our algorithm under specific choices of step size schedules, whose proof is given in Appendix E.2.

Corollary 2 *Consider the setting in Theorem 3. Let* $T = \Theta(\log(d\log(S)))$. *For fixed* $\varepsilon > 0$, *it obeys* $\mathsf{KL}(q_{\mathrm{data}} \,\|\, p_{\mathrm{output}}) \lesssim \varepsilon_{\mathrm{score}} + \varepsilon$,

- *under constant step size schedule,* $t_k - t_{k-1} = T/N$ *for all* $k \in [N]$, *for* $N = \widetilde{O}\left(\mathcal{B}(q_{\mathrm{data}})/\varepsilon\right)$;
- *under exponential-then-constant step size schedule, when* $t_{k+1} - t_k \leqslant \kappa \min(1, T - t_{k+1})$ *for* $k \in \{0, \ldots, N-2\}$, $T - t_{N-1} = \varepsilon/(d\log(S))$, *and* $\kappa = N^{-1}(T + \log(\varepsilon^{-1} d\log(S)))$, *for*
$$N = \widetilde{O}\left(\mathcal{D}(q_{\mathrm{data}})/\varepsilon\right) \leqslant \widetilde{O}(\min\{\mathcal{B}(q_{\mathrm{data}}), \mathcal{C}(q_{\mathrm{data}})\}/\varepsilon).$$

In words, Corollary 2 shows that the sampling complexity of Algorithm 1 required to obtain an ε-accurate distribution is governed by intrinsic complexity measures of the target distribution, which are dimension-independent up to logarithmic factors for distributions such as uniform and sparse mixture of Dirac measures. Under a constant step size schedule, the iteration complexity is controlled by the dual total correlation of the target distribution, whereas under an exponential-then-constant schedule, the effective total correlation becomes the relevant quantity.

Important examples and adaptation. To further illustrate the implications of Theorem 3, we consider some representative distributions for which one or more of the quantities $\mathcal{B}(q_{\mathrm{data}}), \mathcal{C}(q_{\mathrm{data}})$, or $\mathcal{D}(q_{\mathrm{data}})$ are small. Since the iteration complexity scales linearly with these quantities, our result shows that discrete diffusion models can provably achieve efficient sampling. Appendix A develops these examples in detail and provides rigorous proofs of the stated claims.

- **Hidden Markov models.** Here, the observed variables correspond to words or tokens in a sentence, while the hidden states encode latent semantic topics. Under the natural assumption that topics evolve slowly, we show that $\mathcal{B}(q_{\mathrm{data}})$ grows sublinearly with the sequence length.
- **Low-dimensional structures.** Motivated by image generation, when the discrete data arise from a quantization of a continuous distribution with intrinsic dimension k, the dual total correlation $\mathcal{B}(q_{\mathrm{data}})$ scales linearly with k rather than with the ambient dimension d.
- **Random graph models.** Such models define distributions over $d = \binom{n}{2}$ binary variables corresponding to edges of a graph with n vertices. Besides Erdős-Rényi random graphs, which have independent edges and are therefore easy to sample, we consider both sparse random regular graphs and stochastic block models. In these cases, $\mathcal{B}(q_{\mathrm{data}})$ grows at most linearly (up to logarithmic factors) with n, rather than quadratically.
- **Latent parity model.** Finally, we present an example in which both total correlation $\mathcal{C}(q_{\mathrm{data}})$ and dual total correlation $\mathcal{B}(q_{\mathrm{data}})$ are of order d, while $\mathcal{D}(q_{\mathrm{data}})$ remains of constant order. Such distributions are motivated by applications such as error-correcting codes and DNA sequences, where substantial noise may be present, yet the underlying signal is highly structured.

4. Discussion

In this work, we establish novel theoretical results for both uniform and masking discrete diffusions. For uniform diffusion models, we show that a τ-leaping algorithm requires $\widetilde{O}(d/\varepsilon)$ iterations to achieve ε accuracy in KL divergence, improving upon the prior bound $\widetilde{O}\left(d^2 S/\varepsilon\right)$. We further establish the first algorithmic lower bound for the τ-leaping sampler, which shows that the upper bound we achieved is unimprovable for a large class of distributions. For masking discrete diffusion, we derive an upper bound that captures the intrinsic complexity of the data distribution and can scale logarithmically with the ambient dimension. Importantly, our results for both models only require small score estimation error and, in contrast to prior work, do not rely on early stopping or boundedness assumptions of the score estimator.

The improved bound for the masking noising process is achieved via a modification of the τ-leaping algorithm. This modification falls within a structured subclass of τ-leaping strategies that (i) allow for parallel coordinate updates, and thus sublinear rates, and (ii) preserve CTMC dynamics, which facilitates theoretical analysis. We hope that this perspective motivates the development of adaptive samplers for uniform discrete diffusion as well in the future.

Several other open questions remain. Understanding which noising mechanisms — masking, uniform, or discrete Gaussian — are best suited to different classes of target distributions is an important direction for future work. Moreover, the problem of learning accurate score functions in discrete diffusion models remains largely unexplored and warrants further investigation.

Acknowledgments

This work is supported in part by the NSF grants CCF-2106778, CCF-2418156 and CAREER award DMS2143215.

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Tim Austin. Multi-variate correlation and mixtures of product measures. *Kybernetika*, pages 459–499, July 2020. ISSN 0023-5954, 1805-949X. doi: 10.14736/kyb-2020-3-0459.

Francis Bach and Saeed Saremi. Sampling binary data by denoising through score functions. *arXiv preprint arXiv:2502.00557*, 2025.

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

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

Hongrui Chen and Lexing Ying. Convergence analysis of discrete diffusion model: exact implementation through uniformization. *Journal of Machine Learning*, 4(2):108–127, June 2025. ISSN 2790-2048, 2790-203X. doi: 10.4208/jml.240812.

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

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

Sitan Chen, Kevin Cong, and Jerry Li. Optimal inference schedules for masked diffusion models. *arXiv preprint arXiv:2511.04647*, 2025.

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

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Eugene A Feinberg, Manasa Mandava, and Albert N Shiryaev. On solutions of kolmogorov's equations for nonhomogeneous jump markov processes. *Journal of Mathematical Analysis and Applications*, 411(1):261–270, 2014.

Willy Feller. On the integro-differential equations of purely discontinuous markoff processes. *Transactions of the American Mathematical Society*, 48(3):488–515, 1940.

Mark Gales and Steve Young. The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304, 2024.

Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of chemical physics*, 115(4):1716–1733, 2001.

Alexander N Gorban and Ivan Yu Tyukin. Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2118):20170237, 2018.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *arXiv preprint arXiv:2410.18784*, 2024.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in Neural Information Processing Systems*, 32, 2019.

Gen Li and Changxiao Cai. Breaking AR's sampling bottleneck: Provable acceleration via diffusion language models. *Advances in Neural Information Processing Systems*, 38, 2025.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *Advances in Neural Information Processing Systems*, 37:126297–126331, 2024.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

Gen Li, Changxiao Cai, and Yuting Wei. Dimension-free convergence of diffusion models for approximate gaussian mixtures. *arXiv preprint arXiv:2504.05300*, 2025.

Lizao Li, Robert Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13), 2024.

Yuchen Liang, Renxiang Huang, Lifeng Lai, Ness Shroff, and Yingbin Liang. Absorb and converge: Provable convergence guarantee for absorbing discrete diffusion models. *Advances in Neural Information Processing Systems*, 39, 2025a.

Yuchen Liang, Yingbin Liang, Lifeng Lai, and Ness Shroff. Discrete diffusion models: Novel analysis and new sampler guarantees. *Advances in Neural Information Processing Systems*, 39, 2025b.

Anita Liebenau and Nick Wormald. Asymptotic enumeration of graphs by degree sequence, and the degree sequence of a random graph. *Journal of the European Mathematical Society*, 26:1–40, 2024.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2024.

Anuran Makur and Yury Polyanskiy. Comparison of channels: Criteria for domination by a symmetric channel. *IEEE Transactions on Information Theory*, 64(8):5704–5725, 2018.

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

Bhavya Mor, Sunita Garhwal, and Ajay Kumar. A systematic review of hidden markov models and their applications. *Archives of computational methods in engineering*, 28(3), 2021.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yong-Hyun Park, Chieh-Hsin Lai, Satoshi Hayakawa, Yuhta Takida, and Yuki Mitsufuji. Jump your steps: Optimizing sampling schedule of discrete diffusion 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 *International Conference on Machine Learning*, 2025.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

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 *International Conference on Learning Representations*, 2025.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

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

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Nico M Van Dijk. Uniformization for nonhomogeneous markov chains. *Operations research letters*, 12(5):283–291, 1992.

Dimitri von Rütte, Janis Fluri, Omead Pooladzandi, Bernhard Schölkopf, Thomas Hofmann, and Antonio Orvieto. Scaling behavior of discrete diffusion language models. *arXiv preprint arXiv:2512.10858*, 2025.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

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

Appendix Contents

Appendix A. Examples of low intrinsic dimensions

A.1. Details and formal results

In this section, we revisit the examples outlined in Section 3.2.2 and develop them in full detail. We formalize the statements in this section, and provide rigorous proofs in Appendix A.2.

Hidden Markov models. A hidden Markov model (HMM) is a Markov model in which the observations are dependent on a latent Markov process (referred to as z), which is widely used in natural language processing and other pattern recognition tasks (Mor et al., 2021; Gales and Young, 2024). For example, in natural language modeling, the hidden states z^i can represent the underlying grammatical structure or semantic meaning of paragraphs, while the observed states x^i correspond to the actual words or tokens in a sentence.

Consider the following HMM: $\{z^i\}_{i\in[d]}$ is the discrete hidden state, Markov chain supported on $\mathcal{Z}$ and $\{x^i\}_{i\in[d]}$ is the observed chain generated by $x^i = f_i(z^i, \varepsilon^i)$, where $\{\varepsilon^i\}_{i\in[d]}$ are i.i.d. noise variables independent of $\{z^i\}_{i\in[d]}$. If z^i indicates the semantic topic of the i-th paragraph in a document, we can expect that in natural language, the topic transitions are rare, i.e., z^i is likely to remain the same as z^{i-1} for $i > 1$. For data generated from this model, we can show the following proposition:

Proposition 1 *For the above HMM, suppose the transition probability of $\{z^i\}_{i\in[d]}$ is at most p, i.e., $\Pr(z^i \neq z^{i-1}) \leqslant p$ for all $i \in \{2, \ldots, d\}$. When $1/d \lesssim p \ll 1$, it holds that*

$$\mathcal{B}(q_{\mathrm{data}}) \leqslant pd \log\left(\frac{|\mathcal{Z}|}{p}\right).$$

To develop some intuition, consider generating a document with a constant number of paragraphs, where the transition probability scales $p = \Theta(1/d)$. Suppose further that the latent space $\mathcal{Z} \in [S]^k$ for some $k \ll d$ and S denotes the vocabulary size. Then, the above bound yields

$$\mathcal{B}(q_{\mathrm{data}}) \lesssim k \log(Sd),$$

which is much smaller than the ambient dimension $d \log(S)$. As such, with Theorem 3, the sampling complexity scales with the intrinsic topic dimension k rather than the document length d.

Low-dimensional Structures. In image generation tasks and some other tasks with structured data, it is commonly assumed that the data lies on or near a low-dimensional manifold embedded in a high-dimensional space, which is often referred to as the manifold hypothesis (Gorban and Tyukin, 2018; Pope et al., 2021). For example, natural images may be viewed as points on a manifold parameterized by a small number of underlying factors, such as lighting conditions, pose, and object identity.

In discrete settings, the notion of a manifold is not mathematically well defined. To capture low-dimensional structure, we instead model the data as arising from a continuous mapping from a

latent representation into a high-dimensional observation space. For some latent continuous random variable z supported on $\mathcal{Z} \subset \mathbb{R}^k$, consider a decoding procedure $f : [0,1]^k \to \mathbb{R}^d$ as

$$x^{\text{con}} = f(z) + \varepsilon_{\text{noise}},$$

for additive perturbations $\varepsilon_{\text{noise}}$. Thus, data lies close to a manifold $\{f(z) : z \in \mathcal{Z}\}$. The final discrete observation is obtained via a quantization operator $\mathcal{Q}_S$, i.e., $x = \mathcal{Q}_S(x^{\text{con}}) \sim q_{\text{data}}$.

To align the model with standard image processing pipelines, we work with the uniform lattice quantization function $\mathcal{Q}_S : \mathbb{R}^d \to [S]^d$ defined coordinate-wise as $[\mathcal{Q}_S(x)]^i = \text{clip}(\lfloor x^i \rfloor, 0, S)$ for $i \in [d]$, where $\text{clip}(x, a, b) := \min\{\max\{x, a\}, b\}$ is the clip function and $\lfloor \cdot \rfloor$ is the floor function. To ensure regularity of both the manifold and the induced data distribution, we focus on the case where $\mathcal{Z}$ is a compact set and f is a Lipschitz function. The noise $\varepsilon_{\text{noise}}$ is taken to be Gaussian for simplicity of analysis; the arguments extend readily to more general smooth noise distributions.

Proposition 2 *Let $\mathcal{Z} \subset \mathbb{R}^k$ be compact with diameter D, and let $f : \mathcal{Z} \to \mathbb{R}^d$ be L-Lipschitz. Assume the noise satisfies $\varepsilon_{\text{noise}} \sim \mathcal{N}(0, \sigma^2 I_d)$ independenly generated for each obersvation. Then the resulting distribution satisfies*

$$\mathcal{B}(q_{\text{data}}) \leqslant k \log\left(2 + \frac{2\mathsf{DL}}{\sigma}\right). \tag{14}$$

In image generation, the "ideal image" x^{con} may be interpreted as the vector of continuous pixel intensities prior to quantization, while the observed image x is obtained by applying pixel-wise quantization to x^{con}. When $k \ll d$, the above bound yields

$$\mathcal{B}(q_{\text{data}}) = \widetilde{O}(k) = o(d),$$

and hence we can efficiently sample such images despite the high dimensionality of the observation space.

Random graph models. Discrete diffusion models have also found applications in scientific domains such as molecular generation and protein design, where data are naturally represented as random graphs with fixed vertex sets and random edges (Ingraham et al., 2019; Xu et al., 2022). To make this concrete, we consider two widely studied random graph models on n vertices, which can be viewed as a discrete distribution over adjacency matrices of dimension n^2.

- **Regular graphs**: A k-regular graph is a graph in which each vertex has degree exactly k. Suppose we want to sample a random graph $\mathcal{G}$ from some distribution supported on the set of k-regular graphs with n vertices.

  Proposition 3 *For sparse regular graph model, i.e., $k \leqslant n/\log(n)$, we have*

  $$\mathcal{B}(\mathcal{G}) \lesssim kn \log\left(\frac{n}{k}\right) = o(n^2). \tag{15}$$

- **Stochastic block models**: A stochastic block model (SBM) is a generative model for random graphs that captures community structure within networks. In an SBM, n vertices of the graph are partitioned into r distinct communities or blocks, represented by latent variables $\{z^i\}_{i \in [n]}$

taking values in $[r]$. Conditioned on the latent labels, edges are generated independently. For two vertices $i, j \in [n]$, an edge is created with probability

$$p\mathbb{I}\{z^i = z^j\} + q\mathbb{I}\{z^i \neq z^j\},$$

where $p, q \in [0, 1]$ govern the within- and between-community connection probabilities, respectively.

Proposition 4 *Let $\mathcal{G}$ be a random graph drawn from the above r-block SBM. Then*

$$\mathcal{B}(\mathcal{G}) \leqslant n \log(r) = o(n^2).$$

For both random graph models, as the number of vertices n grows large, the complexity satisfies $\mathcal{B}(\mathcal{G}) = o(n^2)$, which is strictly smaller compared to the ambient dimension n^2. This indicates that diffusion-based methods can sample efficiently from such graph distributions.

In fact, the analyses of Propositions 2 and 4 extend naturally to generalized random geometric graphs. Consider the following example. Let each vertex $i \in [n]$ be associated with latent variable $z^i \in \mathcal{Z}$. For distinct vertices i and j, an edge is placed independently with probability

$$\beta \exp\left(-\frac{d(z^i, z^j)}{r_0}\right),$$

where $\beta \in [0, 1]$, $r_0 > 0$ and $d(\cdot, \cdot)$ is an appropriate metric in the latent space $\mathcal{Z}$.

- When latent variables $\{z^i\}$ are discrete with $o(n)$ entropy, as is the case, for example, when it takes value in a fixed-dimensional latent space, the dual total correlation of the resulting random graph is $o(n^2)$.
- For continuous latent variables, suppose $\mathcal{Z} = \mathcal{S}^{d_z - 1}$, the unit sphere in $\mathbb{R}^{d_z}$. Under some regularity conditions, the dual total correlation scales with $d_z \cdot n$, followed by an analogous covering number argument in Proposition 2. In particular, whenever $d_z = o(n)$, the complexity is again subquadratic, leading to sublinear (in n^2) convergence rates for diffusion-based sampling.

Latent parity model A prototypical example of a distribution with small dual total correlation $\mathcal{B}(q_0)$ and large total correlation $\mathcal{C}(q_0)$ is the mixture of two Dirac measures:

$$p_m := \frac{1}{2}\delta_{\mathbf{0}} + \frac{1}{2}\delta_{\mathbf{1}},$$

where $\mathbf{0}$ and $\mathbf{1}$ are vectors of all-zeros and all-ones, respectively. It can be easily computed that $\mathcal{B}(p_m) = \log(2)$, whereas, $\mathcal{C}(p_m) = (d-1)\log(2)$.

The opposite happens, for instance, for the following XOR distribution p_{XOR}:

$$x^1, \ldots, x^{d-1} \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(1/2) \quad \text{and} \quad x^d = \sum_{i=1}^{d-1} x^i \bmod 2.$$

In this case, $\mathcal{B}(p_{\mathrm{XOR}}) = (d-1)\log(2)$, and $\mathcal{C}(p_{\mathrm{XOR}}) = \log(2)$.

Real-world data distributions can combine features of both extremes: a strong low-dimensional signal corrupted by weakly correlated noise. In such cases, both $\mathcal{B}(q_{\mathrm{data}})$ and $\mathcal{C}(q_{\mathrm{data}})$ can be large, while $\mathcal{D}(q_{\mathrm{data}})$ remains small. To illustrate this phenomenon, consider the following entrywise mixture of the two preceding examples.

1. Fix a bi-partition $[d] = I_0 \sqcup I_1$ for non-empty index sets I_0 and I_1;
2. For all indices $i \in I_0$, set $x_i = b$ for $b \sim \mathrm{Bern}(1/2)$;
3. Among all indices $i \in I_1$, sample all but one $x_i \sim \mathrm{Bern}(1/2)$ independently;
4. For the last index i^*, set $x_{i^*} = \big(b + \sum_{i \neq i^*} x_i\big) \bmod 2$.

Denote this distribution as p_{ex}, and let $x = (x^1, \ldots, x^d) \sim p_{\mathrm{ex}}$.

Proposition 5 *Suppose that* $\min\{|I_0|, |I_1|\}/d = \Theta(1)$. *Distribution* p_{ex} *satisfies*

$$\mathcal{B}(p_{\mathrm{ex}}) = \Theta(d), \quad \mathcal{C}(p_{\mathrm{ex}}) = \Theta(d), \quad \textit{and} \quad \mathcal{D}(p_{\mathrm{ex}}) = O(1).$$

By Proposition 5, p_{ex}, which can be viewed as a non-trivial mixing of p_m and p_{XOR}, satisfies

$$\mathcal{D}(p_{\mathrm{ex}}) \ll \min\{\mathcal{B}(p_{\mathrm{ex}}), \mathcal{C}(p_{\mathrm{ex}})\}.$$

This example highlights the fundamental role of the effective total correlation in characterizing sampling efficiency.

A.2. Proofs of results in Section A.1

Variants of the following lemma will be used repeatedly throughout this section. We state it here for convenience and to streamline the proofs.

Lemma 1 *Consider any d-dimensional discrete random variable X and any random variable W such that $X^i \perp\!\!\!\perp X^{-i} \mid W$ for any $i \in [d]$, where $X = (X^1, \ldots, X^d)$ and X^{-i} is the $(d-1)$-dimensional marginal of X with i-th coordinate excluded. Then,*

$$\mathcal{B}(X) \leqslant \mathrm{I}(X; W).$$

If W is discrete, we additionally have $\mathcal{B}(X) \leqslant \mathcal{H}(W)$.

Proof of Lemma 1. We first notice that for any random variable W such that $X^i \perp\!\!\!\perp X^{-i} \mid W$ for any $i \in [d]$, we have

$$\mathcal{H}(X^i \mid X^{-i}) \geqslant \mathcal{H}(X^i \mid X^{-i}, W) = \mathcal{H}(X^i \mid W),$$

where the first inequality follows from the definition of the entropy. Recalling the definition of $\mathcal{B}(\cdot)$, we obtain

$$\mathcal{B}(X) = \mathcal{H}(X) - \sum_{i=1}^{d} \mathcal{H}(X^i | X^{-i}) \leqslant \mathcal{H}(X) - \sum_{i=1}^{d} \mathcal{H}(X^i \mid W).$$

Using the conditional independence condition again, we have

$$\mathcal{H}(X|W) = \mathcal{H}\big((X_1, \ldots, X_d)|W\big) = \sum_{i=1}^{d} \mathcal{H}(X^i \mid W),$$

which implies

$$\mathcal{B}(X) \leqslant \mathcal{H}(X) - \mathcal{H}(X|W) = I(X; W) \overset{\text{(a)}}{=} \mathcal{H}(W) - \mathcal{H}(W|X) \overset{\text{(b)}}{\leqslant} \mathcal{H}(W),$$

where (a) and (b) apply when W is a discrete random variable.

A.2.1. PROOF OF PROPOSITION 1

The hidden Markov structure of $\{(x^i, z^i)\}_{i\in[d]}$ satisfies $x^i \perp\!\!\!\perp x^j \mid (z^i, z^j)$, since $\varepsilon^i \perp\!\!\!\perp \varepsilon^j \mid (z^i, z^j)$. Considering Lemma 1 above, we can upper bound $\mathcal{B}(q_{\text{data}})$ by $\mathcal{H}(z)$, which is the entropy of the latent Markov chain. By the additivity of the entropy, we have

$$\mathcal{B}(q_{\text{data}}) \leqslant \mathcal{H}(z) = \mathcal{H}(z^1) + \sum_{i=2}^{d} \mathcal{H}\big(z^i|\{z^j\}_{j\in[i-1]}\big) = \mathcal{H}(z^1) + \sum_{i=2}^{d} \mathcal{H}(z^i|z^{i-1}).$$

When $\{z^i\}_{i\in[d]}$ is supported on a single point, we have $|\mathcal{Z}| = 1$ and $\mathcal{H}(z) = 0$. When the state space $\mathcal{Z}$ satisfies $2 \leqslant |\mathcal{Z}| < \infty$, the maximum entropy distribution is achieved when

$$z^1 \sim \text{Unif}(\mathcal{Z}) \quad \text{and} \quad z^i|z^{i-1} \sim (1-p)\delta_{z^{i-1}} + p\text{Unif}\big(\mathcal{Z} \setminus \{z^{i-1}\}\big).$$

We obtain

$$\begin{aligned}
\mathcal{H}(z) &\leqslant \log(|\mathcal{Z}|) + \sum_{i=2}^{d} \left[-(1-p)\log(1-p) + -(|\mathcal{Z}|-1)\cdot\frac{p}{|\mathcal{Z}|-1}\log\left(\frac{p}{|\mathcal{Z}|-1}\right)\right] \\
&\overset{\text{(a)}}{\leqslant} \log(|\mathcal{Z}|) + (d-1)\cdot\left(2p + p\log\left(\frac{|\mathcal{Z}|}{p}\right)\right) \\
&\overset{\text{(b)}}{\leqslant} pd\log\left(\frac{|\mathcal{Z}|}{p}\right),
\end{aligned}$$

where in (a), we use $-\log(1-p) \leqslant 2p$, since $p \ll 1$; in (b), we use the condition $p \gtrsim 1/d$ and $|\mathcal{Z}|/p \geqslant 2/p \gg 1$. This completes the proof of the desired result.

