VOCABULARY-SIZE-INDEPENDENT CONVERGENCE OF DISCRETE DIFFUSION MODELS: ADJOINT EQUATIONS INDUCE THE RIGHT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete diffusion has become a leading framework for generative modeling in various applications including language, vision, and biology. Existing convergence theory, however, exhibits fundamental limitations. KL-based analyses diverge under singular priors such as the masked distribution, while bounds in total variation (TV) depend on the vocabulary size S and become vacuous for modern language tasks, where vocabularies contain hundreds of thousands of tokens. We develop a unified adjoint-equation-based framework that establishes vocabulary-size-independent convergence guarantees in any integral probability metric (IPM). To the best of our knowledge, our bounds are the first to be entirely free of S and applicable to both masked and uniform priors. Importantly, our results can extend existing step complexity guarantees to any IPM. Also, our theory relies only on a single standard rate-matrix regularity assumption and applies to general priors.

Five novel techniques drive our improvements: 1. working in the space of observables via adjoint equations rather than directly with probability measures; 2. a regularity analysis that yields bounds on any IPM; 3. a coupling argument that removes S-dependence under uniform transitions; and 4. score–marginal cancellation and 5. exit-routing techniques that remove S-dependence under masked transitions. Our framework thus sharply departs from prior analyses and avoids the shortcomings of pathspace-KL and existing TV-based approaches. Beyond convergence bounds, our framework provides a versatile toolkit for further theoretical study of discrete diffusion models, including principled choices of loss functions and vocabulary-size-independent step complexity.

1 INTRODUCTION

Diffusion-based generative models have achieved remarkable success across continuous-data domains such as images, audio, and video (Ho et al., 2020; Song et al., 2021b; Rombach et al., 2022). For discrete data, where Gaussian-