A.2.2. PROOF OF PROPOSITION 2

Write $\varepsilon_{\text{noise}} = (\varepsilon^1_{\text{noise}}, \ldots, \varepsilon^d_{\text{noise}})$. Since $\varepsilon_{\text{noise}} \sim \mathcal{N}(0, \sigma^2 I_d)$, we have $\varepsilon^i_{\text{noise}} \perp\!\!\!\perp \varepsilon^{-i}_{\text{noise}}$ for any $i \in [d]$. Processing through the decoder f, $[x^{\text{con}}]^i = [f(z)]^i + \varepsilon^i_{\text{noise}}$ for any $i \in [d]$, which leads to

$$[x^{\text{con}}]^i \perp\!\!\!\perp [x^{\text{con}}]^{(-i)} \mid z, \tag{16}$$

where $[x^{\text{con}}]^{(-i)}$ is the $(d-1)$-dimensional marginal of x^{con} with i-th coordinate excluded. Note that $\mathcal{Q}_S$ is a entry-wise quantization, i.e., we can write $\mathcal{Q}_S(x) = (\widetilde{\mathcal{Q}}_S(x^1), \ldots, \widetilde{\mathcal{Q}}_S(x^d))$ for entry-wise deterministic quantization function $\widetilde{\mathcal{Q}}_S : \mathbb{R} \to [S]$, and $x^i = \widetilde{\mathcal{Q}}_S([x^{\text{con}}]^i)$ by the generation process. Eqn. (16) therefore implies that for any $i \in [d]$,

$$x^i \perp\!\!\!\perp x^{-i} \mid z.$$

Applying Lemma 1, we obtain

$$\mathcal{B}(q_{\text{data}}) = \mathcal{B}(x) \leqslant \mathrm{I}(x; z) \leqslant \mathrm{I}(x^{\text{con}}; z), \tag{17}$$

where the last inequality follows from the data processing inequality of the mutual information.

In the following proof, we proceed to control $\mathrm{I}(x^{\text{con}}; z)$. Since $\varepsilon_{\text{noise}}$ is independent noise, using data-processing inequality, we reach

$$\mathrm{I}(x^{\text{con}}; z) \leqslant \mathrm{I}\big(f(z) + \varepsilon_{\text{noise}}; f(z)\big) = \mathrm{I}\big(f(z); f(z) + \varepsilon_{\text{noise}}\big). \tag{18}$$

Without loss of generality, we assume $\mathcal{Z} \subseteq [0, \mathsf{D}]^k$. Partition $[0, \mathsf{D}]^k$ into hypercubes of size $h_J = \sigma/\mathsf{L}$, and write this partition as $\{C_1, \ldots, C_{\lceil \mathsf{D}/h_J \rceil^k}\}$ such that

$$[0, \mathsf{D}]^k \subseteq \bigsqcup_{i=1}^{\lceil \mathsf{D}/h_J \rceil^k} C_i.$$

Define $J = J(z)$ to be the hypercube index $i(z)$ such that $z \in C_{i(z)}$, and $\mathcal{F}_J$ to be σ-algebra generated by $J(z)$. By the chain rule and data processing inequality for mutual information, we have

$$\begin{aligned}
\mathrm{I}\big(f(z); f(z) + \varepsilon_{\text{noise}}\big) &\leqslant \mathrm{I}\big(J(z), f(z); f(z) + \varepsilon_{\text{noise}}\big) \\
&= \mathrm{I}\big(J(z); f(z) + \varepsilon_{\text{noise}}\big) + \mathrm{I}\big(f(z); f(z) + \varepsilon_{\text{noise}} \mid J\big) \\
&\leqslant k \log\left(1 + \frac{\mathsf{D}}{h_J}\right) + \mathrm{I}\big(f(z); f(z) + \varepsilon_{\text{noise}} \mid J\big), \qquad (19)
\end{aligned}$$

where in the last line, we use $\mathrm{I}(J(z); f(z) + \varepsilon_{\text{noise}}) \leqslant \mathcal{H}(J(z)) \leqslant \log(|\mathrm{supp}(J(z))|)$. To upper bound the second term above, we introduce the following lemma on Gaussian channel, whose proof is given in Section G.1.

Lemma 2 *For any random variable $W \in \mathbb{R}^d$ and independent noise $\varepsilon_{\text{noise}} \sim \mathcal{N}(0, \sigma^2 I_d)$, we have*

$$\mathrm{I}(W; W + \varepsilon_{\text{noise}}) \leqslant \frac{\mathrm{Tr}\big(\mathrm{Var}[W]\big)}{2\sigma^2},$$

where $\mathrm{Tr}(\cdot)$ is the trace function.

In Lemma 2, taking $W \overset{d}{=} f(z) \mid \mathcal{F}_J$, we arrive at

$$\mathrm{I}\big(f(z); f(z) + \varepsilon_{\text{noise}} \mid J\big) \leqslant \frac{\mathrm{Tr}\Big(\mathrm{Var}\big[f(z) \mid \mathcal{F}_J\big]\Big)}{2\sigma^2}. \qquad (20)$$

To further control the right hand side, direct calculations show

$$\mathrm{Tr}\Big(\mathrm{Var}\big[f(z) \mid \mathcal{F}_J\big]\Big) = \sum_{i=1}^{d} \mathrm{Var}\big[[f(z)]^i \,\big|\, \mathcal{F}_J\big] = \mathbb{E}\bigg[\Big\|f(z) - \mathbb{E}\big[f(z) \mid \mathcal{F}_J\big]\Big\|_2^2 \,\Big|\, \mathcal{F}_J\bigg]. \qquad (21)$$

It is therefore sufficient to consider the quantity $\|f(z) - \mathbb{E}[f(z) \mid \mathcal{F}_J]\|_2^2$. We make the observation that

$$\begin{aligned}
\Big\|f(z) - \mathbb{E}\big[f(z) \mid \mathcal{F}_J\big]\Big\|_2 &\overset{\text{(a)}}{\leqslant} \sup_{w \in \mathrm{Conv}(f(C_{J(z)}))} \big\|f(z) - w\big\|_2 \\
&\overset{\text{(b)}}{=} \sup_{w \in f(C_{J(z)})} \big\|f(z) - w\big\|_2 \\
&\overset{\text{(c)}}{\leqslant} \|f\|_{\text{Lip}} \cdot \sup_{z' \in C_{J(z)}} \big\|z - z'\big\|_2 \overset{\text{(d)}}{\leqslant} \mathsf{L}\sqrt{k} h_J,
\end{aligned}$$

where $\|\cdot\|_2$ denotes Euclidean norm in $\mathbb{R}^d$, and $\mathrm{Conv}(\cdot)$ denotes the convex null of a given set. In (a), we use the fact that $\mathbb{E}[f(z) \mid \mathcal{F}_J] \in \mathrm{Conv}(f(C_{J(z)}))$; in (b), we adopt f is continuous and hence $f(C_{J(z)}))$ is bounded, and the property of the convex hull that

$$\mathrm{diam}\big(\mathrm{Conv}(A)\big) = \mathrm{diam}(A) \quad \text{for any bounded subset } A \subseteq \mathbb{R}^d;$$

in (c), we recall the Lipschitz condition on f; in (d), we notice that $\mathrm{diam}(C_i) \leqslant \sqrt{k}h_J$ for any hypercube C_i. Putting pieces together gives

$$\mathrm{Tr}\Big(\mathrm{Var}\big[f(z) \mid \mathcal{F}_J\big]\Big) \leqslant \big(\mathsf{L}\sqrt{k}h_J\big)^2 = k\sigma^2. \tag{22}$$

Finally, plugging Eqns. (20) and (22) into Eqn. (19), we obtain

$$\mathrm{I}\big(f(z); f(z) + \varepsilon_{\mathrm{noise}}\big) \leqslant k \log\left(1 + \frac{\mathsf{DL}}{\sigma}\right) + \frac{k}{2} \leqslant k \log\left(2 + \frac{2\mathsf{DL}}{\sigma}\right).$$

Combining the above inequality with Eqns. (17) and (18), we conclude

$$\mathcal{B}(q_{\mathrm{data}}) \leqslant \mathrm{I}(x^{\mathrm{con}}; z) = \mathrm{I}\big(f(z); f(z) + \varepsilon_{\mathrm{noise}}\big) \leqslant k \log\left(2 + \frac{2\mathsf{DL}}{\sigma}\right).$$

A.2.3. PROOF OF PROPOSITION 3

Define the set of all k-regular graphs with n vertices as $G_{n,k}$. Without loss of generality, we assume that nk is even, as otherwise $G_{n,k}$ is empty. By a corollary of Liebenau and Wormald (2024, Theorem 1.4), we have the following asymptotic result:

$$|G_{n,k}| = \Theta\left(\binom{n-1}{k}^n \binom{\frac{n(n-1)}{2}}{m} \binom{n(n-1)}{2m}^{-1}\right).$$

where $m = kn/2$. By Stirling's formula of the form

$$\log(a!) = a\log(a) - a + O(\log(a)),$$

we can compute that

$$\begin{aligned}
\log(|G_{n,k}|) &\lesssim n \log\binom{n-1}{k} + \log\binom{\frac{n(n-1)}{2}}{m} - \log\binom{n(n-1)}{2m} \\
&= \frac{kn}{2}\log\left(\frac{n-1-k}{k}\right) + \frac{n(n-1)}{2}\log\left(\frac{n-1}{n-1-k}\right) \\
&\leqslant \frac{kn}{2}\log\left(\frac{n}{k}\right) + \frac{n^2}{2}\log\left(1 + \frac{k}{n-1-k}\right) \\
&\leqslant \frac{kn}{2}\log\left(\frac{n}{k}\right) + \frac{kn^2}{2(n-1-k)} \lesssim kn\log\left(\frac{n}{k}\right),
\end{aligned}$$

where in the last line, we use the condition that $k \leqslant n/\log(n) \ll n-1-k$. Recalling the definition of $\mathcal{B}(\cdot)$, we have

$$\mathcal{B}(\mathcal{G}) \leqslant \mathcal{H}(\mathcal{G}) \leqslant \log(|G_{n,k}|) \lesssim kn\log\left(\frac{n}{k}\right) = o(n^2).$$

A.2.4. PROOF OF PROPOSITION 4

By the definition of r-block SBM, the latent variable $\{z^i\}_{i\in[n]}$ supported on $[r]^n$, which satisfies

$$\mathcal{H}\big(\{z^i\}_{i\in[n]}\big) \leqslant \log\big(|[r]^n|\big) = n\log(r).$$

Given the latent variable $\{z^i\}_{i\in[n]}$, the block structure is fixed and hence each edge is sampled independently from Bernoulli distributions. Therefore, we have

$$e^{ij} \perp\!\!\!\perp e^{k\ell} \mid \{z^i\}_{i\in[n]}$$

for any $i, j, k, l \in [n]$, where e^{ij} and $e^{k\ell}$ are the indicator variables of the existence of edges between vertices i, j and between vertices k, ℓ. By Lemma 1, we conclude

$$\mathcal{B}(\mathcal{G}) \leqslant \mathcal{H}(\{z^i\}_{i\in[n]}) \leqslant n\log(r) \leqslant n\log(n) = o(n^2),$$

where we use the convention that the block number $r \leqslant n$.

Remark 1 *The setting of Proposition 4 can be viewed as a special case of the generalized random geometric graph model, in which the latent variable corresponds to the block index. More generally, the same conclusion holds under analogous assumptions, with essentially the same proof strategy.*

A.2.5. PROOF OF PROPOSITION 5

Let $r := |I_0|/d$ be the proportion of coordinates in I_0. Throughout, we assume $\min\{r, 1-r\} = \Theta(1)$.

Step 1: Establish $\mathcal{B}(p_{\mathrm{ex}}) = \Theta(d)$ and $\mathcal{C}(p_{\mathrm{ex}}) = \Theta(d)$. For a random variable $x \sim p_{\mathrm{ex}}$, we shall demonstrate that

$$\sum_{i=1}^{d} \mathcal{H}(x^i) = d\log(2), \quad \log(2)(|I_1|-1) \leqslant \mathcal{H}(x) \leqslant \log(2)|I_1| \quad \text{and} \quad \sum_{i=1}^{d} \mathcal{H}(x^i \mid x^{-i}) = 0. \tag{23}$$

Towards this goal, we make the observation that for any $i \in I_0$ or $x \in I_1 \backslash i^*$, $x^i \sim \mathrm{Bern}(1/2)$ and hence $\mathcal{H}(x^i) = \log(2)$. For $i = i^*$, we assert that $x^{i^*} \sim \mathrm{Bern}(1/2)$. In fact, we have

$$\mathbb{P}\left(\sum_{i\in I_1\backslash i^*} x^i \equiv 0 \bmod 2\right) = \mathbb{P}\left(\mathrm{Bin}\left(|I_1|-1, \frac{1}{2}\right) \equiv 0 \bmod 2\right) = \frac{1}{2},$$

where in the last equality, we invoke the following lemma.

Lemma 3 *For any $n \in \mathbb{N}^+$ and $X \sim \mathrm{Bin}(n, 1/2)$, we have*

$$\mathbb{P}(X \equiv 0 \bmod 2) = \mathbb{P}(X \equiv 1 \bmod 2) = \frac{1}{2}.$$

As result, the distribution of x^{i^*} satisfies

$$\mathbb{P}(x^{i^*}=0)=\mathbb{P}(b=0)\cdot\mathbb{P}\left(\sum_{i\in I_1\backslash i^*}x^i\equiv 0 \bmod 2\right)+\mathbb{P}(b=1)\cdot\mathbb{P}\left(\sum_{i\in I_1\backslash i^*}x^i\equiv 1 \bmod 2\right)=\frac{1}{2},$$

which reveals that $x^{i^*}\sim\mathrm{Bern}(1/2)$ and hence $\mathcal{H}(x^{i^*})=\log(2)$. In conclusion, we obtain

$$\sum_{i=1}^{d}\mathcal{H}(x^i)=\sum_{i\in[d]\backslash i^*}\mathcal{H}(x^i)+\mathcal{H}(i^*)=d\log(2). \tag{24}$$

To upper bound $\mathcal{H}(x)$, invoke the simple property for entropy function to get

$$\mathcal{H}(x)\leqslant\log(|\mathrm{supp}(x)|)\leqslant\log\left(2\cdot 2^{|I_1|-1}\right)=\log(2)|I_1|. \tag{25}$$

The lower bound can be obtained through

$$\mathcal{H}(x)\geqslant\mathcal{H}(\{x^i\}_{i\in I_1\backslash i^*})=\log\left(2^{|I_1|-1}\right)=\log(2)(|I_1|-1). \tag{26}$$

For any $i\in[d]$, when x^{-i} is given, we can recover x^i by first observing the value of b from x^j for any $j\in I_0$, then applying the formula

$$x^i=b+\sum_{k\in I_1\backslash i}x^k\mathbb{I}\{i\in I_1\} \bmod 2.$$

Thus, $x^i\mid x^{-i}$ is always a Dirac measure, which leads to

$$\sum_{i=1}^{d}\mathcal{H}(x^i\mid x^{-i})=0. \tag{27}$$

Combining Eqns. (24), (25), (26) and (27) proves Eqn. (23).

Equipped with Eqn. (23), we are ready to bound $\mathcal{B}(p_{\mathrm{ex}})$ and $\mathcal{C}(p_{\mathrm{ex}})$. It can easily seen that

$$\mathcal{B}(p_{\mathrm{ex}})=\mathcal{H}(x)-\sum_{i=1}^{d}\mathcal{H}(x^i\mid x^{-i})\geqslant\log(2)\big((1-r)d-1\big)=\Omega(d),$$

$$\mathcal{C}(p_{\mathrm{ex}})=\sum_{i=1}^{d}\mathcal{H}(x^i)-\mathcal{H}(x)\geqslant\log(2)(d-|I_1|)=\log(2)rd=\Omega(d).$$

For the reverse direction, we can prove the matching lower bound similarly, which leads to

$$\mathcal{B}(p_{\mathrm{ex}})=\Theta(d),\quad\mathcal{C}(p_{\mathrm{ex}})=\Theta(d).$$

Step 2: Show $\mathcal{D}(p_{\rm ex}) = O(1)$. Recall the definition of $\mathcal{D}(\cdot)$ in Eqn. (10):

$$\mathcal{D}(p_{\rm ex}) \coloneqq \int_0^\infty \min(1,t)\mathcal{I}(t)\mathrm{d}t \quad \text{with} \quad \mathcal{I}(t) \coloneqq \sum_{i\neq j\in[d]} \mathrm{I}(x_t^i; x_t^j \mid x_t^{-(i,j)}) \geqslant 0.$$

To upper bound $\mathcal{D}(p_{\rm ex})$, let us write

$$\mathcal{D}(p_{\rm ex}) = \int_0^{\frac{1}{d}} t\,\mathcal{I}(t)\mathrm{d}t + \int_{1/d}^{\log(d)} \min\{1,t\}\,\mathcal{I}(t)\mathrm{d}t + \int_{\log(d)}^\infty \mathcal{I}(t)\mathrm{d}t.$$

By direct calculations, one has

$$\begin{aligned}
\int_0^{\frac{1}{d}} t\,\mathcal{I}(t)\mathrm{d}t &\leqslant \frac{1}{d}\int_0^{\frac{1}{d}} \mathcal{I}(t)\mathrm{d}t \leqslant \frac{\mathcal{B}(p_{\rm ex})}{d} = \Theta(1),\\
\int_{\log(d)}^\infty \mathcal{I}(t)\mathrm{d}t &\leqslant \frac{1}{d-1}\int_{\log(d)}^\infty (e^t-1)\mathcal{I}(t)\mathrm{d}t \leqslant \frac{\mathcal{C}(p_{\rm ex})}{d-1} = \Theta(1).
\end{aligned}$$

Therefore, it obeys

$$\mathcal{D}(p_{\rm ex}) = \int_{1/d}^{\log(d)} \min\{1,t\}\,\mathcal{I}(t)\mathrm{d}t + O(1).$$

To prove $\mathcal{D}(p_{\rm ex}) = O(1)$, it suffices to show that

$$\int_{1/d}^{\log(d)} \min\{1,t\}\,\mathcal{I}(t)\mathrm{d}t = O(1). \tag{28}$$

In view of the definition of $\mathcal{I}(t)$, we can decompose it as

$$\begin{aligned}
\mathcal{I}(t) &= \left(\sum_{i,j\in I_0, i\neq j} + \sum_{i,j\in I_1, i\neq j} + \sum_{i\in I_0, j\in I_1} + \sum_{i\in I_1, j\in I_0}\right) \mathrm{I}(x_t^i; x_t^j \mid x_t^{-(i,j)})\\
&\coloneqq \mathcal{I}_1(t) + \mathcal{I}_2(t) + \mathcal{I}_3(t) + \mathcal{I}_4(t),
\end{aligned}$$

and we shall bound these four terms separately.

Before diving into the proofs, we make the observation that the mutual information can be computed via

$$\mathrm{I}(x_t^i; x_t^j \mid x_t^{-(i,j)}) = \mathcal{H}(x_t^i \mid x_t^{-(i,j)}) - \mathcal{H}(x_t^i \mid x_t^{-i}). \tag{29}$$

To further compute each entropy terms, let us introduce two quantities below

$$\mathcal{H}_t^1 = \mathcal{H}(e^{-t}\delta_0 + (1-e^{-t})\delta_{\rm MASK}) = \mathcal{H}(e^{-t}\delta_1 + (1-e^{-t})\delta_{\rm MASK}) = te^{-t} - \log(1-e^{-t})(1-e^{-t}), \tag{30a}$$

$$\mathcal{H}_t^2 = \mathcal{H}\left(\frac{1}{2}e^{-t}\delta_0 + \frac{1}{2}e^{-t}\delta_1 + (1-e^{-t})\delta_{\rm MASK}\right) = (t+\log(2))e^{-t} - \log(1-e^{-t})(1-e^{-t}). \tag{30b}$$

We shall relate our quantities of interest to these terms below.

Case 1: $i, j \in I_0$**,** $i \neq j$**.** For any given $x_t^{-(i,j)}$, it always holds true that

$$\mathbb{P}(x_t^i = \text{MASK}) = 1 - e^{-t},$$

since the noising process is time-homogeneous and independent between coordinates. Recall the definition $m(x) = \{i \in [d] : x^i = \text{MASK}\}$. Define the event $\mathcal{E}_{t,1}^{i,j} \in \mathcal{F}_t^{-(i,j)}$, where $\mathcal{F}_t^{-(i,j)}$ is the σ-algebra generated by $x_t^{-(i,j)}$, as follows:

$$\mathcal{E}_{t,1}^{i,j} := \left\{ x_t^{-(i,j)} : \left(\bigvee_{k \in I_0 \setminus \{i,j\}} \{k \notin m(x_t)\} \right) \bigvee \left(\bigwedge_{\ell \in I_1} \{\ell \in m(x_t)\} \right) = 1 \right\},$$

where $\wedge$ is the logical operator AND, and $\vee$ is the logical operator OR. By construction of p_{ex}, it can be checked that

$$\begin{aligned}
\left(x_t^i \,\middle|\, x_t^{-(i,j)} \in \mathcal{E}_{t,1}^{i,j} \right) &\sim e^{-t}\delta_{0/1} + (1 - e^{-t})\delta_{\text{MASK}}; \\
\left(x_t^i \,\middle|\, x_t^{-(i,j)} \in (\mathcal{E}_{t,1}^{i,j})^c \right) &\sim \frac{1}{2}e^{-t}\delta_0 + \frac{1}{2}e^{-t}\delta_1 + (1 - e^{-t})\delta_{\text{MASK}},
\end{aligned}$$

where $\delta_{0/1}$ represents either δ_0 or δ_1. Therefore, by the definition of the conditional entropy, we have

$$\mathcal{H}(x_t^i \mid x_t^{-(i,j)}) = \mathcal{H}_t^1 \cdot \mathbb{P}(\mathcal{E}_{t,1}^{i,j}) + \mathcal{H}_t^2 \cdot \big(1 - \mathbb{P}(\mathcal{E}_{t,1}^{i,j})\big). \tag{31}$$

Define the event $\mathcal{E}_{t,1}^i \in \mathcal{F}_t^{-i}$, where $\mathcal{F}_t^{-i}$ is the σ-algebra generated by x_t^{-i}, as follows:

$$\mathcal{E}_{t,1}^i := \left\{ x_t^{-i} : \left(\bigvee_{k \in I_0 \setminus \{i\}} \{k \notin m(x_t)\} \right) \bigvee \left(\bigwedge_{\ell \in I_1} \{\ell \in m(x_t)\} \right) = 1 \right\}.$$

Then, it can be checked similarly that

$$\begin{aligned}
\left(x_t^i \,\middle|\, x_t^{-i} \in \mathcal{E}_{t,1}^i \right) &\sim e^{-t}\delta_{0/1} + (1 - e^{-t})\delta_{\text{MASK}}; \\
\left(x_t^i \,\middle|\, x_t^{-i} \in (\mathcal{E}_{t,1}^i)^c \right) &\sim \frac{1}{2}e^{-t}\delta_0 + \frac{1}{2}e^{-t}\delta_1 + (1 - e^{-t})\delta_{\text{MASK}},
\end{aligned}$$

which leads to

$$\mathcal{H}(x_t^i \mid x_t^{-i}) = \mathcal{H}_t^1 \cdot \mathbb{P}(\mathcal{E}_{t,1}^i) + \mathcal{H}_t^2 \cdot \big(1 - \mathbb{P}(\mathcal{E}_{t,1}^i)\big). \tag{32}$$

Plugging Eqns. (31) and (32) into Eqn. (29) gives that for any $i, j \in I_0$, $i \neq j$,

$$\begin{aligned}
\mathrm{I}(x_t^i; x_t^j \mid x_t^{-(i,j)}) &= \mathcal{H}(x_t^i \mid x_t^{-(i,j)}) - \mathcal{H}(x_t^i \mid x_t^{-i}) \\
&= \mathcal{H}_t^1 \cdot \mathbb{P}(\mathcal{E}_{t,1}^{i,j}) + \mathcal{H}_t^2 \cdot \big(1 - \mathbb{P}(\mathcal{E}_{t,1}^{i,j})\big) - \mathcal{H}_t^1 \cdot \mathbb{P}(\mathcal{E}_{t,1}^i) - \mathcal{H}_t^2 \cdot \big(1 - \mathbb{P}(\mathcal{E}_{t,1}^i)\big) \\
&= (\mathcal{H}_t^2 - \mathcal{H}_t^1)\big(\mathbb{P}(\mathcal{E}_{t,1}^i) - \mathbb{P}(\mathcal{E}_{t,1}^{i,j})\big) \\
&= \log(2) e^{-2t} \left(1 - e^{-t}\right)^{|I_0|-2} \left(1 - e^{-|I_1|t}\right) \\
&= O\left(e^{-2t}(1 - e^{-t})^{rd/2}\right),
\end{aligned}$$

whose value is independent of the indices i and j. Since $|\{i, j \in I_0 : i \neq j\}| = rd(rd-1) = \Theta(d^2)$, quantity $\mathcal{I}_1(t)$ satisfies

$$\mathcal{I}_1(t) = \sum_{i,j \in I_0, i \neq j} \mathrm{I}(x_t^i; x_t^j \mid x_t^{-(i,j)}) = O\left(d^2 e^{-2t}(1 - e^{-t})^{rd/2}\right). \tag{33}$$

Case 2: $i, j \in I_1$**,** $i \neq j$**.** Following the proof strategy in Case 1, for any given $x_t^{-(i,j)}$, it holds that

$$x_t^i \mid x_t^{-(i,j)} \sim \frac{1}{2}e^{-t}\delta_0 + \frac{1}{2}e^{-t}\delta_1 + (1-e^{-t})\delta_{\text{MASK}},$$

which implies that

$$\mathcal{H}(x_t^i \mid x_t^{-(i,j)}) = \mathcal{H}_t^2. \tag{34}$$

Define the event $\mathcal{E}_{t,2}^i \in \mathcal{F}_t^{-i}$ as follows:

$$\mathcal{E}_{t,2}^i := \left\{ x_t^{-i} : \left(\bigvee_{k \in I_0} \{k \notin m(x_t)\} \right) \bigwedge \left(\bigwedge_{\ell \in I_1 \setminus \{i\}} \{\ell \in m(x_t)\} \right) = 1 \right\},$$

which induces

$$\begin{aligned} \left(x_t^i \,\middle|\, x_t^{-i} \in \mathcal{E}_{t,2}^i\right) &\sim e^{-t}\delta_{0/1} + (1-e^{-t})\delta_{\text{MASK}}; \\ \left(x_t^i \,\middle|\, x_t^{-i} \in (\mathcal{E}_{t,2}^i)^c\right) &\sim \frac{1}{2}e^{-t}\delta_0 + \frac{1}{2}e^{-t}\delta_1 + (1-e^{-t})\delta_{\text{MASK}}, \end{aligned}$$

and the conditional entropy formula

$$\mathcal{H}(x_t^i \mid x_t^{-i}) = \mathcal{H}_t^1 \cdot \mathbb{P}(\mathcal{E}_{t,2}^i) + \mathcal{H}_t^2 \cdot \left(1 - \mathbb{P}(\mathcal{E}_{t,2}^i)\right). \tag{35}$$

Plugging Eqns. (34) and (35) into Eqn. (29) gives that for any $i, j \in I_0$, $i \neq j$,

$$\begin{aligned} \mathrm{I}(x_t^i; x_t^j \mid x_t^{-(i,j)}) &= \mathcal{H}(x_t^i \mid x_t^{-(i,j)}) - \mathcal{H}(x_t^i \mid x_t^{-i}) \\ &= \mathcal{H}_t^2 - \mathcal{H}_t^1 \cdot \mathbb{P}(\mathcal{E}_{t,2}^i) - \mathcal{H}_t^2 \cdot \left(1 - \mathbb{P}(\mathcal{E}_{t,2}^i)\right) \\ &= \left(\mathcal{H}_t^2 - \mathcal{H}_t^1\right) \mathbb{P}(\mathcal{E}_{t,2}^i) = O\left(e^{-(1-r)dt}\right), \end{aligned}$$

whose value is, again, independent of the indices i and j. Since $|\{i, j \in I_1 : i \neq j\}| = (1-r)d((1-r)d-1) = \Theta(d^2)$, we reach

$$\mathcal{I}_2(t) = \sum_{i,j \in I_1, i \neq j} \mathrm{I}(x_t^i; x_t^j \mid x_t^{-(i,j)}) = O(d^2 e^{-(1-r)dt}). \tag{36}$$

Case 3: $i \in I_0$**,** $j \in I_1$**.** Define the function $\mathcal{H}_{\mathrm{B}}(p) := -p\log(p) - (1-p)\log(1-p)$ to be the entropy of the distribution $\mathrm{Bern}(p)$. Following the proofs of the two cases above, let us define events

$$\begin{aligned} \mathcal{E}_{t,3}^{i,j} &:= \left\{ x_t^{-(i,j)} : \left(\bigvee_{k \in I_0 \setminus \{i\}} \{k \neq m(x_t)\} \right) \bigvee \left(\bigwedge_{\ell \in I_1 \setminus \{j\}} \{\ell \in m(x_t)\} \right) = 1 \right\}; \\ \mathcal{E}_{t,3}^{i} &:= \left\{ x_t^{-i} : \left(\bigvee_{k \in I_0 \setminus \{i\}} \{k \neq m(x_t)\} \right) \bigvee \left(\bigwedge_{\ell \in I_1} \{\ell \in m(x_t)\} \right) = 1 \right\}. \end{aligned}$$

Similar calculations yield

$$\mathcal{H}(x_t^i \mid x_t^{-(i,j)}) = \mathcal{H}_t^1 \cdot \mathbb{P}(\mathcal{E}_{t,3}^{i,j}) + \mathcal{H}_t^2 \cdot \left(1 - \mathbb{P}(\mathcal{E}_{t,3}^{i,j})\right);$$

$$\mathcal{H}(x_t^i \mid x_t^{-i}) = \mathcal{H}_t^1 \cdot \mathbb{P}(\mathcal{E}_{t,3}^i) + \mathcal{H}_t^2 \cdot \big(1 - \mathbb{P}(\mathcal{E}_{t,3}^i)\big).$$

Therefore, we obtain

$$\begin{aligned}\mathrm{I}(x_t^i; x_t^j \mid x_t^{-(i,j)}) &= (\mathcal{H}_t^2 - \mathcal{H}_t^1)\big(\mathbb{P}(\mathcal{E}_{t,3}^i) - \mathbb{P}(\mathcal{E}_{t,3}^{i,j})\big) \\ &= \log(2) e^{-|I_1|t}\left(1 - e^{-t}\right)^{|I_0|} = O\left(e^{-\mathcal{H}_{\mathrm{B}}(r)d}\right),\end{aligned}$$

where the last equality is due to the fact that $e^{-|I_1|t}\left(1 - e^{-t}\right)^{|I_0|}$ is maximized at $t = -\log(1-r)$. Finally, with $|\{i \in I_0, j \in I_1\}| = r(1-r)d^2$, we can bound

$$\mathcal{I}_3(t) = \sum_{i \in I_0, j \in I_1} I(x_t^i; x_t^j \mid x_t^{-(i,j)}) = O\left(d^2 e^{-\mathcal{H}_{\mathrm{B}}(r)d}\right). \tag{37}$$

Case 4: $i \in I_1$**,** $j \in I_0$**.** Notice that $\mathcal{I}_3(t)$ and $\mathcal{I}_4(t)$ are invariant under swapping i and j. We can show in the same way as above that

$$\mathcal{I}_4(t) = O\left(d^2 e^{-\mathcal{H}_{\mathrm{B}}(r)d}\right). \tag{38}$$

Putting everything together. Combining Eqns. (33), (36), (37) and (38), we arrive at

$$\mathcal{I}(t) \lesssim d^2\left(e^{-2t}(1 - e^{-t})^{rd/2} + e^{-(1-r)dt} + e^{-\mathcal{H}_{\mathrm{B}}(r)d}\right). \tag{39}$$

We are now in a position to prove Eqn. (28). Let us begin with the integration over the time interval $t \in [1/d, 1]$. Direct calculation yields that $e^{-2t}(1 - e^{-t})^{rd/2}$ is maximized at $t^* = \log(1 + \frac{rd}{4}) > 1$, which reveals that

$$d^2 e^{-2t}(1 - e^{-t})^{rd/2} \leqslant d^2 e^{-2t^*} = d^2\left(1 + \frac{rd}{4}\right)^{-2} = O(1). \tag{40}$$

For the term from $\mathcal{I}_2(t)$, we obtain

$$\int_{\frac{1}{d}}^1 t \cdot d^2 e^{-(1-r)dt}\mathrm{d}t \overset{\text{(a)}}{=} \int_1^d s e^{-(1-r)s}\mathrm{d}s \leqslant \int_0^\infty s e^{-(1-r)s}\mathrm{d}s = \frac{1}{(1-r)^2} = O(1), \tag{41}$$

where in (a), we use the change of variable formula with $s = dt$. Similarly, we can show that

$$\int_{\frac{1}{d}}^1 t \cdot d^2 e^{-\mathcal{H}_{\mathrm{B}}(r)d}\mathrm{d}t \leqslant \int_0^1 t \cdot d^2 e^{-\mathcal{H}_{\mathrm{B}}(r)d}\mathrm{d}t = \frac{1}{2} d^2 e^{-\mathcal{H}_{\mathrm{B}}(r)d} = O(1), \tag{42}$$

where the condition $\min\{r, 1-r\} = \Theta(1)$ ensures $\mathcal{H}_{\mathrm{B}}(r) = \Theta(1)$. Taking collectively Eqns. (39), (40), (41) and (42), we arrive at

$$\int_{\frac{1}{d}}^1 \min\{1, t\}\, \mathcal{I}(t)\mathrm{d}t = \int_{\frac{1}{d}}^1 t\, \mathcal{I}(t)\mathrm{d}t = O(1). \tag{43}$$

Let us move on to the integration over time interval $t \in [1, \log(d)]$. The integral computation yields that

$$\int_1^{\log(d)} d^2 e^{-2t}(1 - e^{-t})^{rd/2}\mathrm{d}t \overset{(a)}{=} \int_1^{d/e} s\left(1 - \frac{s}{d}\right)^{rd/2}\mathrm{d}s \overset{(b)}{\leqslant} \int_0^\infty s e^{-rs/2}\mathrm{d}s = O(1), \tag{44}$$

where in (a), we use the change of variable formula with $s = de^{-t}$, and in (b), we use the inequality $(1-x)^{1/x} \leqslant e^{-1}$ for $x \in (0,1]$. For the remaining terms, we have

$$\int_1^{\log(d)} d^2 e^{-(1-r)dt} \mathrm{d}t \leqslant \int_0^{\log(d)} d^2 e^{-(1-r)d} \mathrm{d}t = d^2 \log(d) e^{-(1-r)d} = O(1); \tag{45}$$

$$\int_1^{\log(d)} d^2 e^{-\mathcal{H}_{\mathrm{B}}(r)d} \mathrm{d}t \leqslant d^2 \log(d) e^{-\mathcal{H}_{\mathrm{B}}(r)d} = O(1), \tag{46}$$

where the condition $\min\{r, 1-r\} = \Theta(1)$ ensures $\mathcal{H}_{\mathrm{B}}(r) = \Theta(1)$. Now, combining Eqns. (39), (44), (45) and (46) yields

$$\int_1^{\log(d)} \min\{1,t\}\, \mathcal{I}(t) \mathrm{d}t = \int_1^{\log(d)} \mathcal{I}(t) \mathrm{d}t = O(1). \tag{47}$$

Finally, equipped with Eqns. (43) and (47), we conclude

$$\int_{\frac{1}{d}}^{\log(d)} \min\{1,t\}\, \mathcal{I}(t) \mathrm{d}t = \int_{\frac{1}{d}}^{1} \min\{1,t\}\, \mathcal{I}(t) \mathrm{d}t + \int_1^{\log(d)} \min\{1,t\}\, \mathcal{I}(t) \mathrm{d}t = O(1),$$

which proves $\mathcal{D}(p_{\mathrm{ex}}) = O(1)$.

Appendix B. Proof sketches

B.1. Proof sketch of Theorem 1

By the data-processing inequality and the chain rule for KL divergence, we upper bound the KL divergence between q_0 and p_T by the KL divergence between the paths $q_{T-t_0,\ldots,T-t_N}$ and $p_{t_0,\ldots,t_N}$, which can be decomposed as

$$\begin{aligned}
\mathsf{KL}(q_0 \,\|\, p_T) &\leqslant \mathsf{KL}(q_{T-t_0,\ldots,T-t_N} \,\|\, p_{t_0,\ldots,t_N}) \\
&\leqslant \mathsf{KL}(q_T \,\|\, p_0) + \sum_{k=0}^{N-1} \mathbb{E}_{x_{t_k} \sim \overleftarrow{q}_{t_k}} \left[\mathsf{KL}\left(\overleftarrow{q}_{t_{k+1}|t_k}(\cdot|x_{t_k}) \,\|\, p_{t_{k+1}|t_k}(\cdot|x_{t_k}) \right) \right].
\end{aligned}$$

The first term is the initialization error, which can be upper bounded by the log-Sobolev inequality in Lemma 7. For the second term, we apply Girsanov's change-of-measure theorem for continuous-time Markov chains to obtain the following upper bound:

$$\frac{1}{S} \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k,t}} \sum_{i \in [d]} \sum_{c \in [S]} s_{T-t}(x_t \oplus_i c, x_t) D\big(\widehat{s}_{T-t_k}(x_{t_k} \oplus_i c, x_{t_k}), s_{T-t}(x_t \oplus_i c, x_t)\big) \mathrm{d}t.$$

The details can be found around Eq. (59).

To further control the right hand side, applying the law of cosines for the Bregman divergence, we obtain (with $\ell := t_k$)

$$\sum_{i \in [d]} \sum_{c \in [S]} s_{T-t}(x_t \oplus_i c, x_t) D\big(\widehat{s}_{T-t_k}(x_{t_k} \oplus_i c, x_{t_k}), s_{T-t}(x_t \oplus_i c, x_t)\big)$$

$$= \underbrace{\sum_{y_\ell : \mathrm{d_H}(y_\ell, x_\ell)=1} s_{T-\ell}(y_\ell, x_\ell) D\big(\widehat{s}_{T-\ell}(y_\ell, x_\ell), s_{T-\ell}(y_\ell, x_\ell)\big)}_{\text{Controlled by Assumption 1}}$$
$$+ \underbrace{\sum_{i\in[d]} \sum_{c\in[S]} \big(s_{T-\ell}(x_\ell \oplus_i c, x_\ell) - s_{T-t}(x_t \oplus_i c, x_t)\big) \log \widehat{s}_{T-\ell}(x_\ell \oplus_i c, x_\ell)}_{\text{Expectation controlled by Lemma 8}}$$
$$+ \sum_{y_t : \mathrm{d_H}(y_t, x_t)=1} \big(-\log s_t(y_t, x_t)\big) - \sum_{y_\ell : \mathrm{d_H}(y_\ell, x_\ell)=1} \big(-\log s_{T-\ell}(y_\ell, x_\ell)\big).$$

The first term can be controlled by Assumption 1 after taking expectation over $x_\ell \sim \overleftarrow{q}_\ell$ and integration over time. The second term can be proven to be zero with the help of Lemma 8 after taking expectation over $x_t \sim \overleftarrow{q}_{t|\ell}(\cdot \mid x_\ell)$. Thus, the problem boils down to upper bounding the third term above, whose properties are characterized in Lemma 10. After taking expectation and integration over time, we can upper bound the third term by $\Delta d \log(S/\Delta)$. Combining the bounds for all three terms completes the proof.

B.2. Proof sketch of Theorem 2

The proof is based on a refined analysis of the relative entropy decay along the forward processes for any distribution $q_0 \in \mathcal{P}^\gamma(\mathcal{X})$ and it works for every $\gamma \in (0, 1)$. It can be shown that the relative entropy along the forward process is a differentiable function over time t, and we denote the negative changing rate of it as $\varphi(t)$, i.e.,

$$\varphi(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(q_t \| p_0) = \sum_{x,y:\mathrm{d_H}(x,y)=1} q_t(x) s_t(y, x) \log\left(\frac{s_t(y, x)}{s_t(x, y)}\right),$$

where $p_0 = \mathrm{Unif}(\mathcal{X})$ is the limit distribution of the forward noising process. First, we show that the condition $\mathsf{KL}(Q\|P) \leqslant \varepsilon_{\text{score}} + O(1)$ with the definition of $\mathcal{P}^\gamma(\mathcal{X})$ implies that $T > 1$ and the condition

$$\sum_{k=1}^{N-1} \int_{t_k}^{t_{k+1}} \big(\varphi(T-t) - \varphi(T-t_k)\big)\mathrm{d}t = O(1). \tag{48}$$

Further we can show that $\varphi(t)$ is a non-increasing function and differentiable over t. Thus, Eqn. Eqn. (48) and the Newton-Leibniz formula leads to a stronger condition

$$\begin{aligned}\sum_{k=1}^{N-1} \inf_{t_k \leqslant t \leqslant t_{k+1}} (-\varphi'(T-t)) \cdot \frac{1}{2}(t_{k+1} - t_k)^2 &\leqslant \sum_{k=1}^{N-1} \int_{t_k}^{t_{k+1}} \int_{T-t}^{T-t_k} -\varphi'(u)\mathrm{d}u\mathrm{d}t \\ &= \sum_{k=1}^{N-1} \int_{t_k}^{t_{k+1}} \big(\varphi(T-t) - \varphi(T-t_k)\big) = O(1).\end{aligned} \tag{49}$$

By viewing the forward process as a S-ary symmetric channel (Makur and Polyanskiy, 2018), the strong data process inequality can be applied to get Lemma 12, which further shows that for any

$q_0 \in \mathcal{P}^\gamma(\mathcal{X})$, the function $-\varphi'(t)$ has a lower bound scaling as $\gamma d \log(S)$ for all $t \in (0,1)$. Since $\max_k\{t_{k+1} - t_k\} \leqslant \frac{1}{2}$, we can choose a suitable M, such that $1 < M < N$ and $T - t_M = [\frac{1}{2}, 1]$. Combining this with Eqn. (49), we obtain

$$\sum_{k=M}^{N-1} (t_{k+1} - t_k)^2 \lesssim \frac{1}{\gamma d \log(S)},$$

which implies that $N = \Omega(\gamma d \log(S))$ by Cauchy-Schwarz inequality.

B.3. Proof sketch of Theorem 3

First, Lemma 13 shows that Algorithm 1 outputs a sample from a CTMC with initial distribution p_0 and rate matrices

$$\widehat{Q}_t(x,y) := \begin{cases} \widehat{s}_{T-t_k}(x_{t_k} \odot_i y^i, x_{t_k}) \frac{e^{T-t_k}-1}{e^{T-t}-1} \mathbb{I}\{x^i = \text{MASK}\}, & \text{if } \mathrm{d_H}(x,y) = 1, x^i \neq y^i, \text{ and } x^i_{t_k} = \text{MASK}, \\ -\sum_{z \neq x} \widehat{Q}_t(x,z), & \text{if } y = x, \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

This corresponds to a τ-bridging strategy (Eqn. (7)) with the following function $G_t^i(\widehat{s}_{T-t_k}, x_{t_k})$:

$$G_t^i(\widehat{s}_{T-t_k}, x_{t_k})(a,b) = \frac{e^{T-t_k}-1}{e^{T-t}-1} \widehat{Q}_{T-t_k}(x_{t_k}, x_{t_k} \odot_i b) \mathbb{I}\{x^i_{t_k} = a\} \quad \text{for } a \neq b \in \mathcal{V}.$$

By the data-processing inequality, we upper bound the KL divergence between q_0 and p_T by the KL divergence between the paths $q_{T-t_0,\ldots,T-t_N}$ and $p_{t_0,\ldots,t_N}$. Next, we apply the Markovian property of the paths along with the Girsanov's change-of-measure theorem and upper bound $\mathsf{KL}(q_0 \| p_T)$ by

$$\mathsf{KL}(q_T \| p_0) + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k,t}} \left[\sum_{y_t : Q(y_t, x_t) > 0} s_{T-t}(y_t, x_t) \right.$$
$$\left. D\left(\frac{e^{T-t_k}-1}{e^{T-t}-1} \widehat{s}_{T-t_k}(y_t, x_t), s_{T-t}(y_t, x_t) \right) \mathrm{d}t \right].$$

The first term is the initialization error and is controlled by choosing $T = \Omega(\log d + \log\log(\varepsilon^{-1} S))$. For the second term, we apply the law of cosines for the Bregman divergence and get (with $\ell := t_k$ and $y_\ell := x_\ell \odot_i c$, where $y_t = x_t \odot_i c$):

$$s_{T-t}(y_t, x_t) D\left(\frac{e^{T-\ell}-1}{e^{T-t}-1} \widehat{s}_{T-\ell}(y_t, x_t), s_{T-t}(y_t, x_t) \right)$$
$$= \underbrace{\frac{e^{T-\ell}-1}{e^{T-t}-1} s_{T-\ell}(y_\ell, x_\ell) D\left(\widehat{s}_{T-\ell}(y_\ell, x_t), s_{T-\ell}(y_t, x_t) \right)}_{\text{controlled by Assumption 1}}$$
$$+ \underbrace{\left(s_{T-t}(y_\ell, x_\ell) - s_{T-t}(y_t, x_t) \right) \log \frac{\widehat{s}_{T-\ell}(y_\ell, x_\ell)}{s_{T-\ell}(y_\ell, x_\ell)}}_{\text{after taking } \mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)} \text{ equals zero by Lemma 14}}$$
$$+ s_{T-t}(y_t, x_t) D\left(s_{T-t}(y_\ell, x_\ell), s_{T-t}(y_t, x_t) \right).$$

Similar to the proof of uniform discrete diffusion model, the first term can be controlled by Assumption 1 after taking expectation over $x_{t_k} \sim \overleftarrow{q}_{t_k}$ and integration over time, and the second term can be proven to be zero by the martingale property from Lemma 14. Finally, using Dynkin's formula we relate the third term to the effective total correlation $\mathcal{D}(q_0)$.

Appendix C. Technical preparations

C.1. Score functions

Below, we present an equivalent formulation of the score functions.

Proposition 6 *Let q_0 be an initial distribution on $\mathcal{X}_0$. Let $x, y \in \mathcal{X}$ be such that $Q(y, x) > 0$. Then,*

1. *for the uniform noising process,*

$$s_t(y, x) = \frac{\mathbb{E}_{x_0 \sim q_0} \alpha_t^{\mathrm{d_H}(y, x_0)}}{\mathbb{E}_{x_0 \sim q_0} \alpha_t^{\mathrm{d_H}(x, x_0)}}, \tag{51}$$

   where $\alpha_t := \frac{1-e^{-t}}{1+(S-1)e^{-t}}$.

2. *for the masking noising process,*

$$s_t(y, x) = \frac{1}{e^t - 1} \frac{q_0(y)}{q_0(x)}, \tag{52}$$

   where for $x \in \mathcal{X} \backslash \mathcal{X}_0$, $q_0(x)$ is the marginal probability of the unmasked *coordinates of x under q_0.*

Proof of Proposition 6. By the definition of the score function, one can write

$$s_t(y, x) = \frac{q_t(y)}{q_t(x)} = \frac{\sum_{x_0} q_{t|0}(y \mid x_0) q_0(x_0)}{\sum_{x_0} q_{t|0}(x \mid x_0) q_0(x_0)}.$$

For the uniform noising process, one can solve the Kolmogorov forward equation for every dimension. As a result, the transition can be written as

$$q_{t|0}(y \mid x_0) = \left(\frac{1 - e^{-t}}{S}\right)^{\mathrm{d_H}(y, x_0)} \left(\frac{1 + (S-1)e^{-t}}{S}\right)^{d - \mathrm{d_H}(y, x_0)} = \left(\frac{1 + (S-1)e^{-t}}{S}\right)^{d} \alpha_t^{\mathrm{d_H}(y, x_0)},$$

which proves Eqn. (51). More details of this relation can be found in (e.g., Zhang et al. (2025), Proposition 1).

For the masking noising process, for notational convenience, given any $x \in ([S] \cup \{\mathrm{MASK}\})^d$, define

$$m(x) := \{i \in [d] : x^i = \mathrm{MASK}\}. \tag{53}$$

In view of this piece of notation, as $\Pr(x_t^i = \mathrm{MASK}) = e^{-t}$, and coordinates evolve independently, one can write

$$q_{t|0}(y \mid x_0) = (1 - e^{-t})^{|m(y)|} e^{-t(d - |m(y)|)} \mathbb{I}\{\text{ for all } i \in [d], y^i \in \{x_0^i, \mathrm{MASK}\}\}.$$

As $Q(y,x) > 0$, it must be that $\mathrm{d_H}(x,y) = 1$, and for i, such that $x^i \neq y^i$, $x^i = \mathrm{MASK}$ and $y^i \neq \mathrm{MASK}$. This implies that $|m(x)| = |m(y)| + 1$, and we can write

$$\frac{\sum_{x_0} q_{t|0}(y \mid x_0) q_0(x_0)}{\sum_{x_0} q_{t|0}(x \mid x_0) q_0(x_0)} = \frac{e^{-t}}{1-e^{-t}} \frac{\sum_{x_0} q_0(x_0) \mathbb{I}\{ \text{ for all } i \in [d], y^i \in \{x_0^i, \mathrm{MASK}\} \}}{\sum_{x_0} q_0(x_0) \mathbb{I}\{ \text{ for all } i \in [d], x^i \in \{x_0^i, \mathrm{MASK}\} \}} = \frac{1}{e^t - 1} \frac{q_0(y)}{q_0(x)}.$$

C.2. Technical lemmas

Lemma 4 (Chain rule of KL divergence) *For $N > 0$, let $a_{0:N}, b_{0:N}$ be the joint distributions of two Markov processes. Then,*

$$\mathsf{KL}(a_{0:N} \| b_{0:N}) = \mathsf{KL}(a_0 \| b_0) + \sum_{k=0}^{N-1} \mathbb{E}_{x \sim a_k} \mathsf{KL}\left(a_{k+1|k}(\cdot \mid x) \| b_{k+1|k}(\cdot \mid x) \right).$$

Proof of Lemma 4. Invoking the definition of KL divergence with some direct calculations yields

$$\begin{aligned}
\mathsf{KL}(a_{0:N} \| b_{0:N}) &= \mathbb{E}_{x_{0:N} \sim a_{0:N}} \log \frac{a_{0:N}(x_{0:N})}{b_{0:N}(x_{0:N})} \\
&= \mathbb{E}_{x_{0:N} \sim a_{0:N}} \log \left(\frac{a_0(x_0)}{b_0(x_0)} \prod_{k=0}^{N-1} \frac{a_{k+1|k}(x_{k+1} \mid x_k)}{b_{k+1|k}(x_{k+1} \mid x_k)} \right) \\
&= \mathbb{E}_{x_0 \sim a_0} \log \frac{a_0(x_0)}{b_0(x_0)} + \sum_{k=0}^{N-1} \mathbb{E}_{x_k \sim a_k} \mathbb{E}_{x_{k+1} \sim a_{k+1|k}(\cdot|x_k)} \log \frac{a_{k+1|k}(x_{k+1} \mid x_k)}{b_{k+1|k}(x_{k+1} \mid x_k)} \\
&= \mathsf{KL}(a_0 \| b_0) + \sum_{k=0}^{N-1} \mathbb{E}_{x_k \sim a_k} \mathsf{KL}\left(a_{k+1|k}(\cdot \mid x_k) \| b_{k+1|k}(\cdot \mid x_k) \right).
\end{aligned}$$

Lemma 5 (Derivative of KL: an upper bound) *Let (q_t) and (p_t) be the marginals of CTMCs with rate matrices (Q_t) and $(\widehat{Q}_t)$, respectively, and $\overleftarrow{q}_t \equiv q_{T-t}$ be the marginals of the reverse process. Then, for any $t > t_k$ and for any z,*

$$\begin{aligned}
&\frac{\partial}{\partial t} \mathsf{KL}\left(\overleftarrow{q}_{t|t_k}(\cdot \mid z) \| p_{t|t_k}(\cdot \mid z) \right) \\
&\leqslant \mathbb{E}_{x_t \sim \overleftarrow{q}_{t|t_k}(\cdot|z)} \sum_{y \neq x_t} \left[\widehat{Q}_t(x_t, y) - \overleftarrow{Q}_t(x_t, y) + \overleftarrow{Q}_t(x_t, y) \log \frac{\overleftarrow{Q}_t(x_t, y)}{\widehat{Q}_t(x_t, y)} \right].
\end{aligned}$$

Proof of Lemma 5. Let us omit the conditioning on z for the notation brevity. By direct calculations, one can write

$$A := \frac{\partial}{\partial t} \mathsf{KL}\left(\overleftarrow{q}_{t|t_k} \| p_{t|t_k} \right) = \sum_{x \in \mathcal{X}} \left(\frac{\partial}{\partial t} \overleftarrow{q}_{t|t_k}(x) \right) \log \frac{\overleftarrow{q}_{t|t_k}(x)}{p_{t|t_k}(x)} - \sum_{x \in \mathcal{X}} \overleftarrow{q}_{t|t_k}(x) \frac{\frac{\partial}{\partial t} p_{t|t_k}(x)}{p_{t|t_k}(x)}.$$

Recall the Kolmogorov equation:

$$\frac{\partial}{\partial t} \overleftarrow{q}_{t|t_k}(x) = \sum_{y \in \mathcal{X}} \overleftarrow{Q}_t(y, x) \overleftarrow{q}_{t|t_k}(y) \quad \text{and} \quad \frac{\partial}{\partial t} p_{t|t_k}(x) = \sum_{y \in \mathcal{X}} \widehat{Q}_t(y, x) p_{t|t_k}(y)$$

Putting the above together, we obtain (relabeling x and y)

$$\begin{aligned}
A &= \mathbb{E}_{x\sim \overleftarrow{q}_{t|t_k}} \sum_{y\in\mathcal{X}} \left[\overleftarrow{Q}_t(x,y) \log\left(\frac{\overleftarrow{q}_{t|t_k}(y)}{p_{t|t_k}(y)}\right) - \widehat{Q}_t(x,y) \frac{\overleftarrow{q}_{t|t_k}(y)}{\overleftarrow{q}_{t|t_k}(x)} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)} \right] \\
&= \mathbb{E}_{x\sim \overleftarrow{q}_{t|t_k}} \sum_{y\neq x} \left[\overleftarrow{Q}_t(x,y) \log\left(\frac{\overleftarrow{q}_{t|t_k}(y)}{p_{t|t_k}(y)}\right) - \overleftarrow{Q}_t(x,y) \log\left(\frac{\overleftarrow{q}_{t|t_k}(x)}{p_{t|t_k}(x)}\right) \right. \\
&\qquad\qquad \left. - \widehat{Q}_t(x,y) \frac{\overleftarrow{q}_{t|t_k}(y)}{\overleftarrow{q}_{t|t_k}(x)} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)} \right] - \widehat{Q}_t(x,x) \\
&= \mathbb{E}_{x\sim \overleftarrow{q}_{t|t_k}} \sum_{y\neq x} \left[\overleftarrow{Q}_t(x,y) \log\left(\frac{\overleftarrow{q}_{t|t_k}(y)}{\overleftarrow{q}_{t|t_k}(x)} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)}\right) - \widehat{Q}_t(x,y) \frac{\overleftarrow{q}_{t|t_k}(y)}{\overleftarrow{q}_{t|t_k}(x)} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)} \right] - \widehat{Q}_t(x,x) \\
&= \mathbb{E}_{x\sim \overleftarrow{q}_{t|t_k}} \sum_{y\neq x} \left[\overleftarrow{Q}_t(x,y) \log\left(\frac{\overleftarrow{q}_{t|t_k}(y)}{\overleftarrow{q}_{t|t_k}(x)} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)}\right) - \widehat{Q}_t(x,y) \frac{\overleftarrow{q}_{t|t_k}(y)}{\overleftarrow{q}_{t|t_k}(x)} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)} + \widehat{Q}_t(x,y) \right],
\end{aligned} \tag{54}$$

where we invoke the the property that $\overleftarrow{Q}_t(x,x) = -\sum_{y\neq x} \overleftarrow{Q}_t(x,y)$ and $\widehat{Q}_t(x,x) = -\sum_{y\neq x} \widehat{Q}_t(x,y)$. Then, letting C_{xy} be such that (recall that z is fixed)

$$\frac{\overleftarrow{q}_{t|t_k}(y \mid z)}{\overleftarrow{q}_{t|t_k}(x \mid z)} = \frac{\overleftarrow{Q}_t(x,y)}{C_{xy}}, \tag{55}$$

it satisfies that

$$\begin{aligned}
A = \mathbb{E}_{x\sim \overleftarrow{q}_{t|t_k}} \sum_{y\neq x} \left[\widehat{Q}(x,y) + \overleftarrow{Q}(x,y) \log\left(\frac{\overleftarrow{Q}_t(x,y)}{\widehat{Q}_t(x,y)}\right) + \overleftarrow{Q}_t(x,y) \log\left(\frac{\widehat{Q}_t(x,y)}{C_{xy}} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)}\right) \right. \\
\left. - \widehat{Q}(x,y) \frac{\overleftarrow{Q}_t(x,y)}{C_{xy}} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)} \right]
\end{aligned} \tag{56}$$

Finally, since $\log z \leqslant z - 1$,

$$\begin{aligned}
A &\leqslant \mathbb{E}_{x\sim \overleftarrow{q}_{t|t_k}} \sum_{y\neq x} \left[\widehat{Q}(x,y) + \overleftarrow{Q}(x,y) \log\left(\frac{\overleftarrow{Q}_t(x,y)}{\widehat{Q}_t(x,y)}\right) - \overleftarrow{Q}(x,y) \right. \\
&\qquad \left. + \overleftarrow{Q}_t(x,y) \frac{\widehat{Q}_t(x,y)}{C_{xy}} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)} - \widehat{Q}(x,y) \frac{\overleftarrow{Q}_t(x,y)}{C_{xy}} \cdot \frac{p_{t|t_k}(x)}{p_{t|t_k}(y)} \right] \\
&= \mathbb{E}_{x\sim \overleftarrow{q}_{t|t_k}} \sum_{y\neq x} \left[\widehat{Q}_t(x,y) - \overleftarrow{Q}_t(x,y) + \overleftarrow{Q}_t(x,y) \log\left(\frac{\overleftarrow{Q}_t(x,y)}{\widehat{Q}_t(x,y)}\right) \right].
\end{aligned}$$

Lemma 6 (Itô's Lemma for Poisson jump process) *For the Poisson jump process $\{x_t\}_{t\geqslant 0}$ with generator $\{L_t\}_{t\geqslant 0}$ and rate matrix $\{R_t\}_{t\geqslant 0}$. Itô's Lemma formula can be written as*

$$f(t,x_t) = f(0,x_0) + \int_0^t \left[\partial_s f(s,x_{s^-}) + (L_s f)(s,x_{s^-})\right] \mathrm{d}t + M_t, \tag{57}$$

where $x_{s-} = \lim_{u \to s^-} x_s$*, which exists for almost everywhere* $s \in [0, t)$ *under the Lebesgue measure. The compensation process* $\{M_u\}_{u \in [\ell, t]}$ *is defined as*

$$M_u = \sum_{y_s : y_s \neq x_s} \int_\ell^u \big(f(s, y_s) - f(s, x_s)\big)\big(\mathrm{d}N_s^{x_s, y_s} - \lambda_s^{x_s, y_s}\mathrm{d}s\big),$$

where $N_s^{x,y}$ *is the counting process of jumps from* x *to* y *up to time* t *and* $\lambda_s^{x,y}$ *is the* $\mathcal{F}_{s^-}$*-intensity of* $N_s^{x,y}$*, i.e.,* $\lambda_s^{x,y} = \mathbb{I}\{x_{s^-} = y\} R_t(x, y)$.

Appendix D. Proofs of results in Section 3.1

D.1. Proof of Theorem 1

We first decompose the KL divergence between the output distribution p_T and the target distribution q_0 as

$$\begin{aligned}
\mathsf{KL}(q_0 \| p_T) &\leqslant \mathsf{KL}(q_{T-t_0, \dots, T-t_N} \| p_{t_0, \dots, t_N}) \\
&= \mathsf{KL}(q_T \| p_0) + \sum_{k=0}^{N-1} \mathbb{E}_{x_{t_k} \sim \overleftarrow{q}_{t_k}} \left[\mathsf{KL}\left(\overleftarrow{q}_{t_{k+1}|t_k}(\cdot | x_{t_k}) \| p_{t_{k+1}|t_k}(\cdot | x_{t_k}) \right) \right], \qquad (58)
\end{aligned}$$

where the first inequality follows from the data-processing inequality for KL divergence and the second inequality follows from the chain rule for KL divergence in Lemma 4. The first term is the initialization error, which can be upper bounded by the following lemma.

Lemma 7 *For the uniform noising process, for any initial distribution* $q_0 \in \mathcal{P}(\mathcal{X})$*, time index* $t \geqslant 0$*, we have the same limit distribution*

$$q_t \xrightarrow{d} p_0 = \mathrm{Unif}(\mathcal{X}), \quad \text{as } t \to \infty.$$

Further, the modified log-Sobolev constant [3] *of* q_t *satisfies* $C_{\mathrm{LSI}} = 2$*, which leads to*

$$\mathsf{KL}(q_t \| p_0) \leqslant e^{-t} \mathsf{KL}(q_0 \| p_0) \leqslant e^{-t} d \log(S).$$

The proof of Lemma 7 can be found in previous works, e.g., Zhang et al. (2025, Proposition 2). Applying the above Lemma together with Lemma 5 on the second term in Eqn. (58), we obtain

$$\begin{aligned}
&\mathsf{KL}(p_0 \| q_T) \\
&\leqslant \mathsf{KL}(q_T \| p_0) + \sum_{k=0}^{N-1} \mathbb{E}_{x_{t_k} \sim \overleftarrow{q}_{t_k}} \left[\int_{t_k}^{t_{k+1}} \frac{\partial}{\partial t} \mathsf{KL}\left(\overleftarrow{q}_{t|t_k}(\cdot | x_{t_k}) \| p_{t|t_k}(\cdot | x_{t_k}) \right) \mathrm{d}t \right] \\
&\leqslant e^{-T} d \log(S) \\
&\quad + \sum_{k=0}^{N-1} \mathbb{E}_{x_{t_k} \sim \overleftarrow{q}_{t_k}} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_t \sim \overleftarrow{q}_{t|t_k}} \sum_{y \neq x_t} \left[\widehat{Q}_t(x_t, y) - \overleftarrow{Q}_t(x_t, y) + \overleftarrow{Q}_t(x_t, y) \log \frac{\overleftarrow{Q}_t(x_t, y)}{\widehat{Q}_t(x_t, y)} \right] \mathrm{d}t
\end{aligned}$$

3. C_{LSI} is defined as the smallest number such that for any $q \in \mathcal{P}(\mathcal{X})$, $\mathsf{KL}(q \mid \mathrm{Unif}(\mathcal{X})) \leqslant C_{\mathrm{LSI}}/2 \cdot \mathcal{E}(q, \log(q))$, where $\mathcal{E}$ is the Dirichlet form associated with the uniform noising process, i.e., $\mathcal{E}(f, g) = -(2|\mathcal{X}|)^{-1} \sum_{x, y \in \mathcal{X}} (f(x) - f(y))(g(x) - g(y)) Q(x, y)$.

$$\leqslant e^{-T} d \log(S) + \frac{1}{S} \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k,t}} \Big[\sum_{i \in [d]} \sum_{c \in [S]} s_{T-t}(x_t \oplus_i c, x_t) \\ D\big(\widehat{s}_{T-t_k}(x_{t_k} \oplus_i c, x_{t_k}), s_{T-t}(x_t \oplus_i c, x_t)\big) \mathrm{d}t \Big]. \tag{59}$$

In the following, we focus on the quantity

$$\mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k,t}} \sum_{i \in [d]} \sum_{c \in [S]} s_{T-t}(x_t \oplus_i c, x_t) D\big(\widehat{s}_{T-t_k}(x_{t_k} \oplus_i c, x_{t_k}), s_{T-t}(x_t \oplus_i c, x_t)\big). \tag{60}$$

For simplicity, we write $t_k \coloneqq \ell$. Direct calculations yield the following decomposition

$$\begin{aligned} &\sum_{i \in [d]} \sum_{c \in [S]} s_{T-t}(x_t \oplus_i c, x_t) D\big(\widehat{s}_{T-t_k}(x_{t_k} \oplus_i c, x_{t_k}), s_{T-t}(x_t \oplus_i c, x_t)\big) \\ &= \underbrace{\sum_{y_\ell : \mathrm{d_H}(y_\ell, x_\ell) = 1} s_{T-\ell}(y_\ell, x_\ell) D\left(\widehat{s}_{T-\ell}(y_\ell, x_\ell), s_{T-\ell}(y_\ell, x_\ell)\right)}_{T_1^{t,\ell}} \\ &\quad + \underbrace{\sum_{i \in [d]} \sum_{c \in [S]} \left(s_{T-\ell}(x_\ell \oplus_i c, x_\ell) - s_{T-t}(x_t \oplus_i c, x_t)\right) \log \widehat{s}_{T-\ell}(x_\ell \oplus_i c, x_\ell)}_{T_2^{t,\ell}} \\ &\quad + \underbrace{\sum_{y_t : \mathrm{d_H}(y_t, x_t) = 1} h\left(s_{T-t}(y_t, x_t)\right) - \sum_{y_\ell : \mathrm{d_H}(y_\ell, x_\ell) = 1} h\left(s_{T-\ell}(y_\ell, x_\ell)\right)}_{T_3^{t,\ell}}, \end{aligned}$$

where $h(x) = x \log x - x + 1$. We proceed by bounding each term separately.

- For term $T_1^{t,\ell}$, notice that $T_1^{t,\ell}$ is independent of t. In view of definition of score entropy loss, we have

$$\begin{aligned} &\mathbb{E}_{x_\ell, x_t \sim \overleftarrow{q}_{\ell,t}} \left[T_1^{t,\ell}\right] \qquad (61) \\ &= \mathbb{E}_{x_\ell, x_t \sim \overleftarrow{q}_{\ell,t}} \left[\sum_{y_\ell : \mathrm{d_H}(y_\ell, x_\ell) = 1} s_{T-\ell}(y_\ell, x_\ell) D\left(\widehat{s}_{T-\ell}(y_\ell, x_\ell), s_{T-\ell}(y_\ell, x_\ell)\right) \right] \\ &= S \cdot \mathbb{E}_{x_\ell, x_t \sim \overleftarrow{q}_{\ell,t}} \left[\sum_{y_\ell : \mathrm{d_H}(y_\ell, x_\ell) = 1} Q_{T-\ell}(y_\ell, x_\ell) s_{T-\ell}(y_\ell, x_\ell) D\left(\widehat{s}_{T-\ell}(y_\ell, x_\ell), s_{T-\ell}(y_\ell, x_\ell)\right) \right] \\ &= S \cdot \mathcal{L}_{\mathrm{SE}}(T - \ell, \widehat{s}_{T-\ell}, s_{T-\ell}), \qquad (62) \end{aligned}$$

  where we use the fact that $Q_{T-t}(y, x) = S^{-1}$ for any $\mathrm{d_H}(y, x) = 1$.

- For term $T_2^{t,\ell}$, we establish the following lemma, whose proof is provided in Section F.2.

Lemma 8 *Consider the uniform noising process and let $0 \leqslant \ell < t < T$. Then, for any $c \in [S]$, $i \in [d]$ and $x_\ell \in \mathcal{X}$, it obeys*

$$\mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}\Big[\big(s_{T-\ell}(x_\ell \oplus_i c, x_\ell) - s_{T-t}(x_t \oplus_i c, x_t)\big) \log \widehat{s}_{T-\ell}(x_\ell \oplus_i c, x_\ell)\Big] = 0.$$

With Lemma 8, it is easily seen that

$$\begin{aligned}
&\mathbb{E}_{x_\ell, x_t \sim \overleftarrow{q}_{\ell,t}}\left[T_2^{t,\ell}\right] \\
&= \mathbb{E}_{x_\ell, x_t \sim \overleftarrow{q}_{\ell,t}}\left[\sum_{i\in[d]}\sum_{c\in[S]}\big(s_{T-\ell}(x_\ell \oplus_i c, x_\ell) - s_{T-t}(x_t \oplus_i c, x_t)\big) \log \widehat{s}_{T-\ell}(x_\ell \oplus_i c, x_\ell)\right] \\
&= \sum_{i\in[d]}\sum_{c\in[S]} \mathbb{E}_{x_\ell \sim \overleftarrow{q}_\ell}\Big[\mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}\Big[\big(s_{T-\ell}(x_\ell \oplus_i c, x_\ell) - s_{T-t}(x_t \oplus_i c, x_t)\big) \\
&\qquad\qquad \log \widehat{s}_{T-\ell}(x_\ell \oplus_i c, x_\ell)\Big]\Big] \\
&= 0. \qquad (63)
\end{aligned}$$

- For term $T_3^{t,\ell}$, we make the crucial observation that $\mathbb{E}_{x_t \sim \overleftarrow{q}_t}[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} h(s_{T-t}(y_t, x_t))]$ admits a simple representation. The statement is formalized in the following lemma.

Lemma 9 *For any $t \in [0, T]$, we have*

$$\mathbb{E}_{x_t \sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} h\left(s_{T-t}(y_t, x_t)\right)\right] = \mathbb{E}_{x_t \sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} -\log\left(s_{T-t}(y_t, x_t)\right)\right].$$

In view of this lemma, we can further express the term $T_3^{t,\ell}$ as

$$\begin{aligned}
&\mathbb{E}_{x_\ell, x_t \sim \overleftarrow{q}_{\ell,t}}\left[T_3^{t,\ell}\right] \qquad (64) \\
&= \mathbb{E}_{x_t \sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} h\left(s_{T-t}(y_t, x_t)\right)\right] - \mathbb{E}_{x_\ell \sim \overleftarrow{q}_\ell}\left[\sum_{y_\ell:\mathrm{d_H}(y_\ell,x_\ell)=1} h\left(s_{T-\ell}(y_\ell, x_\ell)\right)\right] \\
&= \mathbb{E}_{x_t \sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} -\log\left(s_{T-t}(y_t, x_t)\right)\right] - \mathbb{E}_{x_\ell \sim \overleftarrow{q}_\ell}\left[\sum_{y_\ell:\mathrm{d_H}(y_\ell,x_\ell)=1} -\log\left(s_{T-\ell}(y_\ell, x_\ell)\right)\right]. \qquad (65)
\end{aligned}$$

Plugging Eqns. (62), (63), and (65) into Eqn. (60), we end up with

$$\begin{aligned}
&\mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k,t}} \sum_{i\in[d]}\sum_{c\in[S]} s_{T-t}(y_t, x_t) D\left(\widehat{s}_{T-t_k}(x_{t_k} \oplus_i c, x_{t_k}), s_{T-t}(y_t, x_t)\right) \\
&= S \cdot \mathcal{L}_{\mathrm{SE}}(T-\ell, \widehat{s}_{T-\ell}, s_{T-\ell}) + \mathbb{E}_{x_t \sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} -\log\left(s_{T-t}(y_t, x_t)\right)\right]
\end{aligned}$$

$$- \mathbb{E}_{x_\ell \sim \overleftarrow{q}_\ell} \left[\sum_{y_\ell : \mathrm{d_H}(y_\ell, x_\ell)=1} - \log\left(s_{T-\ell}(y_\ell, x_\ell)\right) \right]$$
$$= S \cdot \mathcal{L}_{\mathrm{SE}}(T-\ell, \widehat{s}_{T-\ell}, s_{T-\ell}) + S\big(\varphi(T-t) - \varphi(T-\ell)\big),$$

where we define $\varphi(t)$ as

$$\varphi(t) := \frac{1}{S} \mathbb{E}_{x_t \sim q_t} \left[\sum_{y_t : \mathrm{d_H}(y_t, x_t)=1} - \log\left(s_t(y_t, x_t)\right) \right]. \tag{66}$$

Returning to Eqn. (59), we conclude that

$$\mathsf{KL}(p_0 \| q_T) \leqslant e^{-T} d \log(S) + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathcal{L}_{\mathrm{SE}}(T-t_k, \widehat{s}_{T-t_k}, s_{T-t_k}) + \big(\varphi(T-t) - \varphi(T-t_k)\big)\, \mathrm{d}t$$
$$\leqslant \varepsilon_{\mathrm{score}} + e^{-T} d \log(S) + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \big(\varphi(T-t) - \varphi(T-t_k)\big)\, \mathrm{d}t. \tag{67}$$

To establish Theorem 1, it is only left for us to control the last term in Eqn. (67). First, by Jensen's inequality, $\varphi(t)$ is lower bounded by

$$\varphi(t) \geqslant -\frac{1}{S} \sum_{i \in [d]} \sum_{c \in [S]} \log\left(\mathbb{E}_{x_t \sim q_t}\left[s_t(x_t \oplus_i c, x_t)\right]\right) = 0. \tag{68}$$

For the upper bound, it satisfies that

$$\varphi(t) \leqslant \frac{1}{S} \mathbb{E}_{x_t \sim q_t} \left[\big|\{y_t : \mathrm{d_H}(y_t, x_t) = 1\}\big| \cdot \sup_{x,y : \mathrm{d_H}(x,y)=1} \big|\log\left(s_t(y,x)\right)\big| \right], \tag{69}$$

where $|\{y_t : \mathrm{d_H}(y_t, x_t) = 1\}|$ denotes the cardinality of the set $\{y_t : \mathrm{d_H}(y_t, x_t) = 1\}$, which equals $d(S-1)$ for any $x_t \in \mathcal{X}$. It therefore suffices to control the quantity $|\log\left(s_t(y_t, x_t)\right)|$, which is achieved through the following lemma.

Lemma 10 *For any distribution q_0 on $\mathcal{X}$, let q_t be the marginal distribution of the uniform noising process at time t. Then, for any $x, y \in \mathcal{X}$ such that $\mathrm{d_H}(x,y) = 1$, it holds that*

$$|\log s_t(y,x)| \lesssim \log(S) + \max\{\log(t^{-1}), 0\}.$$

As a consequence of Lemma 10, we arrive at

$$\varphi(t) \leqslant \frac{d(S-1)}{S} \cdot \sup_{x,y : \mathrm{d_H}(x,y)=1} |\log\left(s_t(y,x)\right)| \lesssim d\big(\log(S) + \max\{\log(t^{-1}), 0\}\big). \tag{70}$$

In addition, we make the observation in Lemma 12 that $\varphi(t)$ is a non-increasing function in t.

Now we are ready to combine everything and bound the last term of Eqn. (67). Define $\Delta = \max_k\{t_{k+1} - t_k\}$, and choose $1 \leqslant M \leqslant N-1$ such that $T - t_M \in [\Delta, 2\Delta]$. Armed with Eqns. (68), (70) and the monotonicity of $\varphi(t)$, we obtain

$$\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \big(\varphi(T-t) - \varphi(T-t_k)\big)\, \mathrm{d}t$$

$$\leqslant \int_0^{T-t_M} \varphi(t)\,\mathrm{d}t + \sum_{k=0}^{M-1} \int_{t_k}^{t_{k+1}} \big(\varphi(T-t_{k+1}) - \varphi(T-t_k)\big)\,\mathrm{d}t$$
$$\leqslant 2\Delta d\big(\log(S) + \log(2/\Delta) + 1\big) + \Delta \sum_{k=0}^{N-2} \big(\varphi(T-t_{k+1}) - \varphi(T-t_k)\big)$$
$$\leqslant 2\Delta d\big(\log(S) + \log(2/\Delta) + 1\big) + \Delta\varphi(\Delta) \lesssim \Delta d \log(S/\Delta).$$

Combining the inequality above with Eqn. (67) achieves

$$\mathsf{KL}(p_0 \| q_T) \leqslant \varepsilon_{\text{score}} + e^{-T} d \log(S) + \Delta d \log(S/\Delta),$$

which completes the proof of Theorem 1.

D.2. Proof of Corollary 1

Choose time horizon $T = \log(d\log(S)/\varepsilon)$ and number of discretization steps

$$N = \Theta\left(\frac{d\log(S)\log^3(d\log(S)/\varepsilon)}{\varepsilon}\right) = \widetilde{\Theta}\left(\frac{d}{\varepsilon}\right).$$

Adopting the upper bound in Theorem 1 leads to

$$\begin{aligned}\mathsf{KL}(q_{\text{data}} \| p_{\text{output}}) = \mathsf{KL}(q_0 \| p_T) &\lesssim \varepsilon_{\text{score}} + e^T d\log(S) + \frac{Td}{N}\log\left(\frac{SN}{T}\right) \\ &\lesssim \varepsilon_{\text{score}} + \varepsilon + \frac{\varepsilon}{\log(S)T^2}\Big(\log(S) + 3T\Big) \\ &\lesssim \varepsilon_{\text{score}} + \varepsilon.\end{aligned}$$

Remark 2 (Step size schedule) *In Corollary 1, we adopt a uniform step size schedule for simplicity. This choice is optimal in the sense that it minimizes the worst-case upper bound for a fixed number of steps N, and it is also empirically effective (Campbell et al., 2022). That said, other step size schedules commonly used in practice and theory achieve the same iteration complexity, including the exponential-then-constant schedule (defined as in Corollary 2, and used in Liang et al. (2025b)) and the log-linear schedule (Lou et al., 2024). In these works, early stopping is introduced to maintain numerical stability in score estimation during training, and also to ensure a small discretization error. Our result shows that, under Assumption 1, early stopping is not necessary for small discretization error.*

D.3. Proof of Theorem 2

Recall that the path measures of the backward process and the sampling process are denoted by $Q \overset{d}{=} \{\overleftarrow{q}_t\}_{t\in[0,T-\delta]}$ and $P \overset{d}{=} \{p_t\}_{t\in[0,T-\delta]}$, respectively. It can be checked that the path measure Q is absolutely continuous with respect to P. By Girsanov's theorem for the backward process (e.g. Ren et al. (2025, Corollary 3.4)), it satisfies

$\mathsf{KL}(Q \,\|\, P)$

$$= \mathsf{KL}(\overleftarrow{q}_0 \| p_0) + \frac{1}{S} \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k,t}} \Bigg[\sum_{i\in[d]} \sum_{c\in[S]} s_{T-t}(x_t \oplus_i c, x_t) \\ D\left(\widehat{s}_{T-t_k}(x_{t_k} \oplus_i c, x_{t_k}), s_{T-t}(x_t \oplus_i c, x_t)\right) \mathrm{d}t \Bigg].$$

Following the same analysis as in Eqns. (60) to (67) in Appendix D.1, we arrive at

$$\begin{aligned} \mathsf{KL}(Q \,\|\, P) &= \varepsilon_{\text{score}} + \mathsf{KL}(\overleftarrow{q}_0 \| p_0) + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} (\varphi(T-t) - \varphi(T-t_k))\, \mathrm{d}t \\ &= \varepsilon_{\text{score}} + \mathsf{KL}(q_T \| p_0) + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} (\varphi(T-t) - \varphi(T-t_k))\, \mathrm{d}t, \end{aligned}$$

where the function $\varphi(\cdot)$ is defined as in Eqn. (66)

$$\varphi(t) := \frac{1}{S} \mathbb{E}_{x_t \sim q_t} \left[\sum_{y_t : \mathrm{d_H}(y_t, x_t) = 1} - \log\left(s_t(y_t, x_t)\right) \right].$$

Thus, to achieve $\mathsf{KL}(Q \,\|\, P) \leqslant \varepsilon_{\text{score}} + O(1)$, we need to select N, T and step size schedule such that

$$\mathsf{KL}(q_T \| p_0) + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} (\varphi(T-t) - \varphi(T-t_k))\, \mathrm{d}t = O(1). \tag{71}$$

In order to understand the first term in Eqn. (71), let us first consider the case when $T = 1$. By the assumption $q_0 \in \mathcal{P}^\gamma(\mathcal{X})$, therefore, we are ensured that

$$\mathsf{KL}(q_1 \| p_0) = \sum_{x\in\mathcal{X}} q_1(x) \log\big(q_1(x)\big) - \sum_{x\in\mathcal{X}} q_1(x) \log\big(p_0(x)\big) \geqslant \gamma d \log(S) \gg 1.$$

Hence, it implies that T must satisfy $T > 1$.

We then focus on the analysis of the second term in Eqn. (71). We aim to show that the changing rate, i.e., $-\varphi'(t)$, is lower bounded as we come close to the target data distribution (i.e., $t \in [0, 1]$), which in turn leads to a lower bound on the difference $\varphi(T-t) - \varphi(T-t_k)$. We proceed with our analysis under the information-theoretic framework.

For notational convenience, given every $i \in [d]$ and $c \in [S]$, let us define

$$\varphi_{i,c}(t) = \mathbb{E}_{x_t \sim q_t}\left[-\log\left(s_t(x_t \oplus_i c, x_t)\right] = \mathbb{E}_{x_t \sim q_t}\left[-\log\left(s_t(N_{i,c}(x_t), x_t)\right]\right. \tag{72}$$

where the operator $N_{i,c} : \mathcal{X} \to \mathcal{X}$ is defined as $N_{i,c}(x) = x \oplus_i c$. It is easy to check that

$$\varphi(t) = 1/S \sum_{i\in[d]} \sum_{c\in[S]} \varphi_{i,c}(t).$$

Notice that $N_{i,c}$ is a bijection in $\mathcal{X}$. We define $N_{i,-c} = (N_{i,-c})^{-1} = N_{i,S-c}$, where $(N_{i,-c})^{-1}$ is denoted as the inverse function.

Since $\varphi(t)$ can be written as a linear combination of $\varphi_{i,c}(t)$, it suffices to study the properties of the individual $\varphi_{i,c}(t)$ to characterize $\varphi(t)$. To begin with, the following lemma provides a characterization of $\varphi(t)$ and $\varphi_{i,c}(t)$ as information-theoretic quantities.

Lemma 11 *For $\varphi(t)$ and $\varphi_{i,c}(t)$ defined in Eqns. (66) and (72), we have*

$$\varphi_{i,c}(t) = \mathsf{KL}(q_t \,\|\, (N_{i,-c})_{\#} q_t);$$
$$\varphi(t) = -\frac{\partial}{\partial t}\mathsf{KL}(q_t \| p_0) = \sum_{i\in[d]} \sum_{c\in[S]} \mathsf{KL}(q_t \,\|\, (N_{i,c})_{\#} q_t),$$

where $(N_{i,c})_{\#}$ is denoted as the pushforward measure of q_t under operator $N_{i,c}$.

Lemma 11 allows us to write $\varphi_{i,c}(t)$ as the KL divergence between the marginal forward process and its pushforward under $N_{i,c}$. By viewing $N_{i,c}$ as an information channel, we can show it is in a special family of channels, named S-ary symmetric channel (Makur and Polyanskiy, 2018), which satisfies strong data processing inequality. Through this idea, we can prove the following lemma.

Lemma 12 *For $t \in (0, T]$, $\varphi_{i,c}(t)$ is differentiable in t and it holds that*

$$-\varphi'_{i,c}(t) \geqslant \varphi_{i,c}(t).$$

Consequently, Lemma 12 leads to

$$-\varphi'(t) = -\frac{1}{S}\sum_{i\in[d]}\sum_{c\in[S]} \varphi'_{ic}(t) \geqslant \frac{1}{S}\sum_{i\in[d]}\sum_{c\in[S]} \varphi_{ic}(t) = \varphi(t).$$

Recall the log-Sobolev inequality in Lemma 7. We have, for any target distribution $q_0 \in \mathcal{P}^{\gamma}(\mathcal{X})$ and $t \in (0, 1)$,

$$-\varphi'(t) \geqslant \varphi(t) \geqslant \mathsf{KL}(q_t \| p_0) \geqslant \mathsf{KL}(q_1 \| p_0) \geqslant \gamma d \log(S).$$

Let us go back to the second term in Eqn. (71). By the fundamental theorem of calculus, we obtain

$$\begin{aligned}\int_{t_k}^{t_{k+1}} (\varphi(T-t) - \varphi(T-t_k))\,\mathrm{d}t &= \int_{T-t_{k+1}}^{T-t_k}\int_{T-t}^{T-t_k} -\varphi'(\tau)\mathrm{d}\tau\mathrm{d}t \\ &\gtrsim \int_{T-t_{k+1}}^{T-t_k} (t-t_k)\gamma d\log(S)\mathrm{d}t = \frac{1}{2}(t_{k+1}-t_k)^2\gamma d\log(S).\end{aligned}$$

Choose M such that $T - t_M \in [\frac{1}{2}, 1]$. Such M exists due to the fact that $T > 1$ and $\max_k\{t_{k+1} - t_k\} \leqslant \frac{1}{2}$. It holds in this case that

$$\begin{aligned}O(1) &= \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} (\varphi(T-t) - \varphi(T-t_k))\,\mathrm{d}t \\ &\gtrsim \sum_{k=M}^{N-1}\int_{t_k}^{t_{k+1}} (\varphi(T-t) - \varphi(T-t_k))\,\mathrm{d}t \\ &\gtrsim \sum_{k=M}^{N-1} (t_{k+1}-t_k)^2\gamma d\log(S). \qquad (73)\end{aligned}$$

By Cauchy-Schwarz inequality, it is direct to show that

$$\sum_{k=M}^{N-1}(t_{k+1}-t_k)^2 \geqslant \frac{1}{N-M}\left(\sum_{k=M}^{N-1}(t_{k+1}-t_k)\right)^2 = \frac{(T-t_M-\delta)^2}{N-M} \gtrsim \frac{1}{N-M}, \tag{74}$$

where in the last inequality, we use the fact that $T - t_M \geqslant \frac{1}{2}$ and $\delta \ll 1$. Plugging Eqn, (74) into Eqn. (73) leads to

$$N \geqslant N - M = \Omega(\gamma d \log(S)) = \widetilde{\Omega}(d).$$

D.4. Efficient sampling for high-entropy distributions

After Theorem 2, we mentioned that it is possible to get sublinear d iteration complexity with τ-leaping algorithm and uniform noising process when the target distribution is close to the uniform distribution on $\mathcal{X}$. Formally, we have the following result.

Theorem 4 *Let $q_0 \in \mathcal{P}(\mathcal{X})$ be the data distribution . Choose $0 = t_0 < t_1 < \ldots < t_N = T - \delta$ with exponential-then-constant step size schedule, i.e., $t_{k+1} - t_k \leqslant \kappa \min(1, T - t_{k+1})$ for $k = 0, \ldots, N-2$. Suppose $0 < \kappa < 0.9$, we have*

$$\mathsf{KL}(q_{T-\delta} \,\|\, p_{\text{output}}) \lesssim \varepsilon_{\text{score}} + \left(e^{-T} + \kappa \log(\delta^{-1})\right) \cdot \mathsf{KL}(q_0 \,\|\, \mathrm{Unif}(\mathcal{X})).$$

Theorem 4 reveals that, with an exponential-then-constant schedule and early stopping time δ, the error upper bound only depends on the initial KL divergence $\mathsf{KL}(q_0 \,\|\, \mathrm{Unif}(\mathcal{X}))$, which could be small if q_0 is close to the forward limit distribution $\mathrm{Unif}(\mathcal{X})$.

To be more concrete, we can choose $T = \log(\mathsf{KL}(q_0 \,\|\, \mathrm{Unif}(\mathcal{X}))/\varepsilon)$, $\delta^{-1} = \mathrm{poly}(d)$ and $\kappa = e^{-T}/\log(d)$ to achieve

$$\mathsf{KL}(q_{T-\delta} \,\|\, p_{\text{output}}) \lesssim \varepsilon_{\text{score}} + \varepsilon,$$

with iteration complexity

$$N = \widetilde{\Theta}\left(\frac{\mathsf{KL}(q_0 \,\|\, \mathrm{Unif}(\mathcal{X}))}{\varepsilon}\right),$$

which could be sublinear in d when $\mathsf{KL}(q_0 \,\|\, \mathrm{Unif}(\mathcal{X})) = o(d)$.

Proof of Theorem 4. The proof follows similar arguments as in the proof of Theorem 1.

Write $p_0 = \mathrm{Unif}(\mathcal{X})$ to be initial distribution of the sampling process. Following the proof of Eqn. (67), we get

$$\mathsf{KL}(q_{T-\delta} \,\|\, p_{T-\delta}) \leqslant \varepsilon_{\text{score}} + e^{-T}\mathsf{KL}(q_0 \,\|\, p_0) + \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \left(\varphi(T-t) - \varphi(T-t_k)\right) \mathrm{d}t, \tag{75}$$

where as shown in Lemma 11, $\varphi(t) = -\partial_t \mathsf{KL}(q_t \,\|\, p_0) \geqslant 0$. By Lemma 12, $\varphi(t)$ is a non-increasing function of $t \in (0, T]$. Thus, we have

$$\varphi(t) \leqslant \frac{1}{t}\int_0^t \partial_s \mathsf{KL}(q_s \,\|\, p_0)\mathrm{d}s = \frac{\mathsf{KL}(q_0 \,\|\, p_0)}{t}. \tag{76}$$

Without loss of generality, Choose M' such that $1 \leqslant M \leqslant N-1$ such that $T - t_M = 1$. For $1 \leqslant k < M$, $t_{k+1} - t_k = t_k - t_{k-1} = \kappa$. With Eqn. (76), we can show that

$$\begin{aligned}
&\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \big(\varphi(T-t) - \varphi(T-t_k)\big)\,\mathrm{d}t \\
&\leqslant \sum_{k=0}^{N-1} (t_{k+1}-t_k)\big(\varphi(T-t_{k+1}) - \varphi(T-t_k)\big)\,\mathrm{d}t \\
&= (t_N - t_{N-1})\varphi(T-t_N) + \sum_{k=1}^{N-1} \big((t_k - t_{k-1}) + (t_k - t_{k+1})\big)\varphi(T-t_k) - (t_1 - t_0)\varphi(T) \\
&\overset{\text{(a)}}{\leqslant} \frac{\kappa\delta}{1-\kappa} \cdot \frac{\mathsf{KL}(q_0 \,\|\, p_0)}{\delta} + \sum_{k=M}^{N-1} \frac{\kappa^2}{1-\kappa}(T-t_k) \cdot \frac{\mathsf{KL}(q_0 \,\|\, p_0)}{T-t_k} \\
&\lesssim (N-M)\kappa^2 \mathsf{KL}(q_0 \,\|\, p_0) = \log_{(1-\kappa)}(\delta)\kappa^2 \mathsf{KL}(q_0 \,\|\, p_0) \overset{\text{(b)}}{\leqslant} \kappa \log(\delta^{-1}) \mathsf{KL}(q_0 \,\|\, p_0),
\end{aligned}$$

where we apply Eqn. (76) in (a) and $\log(1-\kappa) \leqslant -\kappa$ in (b). Collecting the above bound and Eqn. (75) proves the result.

Appendix E. Proofs of results in Section 3.2

E.1. Proof of Theorem 3

We present the details of Algorithm 1 here.
For $t \in \{t_0, \ldots, t_N\}$, let p_t denote the marginal distribution of x_{t_k} in Algorithm 1. Using the data-processing inequality $\mathsf{KL}(\overleftarrow{q}_T \| p_T) \leqslant \mathsf{KL}(\overleftarrow{q}_{t_0,\ldots,t_N} \| p_{t_0,\ldots,t_N})$ and Lemma 4, we decompose the KL divergence between the target distribution $q_0 \equiv \overleftarrow{q}_T$ and the output distribution p_T as follows:

$$\mathsf{KL}(\overleftarrow{q}_T \| p_T) \leqslant \mathsf{KL}(\overleftarrow{q}_0 \| p_0) + \sum_{k=0}^{N-1} \mathbb{E}_{x_{t_k} \sim \overleftarrow{q}_{t_k}} \left[\mathsf{KL}\left(\overleftarrow{q}_{t_{k+1}|t_k}(\cdot|x_{t_k}) \| p_{t_{k+1}|t_k}(\cdot|x_{t_k}) \right) \right]. \tag{77}$$

The first term, *initialization error*, was bounded in Conforti et al. (2025); Liang et al. (2025a) as follows:

$$\mathsf{KL}(\overleftarrow{q}_0 \| p_0) \leqslant e^{-T} d(1 + \log S + T) \lesssim e^{-T} d \log S. \tag{78}$$

The following lemma states that for each k, conditioned on x_{t_k}, we can consider a CTMC on the interval $[t_k, t_{k+1}]$, with marginals $p_{t_{k+1}|t_k}(\cdot \mid x_{t_k})$ at time t_{k+1}. The proof is given in Section G.3.

Lemma 13 *Fix $k = 0, \ldots, N-1$. Let x_{t_k} and $x_{t_{k+1}}$ be as in Algorithm 1. Let $(y_t)_{t \in [t_k, t_{k+1}]}$ be a CTMC with $y_{t_k} = x_{t_k}$ and the following rate matrix:*

$$\widehat{Q}_t(a,b) := \begin{cases} \widehat{s}_{T-t_k}(y_{t_k} \odot_i b^i, y_{t_k}) \frac{e^{T-t_k}-1}{e^{T-t}-1} \mathbb{I}\{a^i = \mathrm{MASK}\}, & \text{if } \mathrm{d_H}(a,b) = 1, a^i \neq b^i, \text{ and } y^i_{t_k} = \mathrm{MASK}, \\ -\sum_{c \neq a} \widehat{Q}_t(a,c), & \text{if } a = b, \\ 0, & \text{otherwise.} \end{cases} \tag{79}$$

Then, $x_{t_{k+1}}$ has the same distribution as $y_{t_{k+1}}$.

We continue Eqn. (77) with marginals $p_{t|t_k}(\cdot \mid x_{t_k})$ of this CTMC:

$$\begin{aligned}\mathsf{KL}(\overleftarrow{q}_T \| p_T) &\lesssim e^{-T} d \log S + \sum_{k=0}^{N-1} \mathbb{E}_{x_{t_k} \sim \overleftarrow{q}_{t_k}} \left[\mathsf{KL}\left(\overleftarrow{q}_{t_{k+1}|t_k}(\cdot|x_{t_k}) \| p_{t_{k+1}|t_k}(\cdot|x_{t_k}) \right) \right] \\ &= e^{-T} d \log S + \sum_{k=0}^{N-1} \mathbb{E}_{x_{t_k} \sim \overleftarrow{q}_{t_k}} \left[\int_{t_k}^{t_{k+1}} \frac{\partial}{\partial t} \mathsf{KL}\left(\overleftarrow{q}_{t|t_k}(\cdot \mid x_{t_k}) \| p_{t|t_k}(\cdot \mid x_{t_k}) \right) \mathrm{d}t \right]. \end{aligned} \tag{80}$$

Using rate matrices as in Lemma 13, we apply Lemma 5 to upper bound the second term:

$$\begin{aligned}&\mathsf{KL}(\overleftarrow{q}_T \| p_T) \\ &\lesssim e^{-T} d \log S + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k, t}} \sum_{y \neq x_t} \left[\widehat{Q}_t(x_t, y) - \overleftarrow{Q}_t(x_t, y) + \overleftarrow{Q}_t(x_t, y) \log \left(\frac{\overleftarrow{Q}_t(x_t, y)}{\widehat{Q}_t(x_t, y)} \right) \right] \mathrm{d}t. \end{aligned} \tag{81}$$

Fix $k \in \{0, \ldots, N-1\}$ and $t \in [t_k, t_{k+1})$. Let $\ell := t_k$. Invoking Eqn. (79) leads to

$$\begin{aligned}&\sum_{y \neq x_t} \left[\widehat{Q}_t(x_t, y) - \overleftarrow{Q}_t(x_t, y) + \overleftarrow{Q}_t(x_t, y) \log \left(\frac{\overleftarrow{Q}_t(x_t, y)}{\widehat{Q}_t(x_t, y)} \right) \right] \\ &= \sum_{i \in m(x_t)} \sum_{c \in [S]} \left[\widehat{Q}_t(x_t, x_t \odot_i c) - \overleftarrow{Q}_t(x_t, x_t \odot_i c) + \overleftarrow{Q}_t(x_t, x_t \odot_i c) \log \left(\frac{\overleftarrow{Q}_t(x_t, x_t \odot_i c)}{\widehat{Q}_t(x_t, x_t \odot_i c)} \right) \right] \\ &= \sum_{i \in m(x_t)} \sum_{c \in [S]} \left[\frac{e^{T-\ell} - 1}{e^{T-t} - 1} \widehat{s}_{T-\ell}(x_\ell \odot_i c, x_\ell) - s_{T-t}(x_t \odot_i c, x_t) \right. \\ &\qquad \left. + s_{T-t}(x_t \odot_i c, x_t) \log \left(\frac{s_{T-t}(x_t \odot_i c, x_t)}{\frac{e^{T-\ell}-1}{e^{T-t}-1} \widehat{s}_{T-\ell}(x_\ell \odot_i c, x_\ell)} \right) \right] \\ &= \sum_{i \in m(x_t)} \sum_{c \in [S]} s_{T-t}(x_t \odot_i c, x_t) D\left(\frac{e^{T-\ell} - 1}{e^{T-t} - 1} \widehat{s}_{T-\ell}(x_\ell \odot_i c, x_\ell), s_{T-t}(x_t \odot_i c, x_t) \right). \end{aligned} \tag{82}$$

Next, observe that the Bregman divergence satisfies the following law of cosines:

$$D(\alpha, \gamma) = D(\alpha, \beta) + D(\beta, \gamma) + (\alpha - \beta) \frac{\beta - \gamma}{\beta \gamma}.$$

We apply this decomposition to each term of Eqn. (82) with

$$\alpha = \frac{e^{T-\ell} - 1}{e^{T-t} - 1} \widehat{s}_{T-\ell}(x_\ell \odot_i c, x_\ell), \qquad \beta = \frac{e^{T-\ell} - 1}{e^{T-t} - 1} s_{T-\ell}(x_\ell \odot_i c, x_\ell), \quad \text{and} \quad \gamma = s_{T-t}(x_t \odot_i c, x_t).$$

In the following, we slightly abuse the notation and write $x_t := (x_t \odot_i c, x_t)$ and $x_\ell := (x_\ell \odot_i c, x_\ell)$ whenever $i \in m(x_t)$ and $c \in [S]$ are fixed. We proceed for fixed i, c:

$$\begin{aligned}
&s_{T-t}(x_t \odot_i c, x_t) D\left(\frac{e^{T-\ell}-1}{e^{T-t}-1}\widehat{s}_{T-\ell}(x_\ell), s_{T-t}(x_t)\right) \\
&= s_{T-t}(x_t) D(\widehat{s}_{T-\ell}(x_\ell), s_{T-\ell}(x_\ell)) \\
&\quad + s_{T-t}(x_t) D\left(\frac{e^{T-\ell}-1}{e^{T-t}-1} s_{T-\ell}(x_\ell), s_{T-t}(x_t)\right) \\
&\quad\quad + \frac{\widehat{s}_{T-\ell}(x_\ell) - s_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)} \left(\frac{e^{T-\ell}-1}{e^{T-t}-1} s_{T-\ell}(x_\ell) - s_{T-t}(x_t)\right).
\end{aligned}$$

Note that we simplified the first term as $D(\alpha x, \alpha y) = D(x, y)$. Observing that $\frac{e^{T-\ell}-1}{e^{T-t}-1} s_{T-\ell} \equiv s_{T-t}$ by Eqn. (52), this can be rearranged as follows:

$$\begin{aligned}
s_{T-t}(x_t) D\left(\frac{e^{T-\ell}-1}{e^{T-t}-1}\widehat{s}_{T-\ell}(x_\ell), s_{T-t}(x_t)\right) &= \underbrace{\frac{e^{T-\ell}-1}{e^{T-t}-1} s_{T-\ell}(x_\ell) D(\widehat{s}_{T-\ell}(x_\ell), s_{T-\ell}(x_\ell))}_{=: T_1} \\
&\quad + \underbrace{s_{T-t}(x_t) D\left(s_{T-t}(x_\ell), s_{T-t}(x_t)\right)}_{=: T_2} \\
&\quad + \underbrace{(s_{T-t}(x_\ell) - s_{T-t}(x_t)) \log \frac{\widehat{s}_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)}}_{=: T_3}.
\end{aligned}$$

First term, T_1, after summation equals to the score entropy loss:

$$\begin{aligned}
&\mathbb{E}_{x_t, x_\ell \sim \overleftarrow{q}_{t,\ell}} \sum_{i \in m(x_t)} \sum_{c \in [S]} \frac{e^{T-\ell}-1}{e^{T-t}-1} s_{T-\ell}(x_\ell \odot_i c, x_\ell) D(\widehat{s}_{T-\ell}(x_\ell \odot_i c, x_\ell), s_{T-\ell}(x_\ell \odot_i c, x_\ell)) \\
&= \mathbb{E}_{x_\ell \sim \overleftarrow{q}_\ell} \sum_{i \in m(x_\ell)} \sum_{c \in [S]} e^{t-\ell} s_{T-\ell}(x_\ell \odot_i c, x_\ell) D(\widehat{s}_{T-\ell}(x_\ell \odot_i c, x_\ell), s_{T-\ell}(x_\ell \odot_i c, x_\ell)) \\
&= e^{t-\ell} \mathcal{L}_{\mathrm{SE}}(T-\ell, \widehat{s}_{T-\ell}, s_{T-\ell}),
\end{aligned}$$

where we used in the second line that $\Pr(x_t^i = \mathrm{MASK} \mid x_\ell^i = \mathrm{MASK}) = \frac{1-e^{-(T-t)}}{1-e^{-(T-\ell)}}$. Therefore, recalling that $\ell := t_k$,

$$\begin{aligned}
&\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} e^{t-t_k} \mathcal{L}_{\mathrm{SE}}(T-t_k, \widehat{s}_{T-t_k}, s_{T-t_k}) \mathrm{d}t \\
&= \sum_{k=0}^{N-1} (e^{t_{k+1}-t_k} - 1) \mathcal{L}_{\mathrm{SE}}(T-t_k, \widehat{s}_{T-t_k}, s_{T-t_k}) \lesssim \varepsilon_{\mathrm{score}},
\end{aligned} \tag{83}$$

where we used $\Delta = O(1)$ and Assumption 1 in the last inequality. Next lemma describes a martingale property of the score function. The proof is given in Section F.7.

Lemma 14 *Consider the masking noising process and let $0 \leqslant \ell < t < T$. Then, for any $c \in \mathcal{V}$ and $i \in m(x_\ell)$,*

$$\mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)} \left[\left(s_{T-t}(x_\ell \odot_i c, x_\ell) - s_{T-t}(x_t \odot_i c, x_t) \right) \mathbb{I}\{i \in m(x_t)\} \right] = 0.$$

Using Lemma 14, we get that the last term, T_3, contributes zero after conditioning on x_{t_k}:

$$\sum_{i \in [d]} \sum_{c \in [S]} \mathbb{E}_{x_t \sim \overleftarrow{q}_{t|t_k}(\cdot|x_{t_k})} \mathbb{I}\{i \in m(x_t)\} (s_{T-t}(x_{t_k} \odot_i c, x_{t_k}) - s_{T-t}(x_t \odot_i c, x_t)) = 0. \tag{84}$$

To control terms T_2, we use the following lemma, which is proven in Section G.4.

Lemma 15 *Let $\ell < t$. Then, for $\mathcal{I}(t)$ defined in Eqn. (10),*

$$\begin{aligned} &\mathbb{E}_{x_\ell, x_t \sim \overleftarrow{q}_{\ell,t}} \sum_{i \in m(x_t)} \sum_{c \in [S]} s_{T-t}(x_t \odot_i c, x_t) D\left(s_{T-t}(x_\ell \odot_i c, x_\ell), s_{T-t}(x_t \odot_i c, x_t) \right) \\ &= \int_\ell^t e^{t-v} \mathcal{I}(T-v) \mathrm{d}v. \end{aligned}$$

After summation of over $i \in m(x_t)$ and $c \in [S]$, we express the contributions of terms T_2 using Lemma 15 as follows:

$$\begin{aligned} &\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k,t}} \sum_{i \in m(x_t)} \sum_{c \in [S]} s_{T-t}(x_t \odot_i c, x_t) D\left(s_{T-t}(x_{t_k} \odot_i c, x_{t_k}), s_{T-t}(x_t \odot_i c, x_t) \right) \mathrm{d}t \\ &= \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \int_{t_k}^{t} e^{t-v} \mathcal{I}(T-v) \mathrm{d}v \mathrm{d}t \leqslant \sum_{k=0}^{N-1} h_k \int_{T-t_{k+1}}^{T-t_k} \mathcal{I}(t) \mathrm{d}t, \end{aligned} \tag{85}$$

where we used $\Delta = O(1)$ and non-negativity of conditional mutual information in the last inequality.

Collecting Eqns. (78), (83), and (85) proves

$$\mathsf{KL}(q_0 \,\|\, p_T) \lesssim \varepsilon_{\text{score}} + e^{-T} d \log S + \sum_{k=0}^{N-1} h_k \int_{T-t_{k+1}}^{T-t_k} \mathcal{I}(t) \mathrm{d}t.$$

E.2. Proof of Corollary 2

We upper bound the last term of Eqn. (11) under uniform and exponential-then-constant step size schedules. First, under the constant step size schedule, we have

$$\sum_{k=0}^{N-1} h_k \int_{T-t_{k+1}}^{T-t_k} \mathcal{I}(t) \mathrm{d}t = \frac{T}{N} \int_0^T \mathcal{I}(t) \mathrm{d}t \leqslant \frac{T}{N} \int_0^\infty \mathcal{I}(t) \mathrm{d}t = \frac{T}{N} \mathcal{B}(q_0),$$

where we used Lemma 16 in the last equality. Therefore, as long as

$$N \geqslant \frac{T \mathcal{B}(q_0)}{\varepsilon} = \widetilde{O}\left(\frac{\mathcal{B}(q_0)}{\varepsilon} \right),$$

we have that Eqn. (11) implies $\mathsf{KL}(q_0 \| p_T) \lesssim \varepsilon_{\text{score}} + \varepsilon$.

Next, under exponential-then-constant step size schedule, we bound the last term of Eqn. (11) as follows:

$$\sum_{k=0}^{N-1} h_k \int_{T-t_{k+1}}^{T-t_k} \mathcal{I}(t)\mathrm{d}t = \frac{\varepsilon}{d\log(S)} \int_0^{\varepsilon/(d\log S)} \mathcal{I}(t)\mathrm{d}t + \sum_{k=0}^{N-2} \int_{T-t_{k+1}}^{T-t_k} (t_{k+1}-t_k)\mathcal{I}(t)\mathrm{d}t$$
$$\leqslant \frac{\varepsilon^2}{\log(S)} + \kappa \sum_{k=0}^{N-2} \int_{T-t_{k+1}}^{T-t_k} \min(1, T-t_{k+1})\mathcal{I}(t)\mathrm{d}t \leqslant \varepsilon + \kappa \int_0^T \min(1,t)\mathcal{I}(t)\mathrm{d}t \leqslant \varepsilon + \kappa\mathcal{D}(q_0).$$

For $N > 0$, such step size schedule is possible with $\kappa = O\left(\frac{T+\log(\varepsilon^{-1}d\log(S))}{N}\right)$. Thus, choosing

$$N \geqslant \frac{(T+\log(\varepsilon^{-1}d\log(S)))\mathcal{D}(q_0)}{\varepsilon} = \widetilde{O}\left(\frac{\mathcal{D}(q_0)}{\varepsilon}\right)$$

gives $\mathsf{KL}(q_0 \,\|\, p_T) \lesssim \varepsilon_{\text{score}} + \varepsilon$.

E.3. τ-leaping for masking discrete diffusion

In this section, we prove the analogue of Theorem 3 for the truncated τ-leaping algorithm. Note that since applying multiple jumps on a single coordinate is ill-defined in masking noising process (where should we transition if the τ-leaping algorithm requires two transitions $\text{MASK} \to 1$ at some coordinate?), we analyse the truncated version instead of the classical τ-leaping algorithm.

Theorem 5 *Let $q_{\text{data}} = q_0$ be a distribution on $[S]^d$. Let $0 < \delta < T$ and $0 = t_0 < t_1 < \ldots < t_N = T - \delta$. Let $h_k := t_{k+1} - t_k$ be the step size and assume that $\Delta := \max_k h_k = O(1)$. Let*

$$p_0 := \left(\left(1-e^{-T}\right)\delta_{\text{MASK}} + \frac{e^{-T}}{S}\sum_{k=1}^{S}\delta_k\right)^{\otimes d}.$$

Consider the masking noising process given by Eqn. (3) for the initial distribution q_0. Under Assumption 1, truncated τ-leaping Eqn. (8) initialized at p_0 produces a sample from $p_{\text{output}} = p_T$, such that

$$\mathsf{KL}(q_{\text{data}} \,\|\, p_{\text{output}})$$
$$\lesssim \varepsilon_{\text{score}} + e^{-T}d\log(S) + \sum_{k=0}^{N-1} h_k \int_{T-t_{k+1}}^{T-t_k} \mathcal{I}(t)\mathrm{d}t + \frac{d(T+\log(\delta^{-1}))^3}{N^2} + \frac{T+\log(\delta^{-1})}{N}C, \quad (86)$$

where

$$C := \frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}_{x_\ell \sim \overleftarrow{q}_\ell} \sum_{i \in m(x_\ell)} \sum_{c \in [S]} s_{T-\ell}(x_\ell \odot_i c, x_\ell) \left|\log \frac{\widehat{s}_{T-\ell}(x_\ell \odot_i c, x_\ell)}{s_{T-\ell}(x_\ell \odot_i c, x_\ell)}\right|.$$

Note that the bound in Eqn. (86) is similar to Eqn. (11) in the main text. We comment on these differences before proving the proof of the theorem.

Remark 3 *In order to obtain a sharp bound, we would like* $\widehat{Q}_t \approx \overleftarrow{Q}_t$ *for all* $t \in [0, T]$. *In the truncated* τ*-leaping algorithm analysed in this section, this results in*

$$\widehat{s}_{T-t} := \widehat{s}_{T-t_k} \approx s_{T-t}.$$

Informally, we establish this by showing

$$\widehat{s}_{T-t_k} \approx s_{T-t_k} \approx s_{T-t},$$

where the first approximation comes from Assumption 1 and the second from the properties of the score function for the masking noising process. However, in the proof of Theorem 3, the requirement $\widehat{Q}_t \approx \overleftarrow{Q}_t$ *is that*

$$\widehat{s}_{T-t} := \frac{e^{T-t_k}-1}{e^{T-t}-1}\widehat{s}_{T-t_k} \approx s_{T-t}.$$

Again, informally, invoking Assumption 1, we get

$$\widehat{s}_{T-t} := \frac{e^{T-t_k}-1}{e^{T-t}-1}\widehat{s}_{T-t_k} \approx \frac{e^{T-t_k}-1}{e^{T-t}-1}s_{T-t_k} = s_{T-t}.$$

Observe how with a small modification we got rid of the second approximation. This explains why, compared to the theorem in the main text, Theorem 5 has two extra terms and requires early stopping.

We can expect the constant C to be small. Observe that it also appears in the analysis of Conforti et al. (2025), as C_2^M in Theorem 3.2.1, in the form of the maximum rather then the average. Under the (one-sided) boundedness assumption $\widehat{s}_{T-t_k} \geqslant M^{-1}$, the constant C can be upper bounded via the Cauchy-Schwarz inequality.

Proof of Theorem 5. The proof follows closely the proof of Theorem 3 with several extra steps. We begin with Eqn. (81), which we provide below.

$$\begin{aligned}&\mathsf{KL}(\overleftarrow{q}_T \| p_T)\\ &\lesssim e^{-T} d \log S + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k,t}} \sum_{y \neq x_t} \left[\widehat{Q}_t(x_t, y) - \overleftarrow{Q}_t(x_t, y) + \overleftarrow{Q}_t(x_t, y) \log \left(\frac{\overleftarrow{Q}_t(x_t, y)}{\widehat{Q}_t(x_t, y)} \right) \right] \mathrm{d}t.\end{aligned} \tag{87}$$

Next, since for this sampler, the rate matrices $\widehat{Q}_t$ are the following:

$$\widehat{Q}_t(x, y) := \begin{cases} \widehat{s}_{T-t_k}(x_{t_k} \odot_i y^i, x_{t_k}) \mathbb{I}\{x^i = \mathrm{MASK}\}, & \text{if } \mathrm{d_H}(x, y) = 1, x^i \neq y^i, \text{ and } x^i_{t_k} = \mathrm{MASK}, \\ -\sum_{z \neq x} \widehat{Q}_t(x, z), & \text{if } y = x, \\ 0, & \text{otherwise,} \end{cases}$$

we continue with

$$\begin{aligned}
&\sum_{y\neq x_t}\left[\widehat{Q}_t(x_t,y)-\overleftarrow{Q}_t(x_t,y)+\overleftarrow{Q}_t(x_t,y)\log\left(\frac{\overleftarrow{Q}_t(x_t,y)}{\widehat{Q}_t(x_t,y)}\right)\right]\\
&=\sum_{i\in m(x_t)}\sum_{c\in[S]}\left[\widehat{Q}_t(x_t,x_t\odot_i c)-\overleftarrow{Q}_t(x_t,x_t\odot_i c)+\overleftarrow{Q}_t(x_t,x_t\odot_i c)\log\left(\frac{\overleftarrow{Q}_t(x_t,x_t\odot_i c)}{\widehat{Q}_t(x_t,x_t\odot_i c)}\right)\right]\\
&=\sum_{i\in m(x_t)}\sum_{c\in[S]}\left[\widehat{s}_{T-\ell}(x_\ell\odot_i c,x_\ell)-s_{T-t}(x_t\odot_i c,x_t)+s_{T-t}(x_t\odot_i c,x_t)\log\left(\frac{s_{T-t}(x_t\odot_i c,x_t)}{\widehat{s}_{T-\ell}(x_\ell\odot_i c,x_\ell)}\right)\right]\\
&=\sum_{i\in m(x_t)}\sum_{c\in[S]}s_{T-t}(x_t\odot_i c,x_t)D\left(\widehat{s}_{T-\ell}(x_\ell\odot_i c,x_\ell),s_{T-t}(x_t\odot_i c,x_t)\right).
\end{aligned}$$

We apply the law of cosines $D(\alpha,\gamma)=D(\alpha,\beta)+D(\beta,\gamma)+(\alpha-\beta)\frac{\beta-\gamma}{\beta\gamma}$ with

$$\alpha=\widehat{s}_{T-\ell}(x_\ell\odot_i c,x_\ell),\qquad \beta=s_{T-\ell}(x_\ell\odot_i c,x_\ell),\quad\text{and}\quad \gamma=s_{T-t}(x_t\odot_i c,x_t).$$

In the following, we slightly abuse the notation and write $x_t:=(x_t\odot_i c,x_t)$ and $x_\ell:=(x_\ell\odot_i c,x_\ell)$ whenever $i\in m(x_t)$ and $c\in[S]$ are fixed. We proceed for fixed i,c:

$$\begin{aligned}
&s_{T-t}(x_t)D\left(\widehat{s}_{T-\ell}(x_\ell),s_{T-t}(x_t)\right)\\
&=s_{T-t}(x_t)D(\widehat{s}_{T-\ell}(x_\ell),s_{T-\ell}(x_\ell))+s_{T-t}(x_t)D\left(s_{T-\ell}(x_\ell),s_{T-t}(x_t)\right)\\
&\quad+\frac{\widehat{s}_{T-\ell}(x_\ell)-s_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)}\left(s_{T-\ell}(x_\ell)-s_{T-t}(x_t)\right).
\end{aligned}$$

This can be rearranged as follows:

$$\begin{aligned}
s_{T-t}(x_t)D\left(\widehat{s}_{T-\ell}(x_\ell),s_{T-t}(x_t)\right)&=\underbrace{s_{T-\ell}(x_\ell)D(\widehat{s}_{T-\ell}(x_\ell),s_{T-\ell}(x_\ell))}_{=:T_1}\\
&\quad+\underbrace{s_{T-t}(x_t)D\left(s_{T-\ell}(x_\ell),s_{T-t}(x_t)\right)}_{=:T_2}\\
&\quad+\underbrace{\left(s_{T-\ell}(x_\ell)-s_{T-t}(x_t)\right)\log\frac{\widehat{s}_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)}}_{=:T_3}.
\end{aligned}$$

First term, T_1, after summation is upper bounded by the score entropy loss:

$$\begin{aligned}
&\mathbb{E}_{x_t,x_\ell\sim\overleftarrow{q}_{t,\ell}}\sum_{i\in m(x_t)}\sum_{c\in[S]}s_{T-\ell}(x_\ell\odot_i c,x_\ell)D(\widehat{s}_{T-\ell}(x_\ell\odot_i c,x_\ell),s_{T-\ell}(x_\ell\odot_i c,x_\ell))\\
&\leqslant\mathbb{E}_{x_\ell\sim\overleftarrow{q}_\ell}\sum_{i\in m(x_\ell)}\sum_{c\in[S]}s_{T-\ell}(x_\ell\odot_i c,x_\ell)D(\widehat{s}_{T-\ell}(x_\ell\odot_i c,x_\ell),s_{T-\ell}(x_\ell\odot_i c,x_\ell))\\
&=\mathcal{L}_{\mathrm{SE}}(T-\ell,\widehat{s}_{T-\ell},s_{T-\ell}),
\end{aligned}$$

and, by Assumption 1,

$$\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathcal{L}_{\mathrm{SE}}(T-\ell,\widehat{s}_{T-\ell},s_{T-\ell})\mathrm{d}t\leqslant\varepsilon_{\mathrm{score}}.\tag{88}$$

Now, we proceed with terms T_2 and T_3 differently compared to the proof of Theorem 3.

For the term T_2 we again apply the law of cosines for

$$\alpha = s_{T-\ell}(x_\ell), \qquad \beta = s_{T-t}(x_\ell), \quad \text{and} \quad \gamma = s_{T-t}(x_t).$$

This gives

$$\begin{aligned} s_{T-t}(x) D(s_{T-\ell}(x_\ell), s_{T-t}(x_t)) = &\underbrace{s_{T-t}(x_t) D(s_{T-\ell}(x_\ell), s_{T-t}(x_\ell))}_{=: T_{21}} \\ &+ \underbrace{s_{T-t}(x_t) D(s_{T-t}(x_\ell), s_{T-t}(x_t))}_{=: T_{22}} \\ &+ \underbrace{(s_{T-t}(x_\ell) - s_{T-t}(x_t)) \frac{s_{T-\ell}(x_\ell) - s_{T-t}(x_\ell)}{s_{T-t}(x_\ell)}}_{=: T_{23}}. \end{aligned}$$

For T_{21}, using Eqn. (52), observe that

$$D(s_{T-\ell}(x_\ell), s_{T-t}(x_\ell)) = \frac{e^{T-t}-1}{e^{T-\ell}-1} - 1 - \log \frac{e^{T-t}-1}{e^{T-\ell}-1} \leqslant \frac{(e^{T-\ell}-e^{T-t})^2}{2(e^{T-\ell}-1)(e^{T-t}-1)} \lesssim \kappa^2,$$

where κ is a parameter of the step size schedule: $t_{k+1} - t_k \lesssim \kappa \min(1, T - t_{k+1})$. It is possible to take $\kappa = O\left(\frac{T+\log(\delta^{-1})}{N}\right)$, where δ is the early stopping parameter. Therefore, the total contribution of terms T_{21} is:

$$\begin{aligned} &\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_t, x_{t_k} \sim \overleftarrow{q}_{t,t_k}} \sum_{i \in m(x_t)} \sum_{c \in [S]} s_{T-t}(x_t \odot_i c, x_t) D(s_{T-t_k}(x_{t_k} \odot_i c, x_{t_k}), s_{T-t}(x_{t_k} \odot_i c, x_{t_k})) \mathrm{d}t \\ &\lesssim \kappa^3 \sum_{k=0}^{N-1} \mathbb{E}_{x_t \sim \overleftarrow{q}_t} \sum_{c \in [S]} s_{T-t}(x_t \odot_i c, x_t) \\ &= \kappa^3 \sum_{k=0}^{N-1} \mathbb{E}_{x_t \sim \overleftarrow{q}_t} \sum_{i \in m(x_t)} \frac{\sum_{c \in [S]} q_t(x_t \odot_i c)}{q_t(x_t)} \leqslant \kappa^2 \big(T + \log(\delta^{-1})\big) d, \end{aligned} \tag{89}$$

as $\sum_{c \in [S]} q_t(x_t \odot_i c) = q_t(x_t)$.

The term T_{22} is identical to the term T_2 from the proof of Theorem 3, thus we use Lemma 15 and obtain after summation of over $i \in m(x_t)$ and $c \in [S]$:

$$\begin{aligned} &\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_{t_k}, x_t \sim \overleftarrow{q}_{t_k,t}} \sum_{i \in m(x_t)} \sum_{c \in [S]} s_{T-t}(x_t \odot_i c, x_t) D\left(s_{T-t}(x_{t_k} \odot_i c, x_{t_k}), s_{T-t}(x_t \odot_i c, x_t)\right) \mathrm{d}t. \\ &= \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \int_{t_k}^{t} e^{t-v} \mathcal{I}(T-v) \mathrm{d}v \mathrm{d}t \leqslant \sum_{k=0}^{N-1} h_k \int_{T-t_{k+1}}^{T-t_k} \mathcal{I}(t) \mathrm{d}t, \end{aligned} \tag{90}$$

where we used $\Delta = O(1)$ and non-negativity of conditional mutual information in the last inequality.

To control T_{23}, observe that by Eqn. (52),

$$\frac{s_{T-\ell}(x_\ell) - s_{T-t}(x_\ell)}{s_{T-t}(x_\ell)} = \frac{e^{T-t} - e^{T-\ell}}{e^{T-\ell} - 1},$$

and importantly does not depend on x_ℓ. This implies that upon summation over $i \in m(x_t)$ and $c \in [S]$ terms T_{23} contribute zero:

$$\begin{aligned}
&\sum_{i \in m(x_t)} \sum_{c \in [S]} (s_{T-t}(x_\ell \odot_i c, x_\ell) - s_{T-t}(x_t \odot_i c, x_t)) \frac{e^{T-t} - e^{T-\ell}}{e^{T-\ell} - 1} \\
&= \frac{e^{T-t} - e^{T-\ell}}{e^{T-\ell} - 1} \sum_{i \in m(x_t)} \left(\frac{\sum_{c \in [S]} q_t(x_\ell \odot_i c)}{q_t(x_\ell)} - \frac{\sum_{c \in [S]} q_t(x_t \odot_i c)}{q_t(x_t)} \right) \\
&= \frac{e^{T-t} - e^{T-\ell}}{e^{T-\ell} - 1} \sum_{i \in m(x_t)} (1 - 1) \\
&= 0.
\end{aligned}$$

Thus, we remain with term T_3:

$$T_3 \coloneqq (s_{T-\ell}(x_\ell) - s_{T-t}(x_t)) \log \frac{\widehat{s}_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)}.$$

Crucially, unlike in the proof of Theorem 3, we no longer have a martingale property for this term. However, we can decompose

$$T_3 = (s_{T-\ell}(x_\ell) - s_{T-t}(x_\ell)) \log \frac{\widehat{s}_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)} + \underbrace{(s_{T-t}(x_\ell) - s_{T-t}(x_t)) \log \frac{\widehat{s}_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)}}_{\text{contributes zero by Lemma 14}}.$$

Thus, it remains to control

$$(s_{T-\ell}(x_\ell) - s_{T-t}(x_\ell)) \log \frac{\widehat{s}_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)} = \frac{e^{T-t} - e^{T-\ell}}{e^{T-t} - 1} s_{T-\ell}(x_\ell) \log \frac{\widehat{s}_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)}.$$

With exponential-then-constant step size schedule $t_{k+1} - t_k \lesssim \kappa \min(1, T - t_{k+1})$, we have

$$\left| \frac{e^{T-t} - e^{T-\ell}}{e^{T-t} - 1} \right| \lesssim \kappa, \quad \text{for} \quad \kappa = O\left(\frac{T + \log(\delta^{-1})}{N} \right).$$

The total contribution of terms T_3 can be upper bounded by the following:

$$\begin{aligned}
&\kappa \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x_\ell \sim \overleftarrow{q}_\ell} s_{T-\ell}(x_\ell) \left| \log \frac{\widehat{s}_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)} \right| \mathrm{d}t \\
&\lesssim \kappa \big(T + \log(\delta^{-1})\big) \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}_{x_\ell \sim \overleftarrow{q}_\ell} s_{T-\ell}(x_\ell) \left| \log \frac{\widehat{s}_{T-\ell}(x_\ell)}{s_{T-\ell}(x_\ell)} \right|. \qquad (91)
\end{aligned}$$

Collecting Eqns. (87), (88), (89), (90), and (91) proves

$$\mathsf{KL}(q_{\mathrm{data}} \,\|\, p_{\mathrm{output}})$$
$$\lesssim \varepsilon_{\mathrm{score}} + e^{-T} d \log(S) + \sum_{k=0}^{N-1} h_k \int_{T-t_{k+1}}^{T-t_k} \mathcal{I}(t)\mathrm{d}t + \frac{d(T+\log(\delta^{-1}))^3}{N^2} + \frac{T+\log(\delta^{-1})}{N} C,$$

where

$$C := \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}_{x_\ell \sim \overleftarrow{q}_\ell} \sum_{i \in m(x_\ell)} \sum_{c \in [S]} s_{T-\ell}(x_\ell \odot_i c, x_\ell) \left| \log \frac{\widehat{s}_{T-\ell}(x_\ell \odot_i c, x_\ell)}{s_{T-\ell}(x_\ell \odot_i c, x_\ell)} \right|.$$

Appendix F. Proofs of the main lemmas

F.1. Characterization of $\mathcal{B}(q_{\mathrm{data}})$ and $\mathcal{C}(q_{\mathrm{data}})$

The characterization of $\mathcal{B}(q_{\mathrm{data}})$ and $\mathcal{C}(q_{\mathrm{data}})$ is summarized in the following lemma.

Lemma 16 *Consider a masking noising process with initial distribution $q_0 = q_{\mathrm{data}}$. Let $\mathcal{C}(q_{\mathrm{data}})$ and $\mathcal{B}(q_{\mathrm{data}})$ be the total correlation and dual total correlation. Then,*

$$\mathcal{B}(q_{\mathrm{data}}) = \int_0^\infty \mathcal{I}(t)\mathrm{d}t \quad \textit{and} \quad \mathcal{C}(q_{\mathrm{data}}) = \int_0^\infty (e^t - 1)\mathcal{I}(t)\mathrm{d}t.$$

Consequently, $\mathcal{D}(q_{\mathrm{data}}) \leqslant \min(\mathcal{B}(q_{\mathrm{data}}), \mathcal{C}(q_{\mathrm{data}}))$.

Proof of Lemma 1. Let $p \equiv p(t) = e^{-t}$ be the probability that at time t a coordinate is unmasked and $X(p) \equiv (X_1, \ldots, X_d) := (x_t^1, \ldots, x_t^d)$. We also denote $X_{\mathcal{R}} := x_t^{-(i,j)}$ and $X_R := (X_i)_{i \in R}$ for $R \subseteq [d]$. With a slight abuse of notation we write $\mathcal{I}(p) := \mathcal{I}(t(p))$. We have

$$\mathcal{I}(p) := \sum_{i \neq j} \mathrm{I}(X_i; X_j \mid X_{\mathcal{R}}) = \sum_{i \neq j} p^2 \sum_{R \subseteq [d] \setminus \{i,j\}} p^{|R|}(1-p)^{d-2-|R|} \mathrm{I}(X_i; X_j \mid X_R),$$

where p^2 appears since for the term $\mathrm{I}(X_i; X_j \mid X_{\mathcal{R}})$ to be non-zero, both X_i and X_j must be unmasked. For $i \in [d]$, define

$$h_i(p) := \sum_{R \subseteq [d] \setminus \{i\}} p^{|R|}(1-p)^{d-1-|R|} \mathcal{H}(X_i \mid X_R),$$

with

$$\begin{aligned}
\frac{\mathrm{d}h_i(p)}{\mathrm{d}p} &= \sum_{R \subseteq [d] \setminus \{i\}} \left[|R| p^{|R|-1}(1-p)^{d-1-|R|} - (d-1-|R|) p^{|R|}(1-p)^{d-2-|R|} \right] \mathcal{H}(X_i \mid X_R) \\
&= \sum_{j \neq i} \sum_{R \subseteq [d] \setminus \{i,j\}} p^{|R|}(1-p)^{d-2-|R|} (\mathcal{H}(X_i \mid X_{R \cup \{j\}}) - \mathcal{H}(X_i \mid X_R)) \\
&= -\sum_{j \neq i} \sum_{R \subseteq [d] \setminus \{i,j\}} p^{|R|}(1-p)^{d-2-|R|} \mathrm{I}(X_i; X_j \mid X_R).
\end{aligned}$$

Therefore,

$$\mathcal{I}(p) = \sum_{i=1}^d p^2 \left(-\frac{\mathrm{d}h_i(p)}{\mathrm{d}p} \right).$$

Since $p = e^{-t}$ we have that $\mathrm{d}t = -\frac{\mathrm{d}p}{p}$ and we can write

$$\int_0^\infty \sum_{i\neq j} \mathrm{I}(X_i; X_j \mid X_{\mathcal{R}})\mathrm{d}t = \int_0^1 \sum_{i=1}^d p\left(-\frac{\mathrm{d}h_i(p)}{\mathrm{d}p}\right)\mathrm{d}p = \sum_{i=1}^d \big(-ph_i(p)\big)\Big|_0^1 + \int_0^1 \sum_{i=1}^d h_i(p)\mathrm{d}p.$$

Observe that

$$\frac{\mathrm{d}\mathcal{H}(X(p))}{\mathrm{d}p} = \sum_{i=1}^d h_i(p),$$

therefore,

$$\int_0^1 \sum_{i=1}^d h_i(p)\mathrm{d}p = \mathcal{H}(X(1)) - \mathcal{H}(X(0)) = \mathcal{H}(x_0).$$

Since $\sum_{i=1}^d h_i(1) = \sum_{i=1}^d \mathcal{H}(x_0^i \mid x_0^{-i})$, we proved the first part:

$$\int_0^\infty \mathcal{I}(t)\mathrm{d}t = \mathcal{H}(x_0) - \sum_{i=1}^d \mathcal{H}(x_0^i \mid x_0^{-i}) = \mathcal{B}(q_0).$$

We proceed similarly for the total correlation:

$$\begin{aligned}\int_0^\infty (e^t - 1)\sum_{i\neq j} \mathcal{I}(t)\mathrm{d}t &= \int_0^1 (1-p)\sum_{i=1}^d \left(-\frac{\mathrm{d}h_i(p)}{\mathrm{d}p}\right)\mathrm{d}p \\ &= -\left(\sum_{i=1}^d (1-p)h_i(p)\Big|_0^1 + \int_0^1 \sum_{i=1}^d h_i(p)\mathrm{d}p\right) = \sum_{i=1}^d \mathcal{H}(x_0^i) - \mathcal{H}(x_0) = \mathcal{C}(q_0).\end{aligned}$$

This concludes the proof.

F.2. Proof of Lemma 8

For any $i \in [d]$ and $c \in [S]$, let us define

$$f_{i,c}(t, x_t) \coloneqq s_{T-t}(x_t \oplus_i c, x_t). \tag{92}$$

The following analysis holds for all $i \in [d]$ and $c \in [S]$, so we may omit the index i, c in the following analysis, and write it as $f(t, x_t)$.

Consider the case that the process backward $\{x_t\}_{t\in[0,T]} \sim \{\overleftarrow{q}_t\}_{t\in[0,T]}$, which is a Poisson jump process with generator $\overleftarrow{L}_t$ such that

$$\begin{aligned}\left(\overleftarrow{L}_t f\right)(t,x) &= \sum_{y:\mathrm{d_H}(y,x)\leqslant 1} Q_{T-t}(y,x)s_{T-t}(y,x)\big(f(t,y) - f(t,x)\big) \\ &= \frac{1}{S}\sum_{y:\mathrm{d_H}(y,x)=1} s_{T-t}(y,x)\big(f(t,y) - f(t,x)\big).\end{aligned}$$

By Itô's formula for Poisson point process in Lemma 6, $f(t, x_t)$ satisfies the following stochastic differential equation: for $0 \leqslant \ell \leqslant t < T$,

$$f(t, x_t) = f(\ell, x_\ell) + \int_\ell^t \left[\partial_t f(s, x_{s^-}) + \left(\overleftarrow{L}_s f\right)(s, x_{s^-})\right] \mathrm{d}s + M_t, \tag{93}$$

where $x_{s^-} = \lim_{u \to s^-} x_s$, which exists for almost everywhere $s \in [0, T)$ under the Lebesgue measure, since we have finite number of jumps almost surely. The compensation process $\{M_u\}_{u \in [\ell, t]}$ in Eqn. (93) is defined as

$$M_u = \sum_{y_s : \mathrm{d_H}(y_s, x_s) = 1} \int_\ell^u \big(f(s, y_s) - f(s, x_s)\big)\big(\mathrm{d}N_s^{x_s, y_s} - \lambda_s^{x_s, y_s} \mathrm{d}s\big), \tag{94}$$

where $N_s^{x,y}$ is the counting process of jumps from x to y and we write the random counting measure as $\mathrm{d}N_s^{x,y}$. We define $\lambda_s^{x,y} = S^{-1} s_{T-t}(y, x) \mathbb{I}\{x_{s^-} = x\}$ to be intensity of the process $N_s^{x,y}$. Since $x_{s^-} = x_s$ almost everywhere $s \in (0, T)$ due to the finite number of jumps for each path almost surely, we can rewrite Eqn. (93) as

$$f(t, x_t) - f(\ell, x_\ell) = \int_\ell^t \left[\partial_t f(s, x_s) + \left(\overleftarrow{L}_s f\right)(s, x_s)\right] \mathrm{d}s + M_t. \tag{95}$$

To further simplify the right hand side, we assert that

$$\partial_t f(s, x_s) + \left(\overleftarrow{L}_s f\right)(s, x_s) = 0. \tag{96}$$

In order to see this, first, recall the definition Eqn. (92) and direct calculations give,

$$\begin{aligned}
&\partial_t f(s, x_s) + \left(\overleftarrow{L}_s f\right)(s, x_s) \\
&= \frac{\partial}{\partial s}\left(\frac{q_{T-s}(x_s \oplus_i c)}{q_{T-s}(x_s)}\right) \\
&\qquad + \frac{1}{S} \sum_{i' \in [d]} \sum_{c' \in [S]} s_{T-s}(x_s \oplus_{i'} c', x_s)\Big(s_{T-s}(x_s \oplus_{i'} c' \oplus_i c, x_s \oplus_{i'} c') - s_{T-s}(x_s \oplus_i c, x_s)\Big) \\
&\overset{\text{(a)}}{=} \frac{1}{S} \sum_{i' \in [d]} \sum_{c' \in [S]} s_{T-s}(x_s \oplus_i c, x_s)\Big(s_{T-s}(x_s \oplus_{i'} c', x_s) - s_{T-s}(x_s \oplus_i c \oplus_{i'} c', x_s \oplus_i c)\Big) \\
&\qquad + \frac{1}{S} \sum_{i' \in [d]} \sum_{c' \in [S]} s_{T-s}(x_s \oplus_{i'} c', x_s)\Big(s_{T-s}(x_s \oplus_{i'} c' \oplus_i c, x_s \oplus_{i'} c') - s_{T-s}(x_s \oplus_i c, x_s)\Big) \\
&= \frac{1}{S} \sum_{i' \in [d]} \sum_{c' \in [S]} \Big(s_{T-s}(x_s \oplus_i c, x_s) s_{T-s}(x_s \oplus_{i'} c', x_s) - s_{T-s}(x_s \oplus_{i'} c', x_s) s_{T-s}(x_s \oplus_i c, x_s)\Big) \\
&\quad + \frac{1}{S} \sum_{i' \in [d]} \sum_{c' \in [S]} \Big(s_{T-s}(x_s \oplus_{i'} c', x_s) s_{T-s}(x_s \oplus_{i'} c' \oplus_i c, x_s \oplus_{i'} c') \\
&\qquad\qquad - s_{T-s}(x_s \oplus_i c, x_s) s_{T-s}(x_s \oplus_i c \oplus_{i'} c', x_s \oplus_i c)\Big) \\
&\overset{\text{(b)}}{=} \frac{1}{S} \sum_{i' \in [d]} \sum_{c' \in [S]} \Big(s_{T-s}(x_s \oplus_{i'} c' \oplus_i c, x_s) - s_{T-s}(x_s \oplus_i c \oplus_{i'} c', x_s)\Big),
\end{aligned}$$

where in equality (a), we apply the Kolmogorov forward equation on q_{T-t}; in equality (b), we use the fact that $s_{T-t}(x,y)s_{T-t}(y,z) = s_{T-t}(x,z)$ for any $x, y, z \in \mathcal{X}$. It is direct to check that the $\oplus$ operators commute, i.e., for any $x_s \in \mathcal{X}$,

$$x_s \oplus_{i'} c' \oplus_i c = x_s \oplus_i c \oplus_{i'} c'.$$

This directly reveals that

$$\begin{aligned}
&\partial_t f(s, x_s) + \left(\overleftarrow{L}_s f\right)(s, x_s) \\
&= \frac{1}{S} \sum_{i' \in [d]} \sum_{c' \in [S]} \Big(s_{T-s}(x_s \oplus_{i'} c' \oplus_i c, x_s) - s_{T-s}(x_s \oplus_i c \oplus_{i'} c', x_s, x_s) \Big) = 0,
\end{aligned}$$

which proves Eqn. (96).

Taking $u = \ell$ in Eqn. (94), we have $M_\ell = 0$ almost surely, and M_u is a local martingale for $u \in [\ell, t]$ by definition. Recalling Lemma 10, we have

$$\sup_{s \in [\ell,t]} \sup_{x \in \mathcal{X}} f(s,x) \leqslant \log(S) + \max\Big\{ \log\big((T-t)^{-1}\big), 0 \Big\} < \infty.$$

Similarly, for the intensity of the counting process, we have

$$\sup_{s \in [\ell,t]} \sup_{x,y \in \mathcal{X}} \lambda_s^{x,y} \leqslant \frac{1}{S} s_{T-t}(y,x) \leqslant \frac{1}{S} \left(\log(S) + \max\Big\{ \log\big((T-t)^{-1}\big), 0 \Big\} \right) < \infty.$$

Now it is direct to check that

$$\sup_{s \in [\ell,t]} \mathbb{E}[|M_s|] \lesssim (t-\ell) d(S-1) \cdot \sup_{s \in [\ell,t]} \sup_{x,y \in \mathcal{X}} [f(s,x) \cdot \lambda_s^{x,y}] < \infty,$$

which ensures that $\{M_u\}_{u \in [\ell,t]}$ is L^1 and hence a martingale. By the definition of the martingale, we have

$$\mathbb{E}_{\overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}[M_t] = M_\ell = 0.$$

Go back to Eqn. (95). We obtain

$$\mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}\big[f(t,x_t) - f(\ell,x_\ell)\big] = \mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}[M_t] = 0. \tag{97}$$

Thus, we conclude that

$$\begin{aligned}
&\mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}\Big[\big(s_{T-\ell}(x_\ell \oplus_i c, x_\ell) - s_{T-t}(x_t \oplus_i c, x_t)\big) \log \widehat{s}_{T-\ell}(x_\ell \oplus_i c, x_\ell) \Big] \\
&= \mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}\big[f(\ell,x_\ell) - f(t,x_t)\big] \cdot \log \widehat{s}_{T-\ell}(x_\ell \oplus_i c, x_\ell) = 0,
\end{aligned}$$

where we plug in Eqn. (97) in the last line.

F.3. Proof of Lemma 9

The proof of Lemma 9 follows directly from exchanging the order of summation. Specifically, we can write

$$
\begin{aligned}
&\mathbb{E}_{x_t\sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} h\left(s_{T-t}(y_t,x_t)\right)\right] \\
&= \mathbb{E}_{x_t\sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} s_{T-t}(y_t,x_t)\log(s_{T-t}(y_t,x_t)) - s_{T-t}(y_t,x_t) + 1\right] \\
&= \mathbb{E}_{x_t\sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} \left(\frac{q_{T-t}(y_t)}{q_{T-t}(x_t)}\right)\log(s_{T-t}(y_t,x_t))\right] \\
&\qquad\qquad - \mathbb{E}_{x_t\sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} \left(\frac{q_{T-t}(y_t)}{q_{T-t}(x_t)}\right)\right] + d(S-1) \\
&= \sum_{x_t\in[S]^d}\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} q_{T-t}(y_t)\log(s_{T-t}(y_t,x_t)) - \sum_{x_t\in[S]^d}\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} q_{T-t}(y_t) + d(S-1) \\
&\overset{\text{(a)}}{=} -\sum_{x_t\in[S]^d}\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} q_{T-t}(x_t)\log(s_{T-t}(y_t,x_t)) - \sum_{y_t\in[S]^d}\sum_{x_t:\mathrm{d_H}(y_t,x_t)=1} q_{T-t}(y_t) + d(S-1) \\
&= -\mathbb{E}_{x_t\sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} \log(s_{T-t}(y_t,x_t))\right] - d(S-1) + d(S-1) \\
&= \mathbb{E}_{x_t\sim \overleftarrow{q}_t}\left[\sum_{y_t:\mathrm{d_H}(y_t,x_t)=1} -\log\left(s_t(y_t,x_t)\right)\right],
\end{aligned}
$$

where in equality (a), we switch the positions of x_t and y_t in the summations.

F.4. Proof of Lemma 10

Lemma 10 is a direct consequence of Liang et al. (2025b, Lemma 2). Here, we present a simplified proof based on Proposition 6. It is easy to check that

$$
\alpha_t = \frac{1-e^{-t}}{1+(S-1)e^{-t}} \in (0,1).
$$

By Eqn. (51), we have, for $\mathrm{d_H}(x,y)=1$,

$$
\begin{aligned}
s_t(y,x) = \frac{\mathbb{E}_{x_0\sim q_0}\alpha_t^{\mathrm{d_H}(y,x_0)}}{\mathbb{E}_{x_0\sim q_0}\alpha_t^{\mathrm{d_H}(x,x_0)}} &\leqslant \exp\left(-\log(\alpha_t)\cdot \sup_{y,x,x_0} |\,\mathrm{d_H}(y,x_0) - \mathrm{d_H}(y,x_0)|\right) \\
&\leqslant \exp\left(-\log(\alpha_t)\cdot \sup_{y,x}\mathrm{d_H}(y,x)\right) = \exp\big(-\log(\alpha_t)\big).
\end{aligned}
$$

With similar calculation, one can establish the reversed inequality

$$s_t(y,x) \geqslant \exp\left(\log(\alpha_t)\cdot \sup_{y,x} \mathrm{d_H}(y,x)\right) = \exp\big(\log(\alpha_t)\big).$$

As a result, we conclude

$$|\log\big(s_t(y,x)\big)| \leqslant -\log(\alpha_t) \lesssim \log(S) + \max\big\{\log(t^{-1}), 0\big\}.$$

F.5. Proof of Lemma 11

We first prove the first equation, i.e.,

$$\varphi_{i,c}(t) = \mathsf{KL}(q_t \,\|\, (N_{i,-c})_{\#} q_t).$$

Recall the definition of $\varphi_{i,c}(t)$ that

$$\varphi_{i,c}(t) = \mathbb{E}_{x_t\sim q_t}\left[-\log\left(\frac{q_t\left(N_{i,c}(x_t)\right)}{q_t(x_t)}\right)\right] = \sum_{x_t\in\mathcal{X}} q_t(x)\log\left(\frac{q_t(x)}{q_t\left(N_{i,c}(x_t)\right)}\right). \tag{98}$$

The definition of the pushforward measure gives, for any $x \in \mathcal{X}$,

$$(N_{i,-c})_{\#} q_t(x) = q_t\left(N_{i,c}(x)\right). \tag{99}$$

By Eqn. (99), we can rewrite Eqn. (98) as

$$\varphi_{i,c}(t) = \sum_{x_t\in\mathcal{X}} q_t(x)\log\left(\frac{q_t(x)}{(N_{i,-c})_{\#} q_t(x_t)}\right) = \mathsf{KL}(q_t \,\|\, (N_{i,-c})_{\#} q_t),$$

which proves the equation.

For the second line, the definition of KL divergence gives

$$\begin{aligned} -\frac{\partial}{\partial t}\mathsf{KL}(q_t\|p_0) = -\frac{\partial}{\partial t}\sum_{x\in[S]^d} q_t(x)\log(q_t(x)) &= -\sum_{x\in[S]^d}\frac{\mathrm{d}q_t(x)}{\mathrm{d}t}\big(\log(q_t(x))+1\big) \\ &= -\sum_{x\in[S]^d}\frac{\mathrm{d}q_t(x)}{\mathrm{d}t}\log(q_t(x)). \end{aligned} \tag{100}$$

Using the Kolmogorov forward equation for the forward noising process, we have

$$\frac{\mathrm{d}q_t(x)}{\mathrm{d}t} = \sum_{y\in\mathcal{X}} Q(x,y)q_t(x) = \frac{1}{S}\sum_{y:(y,x)=1} q_t(y) - \frac{d(S-1)}{S}q_t(x).$$

Plugging the equation above into Eqn. (100), we arrive at

$$\begin{aligned} -\frac{\partial}{\partial t}\mathsf{KL}(q_t\|p_0) &= -\sum_{x\in[S]^d}\left(\sum_{y:(y,x)=1}\left(\frac{1}{S}q_t(y)\right) - \frac{d(S-1)}{S}q_t(x)\right)\log(q_t(x)) \\ &= -\frac{1}{S}\sum_{x\in[S]^d}\sum_{y:(y,x)=1}(q_t(y)-q_t(x))\log(q_t(x)) \end{aligned}$$

$$= -\frac{1}{S}\sum_{x\in[S]^d}\sum_{y:(y,x)=1} q_t(x)\big(\log(q_t(y)) - \log(q_t(x))\big) = \varphi(t).$$

In addition, recall $\varphi(t) = 1/S\sum_{i\in[d]}\sum_{c\in[S]}\varphi_{i,c}(t)$. We reach

$$\varphi(t) = \frac{1}{S}\sum_{i\in[d]}\sum_{c\in[S]}\mathsf{KL}(q_t \,\|\, (N_{i,-c})_{\#}q_t) = \frac{1}{S}\sum_{i\in[d]}\sum_{c\in[S]}\mathsf{KL}(q_t \,\|\, (N_{i,c})_{\#}q_t).$$

F.6. Proof of Lemma 12

Let L be the time-homogeneous infinitesimal generator of the forward process. Since each coordinate $i \in [d]$ is updated independently in the forward process, we can write $L = L_i + L_{-i}$, where L_i only updates coordinate i, and L_{-i} updates all other coordinates. It is direct to show that L_i and L_{-i} commute, therefore, we have for any $u \geqslant 0$,

$$q_{t+u} = q_t e^{uL_i}e^{uL_{-i}}, \quad (N_{i,-c})_{\#}q_{t+u} = ((N_{i,-c})_{\#}q_t)e^{uL_i}e^{uL_{-i}},$$

where the second equation is due to the operator $N_{i,-c}$ commutes with the semigroup $\{e^{uL}\}_{u\geqslant 0}$. With this formulation, we reach

$$\varphi_{i,c}(t+u) = \mathsf{KL}(q_{t+u} \,\|\, (N_{i,-c})_{\#}q_{t+u}) \leqslant \mathsf{KL}(q_t e^{uL_i} \,\|\, ((N_{i,-c})_{\#}q_t)e^{uL_i}), \tag{101}$$

where in the last inequality, we apply the weak data processing inequality for KL divergence. Since both $N_{i,-c}$ and L_i only operate on the coordinate i, we have the decomposition

$$\mathsf{KL}(q_t e^{uL_i} \,\|\, ((N_{i,-c})_{\#}q_t)e^{uL_i}) = \mathbb{E}_{x^{-i}\sim(q_t)^{-i}}\left[\mathsf{KL}\big(q_t(\cdot|x^{-i})e^{uL_i} \,\|\, ((N_{i,-c})_{\#}q_t(\cdot|x^{-i}))e^{uL_i}\big)\right], \tag{102}$$

where $(q_t)^{-i}$ is the marginal distribution of q_t with coordinate i excluded. Define K_u to be the transition kernel on $[S]\times[S]$ induced by e^{uL_i}. It is shown that

$$K_u(v_1,v_2) = \begin{cases} \frac{1}{S} + (1-\frac{1}{S})e^{-u} & \text{if } v_1 = v_2; \\ \frac{1}{S}(1-e^{-u}) & \text{if } v_1 \neq v_2. \end{cases}$$

It can be directly checked that K_u is a S-ary symmetric channel with noise scale $\sigma_u = (1-S^{-1})(1-e^{-u})$. By Makur and Polyanskiy (2018, Proposition 12), a strong data processing inequality holds for the channel K_u, i.e., for any distribution p, q supported on $[S]$,

$$\mathsf{KL}(pe^{uL_i} \,\|\, qe^{uL_i}) \leqslant \eta_{\mathsf{KL}}(K_u)\mathsf{KL}(p \,\|\, q),$$

where $\eta_{\mathsf{KL}}(K_u)$ satisfies

$$\eta_{\mathsf{KL}}(K_u) \leqslant \left|1 - \sigma_u - \frac{\sigma_u}{S-1}\right| = 1 - \frac{S}{S-1}(1-S^{-1})(1-e^{-u}) = e^{-u}.$$

Applying this strong data processing inequality on Eqn. (102), we have

$$\mathsf{KL}(q_t e^{uL_i} \| ((N_{i,-c})_{\#}q_t)e^{uL_i}) = \mathbb{E}_{x^{-i}\sim(q_t)^{-i}}\left[\mathsf{KL}(q_t(\cdot|x^{-i})e^{uL_i} \,\|\, ((N_{i,-c})_{\#}q_t(\cdot|x^{-i}))e^{uL_i})\right]$$

$$\begin{aligned}
&\leqslant \mathbb{E}_{x_{-i}\sim(q_t)^{-i}}\left[e^{-u}\,\mathsf{KL}(q_t(\cdot|x^{-i})\,\|\,((N_{i,-c})_{\#}q_t(\cdot|x_{-i}))\right]\\
&\leqslant e^{-u}\,\mathbb{E}_{x^{-i}\sim(q_t)^{-i}}\left[\mathsf{KL}(q_t(\cdot|x^{-i})\,\|\,((N_{i,-c})_{\#}q_t(\cdot|x^{-i}))\right]\\
&= e^{-u}\mathsf{KL}(q_t\,\|\,(N_{i,-c})_{\#}q_t).
\end{aligned}$$

Then, by Eqn. (101), we have

$$\varphi_{i,c}(t+u) \leqslant \mathsf{KL}(q_t e^{uL_i}\,\|\,((N_{i,-c})_{\#}q_t)e^{uL_i}) \leqslant e^{-u}\mathsf{KL}(q_t\,\|\,(N_{i,-c})_{\#}q_t) = e^{-u}\varphi_{i,c}(t),$$

which holds for any $u \geqslant 0$. Therefore, the derivative can be bound as

$$\varphi'_{i,c}(t) = \lim_{u\to 0^+}\frac{\varphi_{i,c}(t+u)-\varphi_{i,c}(t)}{u} \leqslant \lim_{u\to 0^+}\frac{e^{-u}}{u}\varphi_{i,c}(t) = -\varphi_{i,c}(t),$$

which induces the target result

$$-\varphi'_{i,c}(t) \geqslant \varphi_{i,c}(t).$$

F.7. Proof of Lemma 14

The proof follows from Conforti et al. (2025), Lemma 5.2.2. We add the argument below for completeness. Let us define

$$f(t,x_t) := s_{T-t}(x_t \odot_i c, x_t)\mathbb{I}\{i \in m(x_t)\},$$

where the dependence on i and c is omitted for simplicity. By Lemma 6, we have for $0 \leqslant \ell \leqslant t < T$,

$$f(t,x_t) = f(\ell,x_\ell) + \int_\ell^t \left[\partial_t f(s,x_s) + (\overleftarrow{Q}_s f)(s,x_s)\right]\mathrm{d}s + M_t,$$

where $\{M_u\}_{u\in[\ell,t]}$ is the compensation process defined as

$$M_u = \int_\ell^u \sum_{i'\in m(x_s)}\sum_{c'\in[S]}\left(f(s,x_s\odot_{i'}c') - f(s,x_s)\right)\left(\mathrm{d}N_s^{x_s,x_s\odot_{i'}c'} - \lambda_s^{x_s,x_s\odot_{i'}c'}\mathrm{d}s\right),$$

With similar argument as in the proof of Lemma 8, we have $\mathbb{E}_{\overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}[M_t] = 0$, which leads to

$$\mathbb{E}_{x_t\sim\overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}[f(t,x_t) - f(\ell,x_\ell)] = \mathbb{E}_{x_t\sim\overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}\left[\int_\ell^t\left(\partial_t f(s,x_s) + (\overleftarrow{Q}_s f)(s,x_s)\right)\mathrm{d}s\right].$$

Taking derivative with respect to t on both side, we arrive

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{x_t\sim\overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}[f(t,x_t)] = \mathbb{E}_{x_t\sim\overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}\left[\partial_t f(t,x_t) + (\overleftarrow{Q}_t f)(t,x_t)\right].$$

By Proposition 6, $s_{T-t}(x_t\odot_i c, x_t) = \frac{1}{e^{T-t}-1}\frac{q_0(x_t\odot_i c)}{q_0(x_t)}$, and we have

$$\frac{\partial}{\partial t}f(t,x_t) = \frac{e^{T-t}}{e^{T-t}-1}f(t,x_t).$$

Next,

$$
\begin{aligned}
&(\overleftarrow{Q}_t f)(t, x_t) \\
&\quad = \sum_{i' \in m(x_t)} \sum_{c' \in [S]} s_{T-t}(x_t \oplus_{i'} c', x_t) \Big(s_{T-t}(x_t \odot_i c \odot_{i'} c', x_t \odot_{i'} c') \mathbb{I}\{i \in m(x_t \odot_{i'} c')\} \\
&\qquad\qquad - s_{T-t}(x_t \odot_i c, x_t) \mathbb{I}\{i \in m(x_t)\} \Big) \\
&\quad = \frac{1}{e^{T-t}-1} f(t, x_t) \left(\sum_{i' \in m(x_t) \setminus \{i\}} \sum_{c' \in [S]} \frac{q_0(x_t \odot_i c \odot_{i'} c')}{q_0(x_t \odot_i c)} - \sum_{i' \in m(x_t)} \sum_{c' \in [S]} \frac{q_0(x_t \odot_{i'} c')}{q_0(x_t)} \right) \\
&\quad = -\frac{1}{e^{T-t}-1} f(t, x_t).
\end{aligned}
$$

Thus, we have shown that

$$
\frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}[f(t, x_t)] = \mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}[f(t, x_t)],
$$

and therefore,

$$
\mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}[f(t, x_t)] = e^{t-\ell} \cdot f(\ell, x_\ell).
$$

Finally, in view of the relation $\Pr(x_t^i = \mathrm{MASK} \mid x_\ell^i = \mathrm{MASK}) = \frac{1-e^{t-T}}{1-e^{\ell-T}}$, we conclude the following

$$
\begin{aligned}
\mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)} \big[s_{T-t}(x_t \odot_i c, x_t)\, \mathbb{I}\{i \in m(x_t)\} \big] &= e^{t-\ell} \cdot s_{T-\ell}(x_\ell \odot_i c, x_\ell) \mathbb{I}\{i \in m(x_\ell)\} \\
&= \mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)} \big[s_{T-t}(x_\ell \odot_i c, x_\ell)\, \mathbb{I}\{i \in m(x_t)\} \big],
\end{aligned}
$$

which completes the proof of the desired result.

Appendix G. Proofs of the auxiliary lemmas

G.1. Proof of Lemma 2

For a continuous random variable U in $\mathbb{R}^d$ with density function p_U with respect to Lebesgue measure, define the differential entropy of U as

$$
\mathcal{H}^{\mathrm{diff}}(U) = -\int_{\mathbb{R}^d} p_U \log(p_U) \mathrm{d}x, \tag{103}
$$

where we adopt the convention $0 \log(0) = 0$ again. By the definition of mutual information, we have

$$
\begin{aligned}
I(W; W + \varepsilon_{\mathrm{noise}}) &= \mathcal{H}^{\mathrm{diff}}(W + \varepsilon_{\mathrm{noise}}) - \mathcal{H}^{\mathrm{diff}}(W + \varepsilon_{\mathrm{noise}} \mid W) \\
&\overset{\text{(a)}}{=} \mathcal{H}^{\mathrm{diff}}(W + \varepsilon_{\mathrm{noise}}) - \mathbb{E}_w \big[\mathcal{H}^{\mathrm{diff}}(w + \varepsilon_{\mathrm{noise}} \mid W = w) \big] \\
&\overset{\text{(b)}}{=} \mathcal{H}^{\mathrm{diff}}(W + \varepsilon_{\mathrm{noise}}) - \mathbb{E}_w \big[\mathcal{H}^{\mathrm{diff}}(\varepsilon_{\mathrm{noise}} \mid W = w) \big] \\
&\overset{\text{(c)}}{=} \mathcal{H}^{\mathrm{diff}}(W + \varepsilon_{\mathrm{noise}}) - \mathcal{H}^{\mathrm{diff}}(\varepsilon_{\mathrm{noise}}),
\end{aligned} \tag{104}
$$

where in (a), we use the chain rule of differential entropy; in (b), we apply the translation invariance property, i.e., $\mathcal{H}^{\text{diff}}(U) = \mathcal{H}^{\text{diff}}(c_0 + U)$ for any constant c_0; in (c), we use the condition that $\varepsilon_{\text{noise}} \perp\!\!\!\perp W$.

Denote the Gaussian density function with mean 0 and variance $\sigma^2 I_d$ as $\phi_\sigma(\cdot)$. Since $\varepsilon_{\text{noise}} \sim \mathcal{N}(0, \sigma^2 I_d)$, we can compute with Eqn. (103) that

$$\begin{aligned}
\mathcal{H}^{\text{diff}}(\varepsilon_{\text{noise}}) &= -\int_{\mathbb{R}^d} \phi_\sigma(x) \log\big(\phi_\sigma(x)\big) \mathrm{d}x \\
&= -\int_{\mathbb{R}^d} \phi_\sigma(x) \left(-\frac{d}{2}\log(2\pi\sigma^2) - \frac{\|x\|_2^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \frac{d}{2}\log(2\pi\sigma^2) + \frac{\mathbb{E}[\|\varepsilon_{\text{noise}}\|_2^2]}{2\sigma^2} \\
&= \frac{d}{2}\log(2\pi e\sigma^2), \qquad (105)
\end{aligned}$$

where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^d$. For $\mathcal{H}^{\text{diff}}(W + \varepsilon_{\text{noise}})$, notice that

$$\operatorname{Var}\big[W + \varepsilon_{\text{noise}}\big] = \operatorname{Var}[W] + \operatorname{Var}[\varepsilon_{\text{noise}}] + 2\operatorname{Cov}\big[W, \varepsilon_{\text{noise}}\big] = \operatorname{Var}[W] + \sigma^2 I_d.$$

By Cover (1999, Page 255), for distributions with the same finite variance, $\mathcal{H}^{\text{diff}}$ is maximized at the centered Gaussian random variable. Therefore, we have

$$\mathcal{H}^{\text{diff}}(W + \varepsilon_{\text{noise}}) \leqslant \mathcal{H}^{\text{diff}}\Big(\mathcal{N}\big(0, \operatorname{Var}[W] + \sigma^2 I_d\big)\Big) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\Big(\det\big(\operatorname{Var}[W] + \sigma^2 I_d\big)\Big),$$

where $\det(\cdot)$ is the determinant of matrices, and the calculation is the same as in Eqn. (105). Since $\operatorname{Var}[W]$ is a positive semidefinite matrix, we can apply the matrix inequality that

$$\begin{aligned}
\log\Big(\det\big(\operatorname{Var}[W] + \sigma^2 I_d\big)\Big) &= d\log(\sigma^2) + \log\Big(\det\big(I_d + \operatorname{Var}[W/\sigma^2]\big)\Big) \\
&\leqslant d\log(\sigma^2) + \operatorname{Tr}\big(\operatorname{Var}[W/\sigma^2]\big) \\
&= d\log(\sigma^2) + \frac{\operatorname{Tr}\big(\operatorname{Var}[W]\big)}{\sigma^2},
\end{aligned}$$

which leads to

$$\mathcal{H}^{\text{diff}}(W + \varepsilon_{\text{noise}}) \leqslant \frac{d}{2}\log(2\pi e\sigma^2) + \frac{\operatorname{Tr}\big(\operatorname{Var}[W]\big)}{2\sigma^2}. \qquad (106)$$

Plugging Eqns. (105) and (106) into Eqn. (104), we conclude that

$$I(W; W + \varepsilon_{\text{noise}}) \leqslant \frac{d}{2}\log(2\pi e\sigma^2) + \frac{\operatorname{Tr}\big(\operatorname{Var}[W]\big)}{2\sigma^2} - \frac{d}{2}\log(2\pi e\sigma^2) = \frac{\operatorname{Tr}\big(\operatorname{Var}[W]\big)}{2\sigma^2}.$$

G.2. Proof of Lemma 3

For $X \sim \operatorname{Bin}(n, 1/2)$, its pmf is given by

$$\mathbb{P}(X = x) = \binom{n}{x}\left(\frac{1}{2}\right)^n \propto \binom{n}{x}.$$

Notice that the equation to prove is equivalent to the following equation:

$$\sum_{x : x \bmod 2 = 0} \binom{n}{x} = \sum_{x : x \bmod 2 = 1} \binom{n}{x},$$

which follows from the binomial theorem for $0 = (1-1)^n$.

G.3. Proof of Lemma 13

The CTMC (79) in the lemma statement can be decomposed into d independent CTMCs for each coordinate. For coordinates i such that $x^i_{t_k} \neq \text{MASK}$ clearly neither Eqn. (79) nor Algorithm 1 makes a change. Next, we fix $i \in m(x_{t_k})$. First, we compute the probability that the i-th coordinate remains masked for $y_{t_{k+1}}$:

$$\begin{aligned}\Pr(y^i_{t_{k+1}} = \text{MASK} \mid y_{t_k}) &= \exp\left(\int_{t_k}^{t_{k+1}} \left(-\sum_{c\in[S]} \widehat{s}_{T-t_k}(y_{t_k} \odot_i c, y_{t_k}) \frac{e^{T-t_k}-1}{e^{T-t}-1}\right) \mathrm{d}t\right) \\ &= \exp(\widehat{Q}^i_k(\text{MASK})\Delta_k) \\ &= \mathcal{P}_k,\end{aligned}$$

where $\widehat{Q}^i_k(\text{MASK})$, Δ_k, and $\mathcal{P}_k$ are defined in Algorithm 1. Next, for $c \in [S]$ we can write

$$\Pr(y^i_{t_{k+1}} = c \mid x_{t_k}) = \Pr(x^i_{t_{k+1}} = c \mid x_{t_k} \text{ and } x^i_{t_{k+1}} \neq \text{MASK})(1-\mathcal{P}_k).$$

Since for any $t \in [t_k, t_{k+1})$ the rates $\widehat{Q}_t(x, x \odot_i c)$ are proportional to $\widehat{Q}^i_k(c)$, we get that

$$\Pr(y^i_{t_{k+1}} = c \mid x_{t_k} \text{ and } y^i_{t_{k+1}} \neq \text{MASK}) = \frac{\widehat{Q}^i_k(c)}{\sum_{b\in[S]} \widehat{Q}^i_k(b)},$$

which matches the expression in Algorithm 1. Therefore, the distribution of $y_{t_{k+1}}$ defined by the CTMC matches the distribution of $x_{t_{k+1}}$ from the algorithm.

G.4. Proof of Lemma 15

In view of the definition of $D(\cdot,\cdot)$, one can write

$$\begin{aligned}&s_{T-t}(x_t \odot_i c, x_t) D\left(s_{T-t}(x_\ell \odot_i c, x_\ell), s_{T-t}(x_t \odot_i c, x_t)\right) \\ &= s_{T-t}(x_\ell \odot_i c, x_\ell) - s_{T-t}(x_t \odot_i c, x_t) + s_{T-t}(x_t \odot_i c, x_t) \log \frac{s_{T-t}(x_t \odot_i c, x_t)}{s_{T-t}(x_\ell \odot_i c, x_\ell)}.\end{aligned}$$

The first two terms cancel out in expectation by Lemma 14; i.e., for any $c \in [S]$, one has

$$\mathbb{E}_{x_t \sim \overleftarrow{q}_{t|\ell}(\cdot|x_\ell)}\Big[\sum_{i\in m(x_t)} \left(s_{T-t}(x_\ell \odot_i c, x_\ell) - s_{T-t}(x_t \odot_i c, x_t)\right)\Big] = 0.$$

Next, using Eqn. (52), we obtain

$$\frac{s_{T-t}(x_t \odot_i c, x_t)}{s_{T-t}(x_\ell \odot_i c, x_\ell)} = \frac{q_0(x_t \odot_i c) q_0(x_\ell)}{q_0(x_t) q_0(x_\ell \odot_i c)}.$$

Using this relation, we continue

$$\mathbb{E}_{x_\ell, x_t \sim \overleftarrow{q}_{\ell,t}} \sum_{i\in m(x_t)} \sum_{c\in[S]} s_{T-t}(x_t \odot_i c, x_t) \log \frac{q_0(x_t \odot_i c) q_0(x_\ell)}{q_0(x_t) q_0(x_\ell \odot_i c)}$$

$$= \mathbb{E}_{y_\ell, y_t \sim \overleftarrow{q}_{\ell,t}} \sum_{i \notin m(y_t)} \log \frac{q_0(y_t) q_0(y_\ell \odot_i \text{MASK})}{q_0(y_i \odot_i \text{MASK}) q_0(y_\ell \odot_i y_t^i)}$$
$$= \sum_{i \in [d]} \mathbb{E}_{y_\ell, y_t \sim \overleftarrow{q}_{\ell,t}} \log \frac{q_0(y_t) q_0(y_\ell \odot_i \text{MASK})}{q_0(y_i \odot_i \text{MASK}) q_0(y_\ell \odot_i y_t^i)}, \tag{107}$$

where in the second line we used the definition of score function along with the natural bijection between the sets $\{(x, i, c)$, for $x \in \mathcal{X}$, $i \in m(x)$, and $c \in [S]\}$ and $\{(y, i)$, for $y \in \mathcal{X}$ and $i \notin m(y)\}$ to change the measure under the expectation:

$$x_t \to y_t \odot_i \text{MASK}$$
$$x_\ell \to y_\ell \odot_i \text{MASK}$$
$$x_t \odot_i c \to y_t$$
$$x_\ell \odot_i c \to y_\ell \odot_i y_t^i.$$

Note that since y_ℓ appears earlier in the backward process, y_ℓ^i can be masked or unmasked. Since the i-th element of $x_\ell \odot_i c$ is unmasked by construction, we explicitly set the i-th element of y_ℓ to y_t^i. The third line of Eqn. (107) follows from the fact that, for $i \in m(y_t)$, the term is equal to zero.

Next, we define, for fixed t, y_t, and $i \in [d]$,

$$f_i(y) := \log \frac{q_0(y \odot_i y_t^i)}{q_0(y \odot_i \text{MASK})},$$

and rewrite Eqn. (107) as follows:

$$\sum_{i \in [d]} \mathbb{E}_{y_\ell, y_t \sim \overleftarrow{q}_{\ell,t}} \log \frac{q_0(y_t) q_0(y_\ell \odot_i \text{MASK})}{q_0(y_i \odot_i \text{MASK}) q_0(y_\ell \odot_i y_t^i)} = \sum_{i \in [d]} \mathbb{E}_{y_\ell, y_t \sim \overleftarrow{q}_{\ell,t}} \left[f_i(y_t) - f_i(y_\ell) \right].$$

We observe that as the value $f_i(y)$ does not depend on the i-th coordinate of y, we can apply Dynkin's formula, Lemma 6 to the remaining $d-1$ coordinates for the forward process: $\mathbb{E}\left[f_i(y_t) - f_i(y_\ell)\right] = \mathbb{E} \int_t^\ell \sum_{j \neq i} \left[f_i(y_v) - f_i(y_v \odot_j \text{MASK})\right] \mathrm{d}v$. With this, we continue:

$$\sum_{i \in [d]} \mathbb{E}_{y_\ell, y_t \sim \overleftarrow{q}_{\ell,t}} \left[f_i(y_t) - f_i(y_\ell) \right]$$
$$= \sum_{i \in [d]} \int_\ell^t \mathbb{E}_{y_v, y_t \sim \overleftarrow{q}_{v,t}} \sum_{j \notin m(y_v) \cup \{i\}} \log \frac{q_0(y_v \odot_i y_t^i) q_0(y_v \odot_i \text{MASK} \odot_j \text{MASK})}{q_0(y_v \odot_i \text{MASK}) q_0(y_v \odot_i y_t^i \odot_j \text{MASK})} \mathrm{d}v$$
$$= \sum_{i \neq j \in [d]} \int_\ell^t \mathbb{E}_{y_v, y_t \sim \overleftarrow{q}_{v,t}} \log \frac{q_0(y_v \odot_i y_t^i) q_0(y_v \odot_i \text{MASK} \odot_j \text{MASK})}{q_0(y_v \odot_i \text{MASK}) q_0(y_v \odot_i y_t^i \odot_j \text{MASK})} \mathrm{d}v$$
$$= \sum_{i \neq j \in [d]} \int_\ell^t e^{t-v} \mathbb{E}_{y_v \sim \overleftarrow{q}_v} \log \frac{q_0(y_v) q_0(y_v \odot_i \text{MASK} \odot_j \text{MASK})}{q_0(y_v \odot_i \text{MASK}) q_0(y_v \odot_j \text{MASK})} \mathrm{d}v, \tag{108}$$

where in the third line, as before, we extended the sum to $j \in m(y_v) \backslash \{i\}$ since additional terms equal zero. The last line follows from

$$\Pr(y_v^i \neq \text{MASK} \mid y_t^i \neq \text{MASK}) = e^{v-t}.$$

Next, let $y_v^{-(i,j)}$ denote all unmasked elements of y_v, except i-th and j-th. We can write

$$\frac{q_0(y_v)q_0(y_v \odot_i \text{MASK} \odot_j \text{MASK})}{q_0(y_v \odot_i \text{MASK})q_0(y_v \odot_j \text{MASK})} = \frac{q_0(y_v^i, y_v^j \mid y_v^{-(i,j)})}{q_0(y_v^i \mid y_v^{-(i,j)})q_0(y_v^j \mid y_v^{-(i,j)})},$$

and thus,

$$\begin{aligned}
&\sum_{i \neq j \in [d]} \int_\ell^t e^{t-v} \mathbb{E}_{y_v \sim \overleftarrow{q}_v} \log \frac{q_0(y_v)q_0(y_v \odot_i \text{MASK} \odot_j \text{MASK})}{q_0(y_v \odot_i \text{MASK})q_0(y_v \odot_j \text{MASK})} \mathrm{d}v \\
&= \sum_{i \neq j} \int_\ell^t e^{t-v} \mathrm{I}(y_v^i; y_v^j \mid y_v^{-(i,j)}) \mathrm{d}v \\
&= \int_\ell^t e^{t-v} \mathcal{I}(T-v) \mathrm{d}v, \qquad (109)
\end{aligned}$$

as $y_v \sim q_{T-v}$. Combining Eqns. (107), (108), and (109) concludes the proof.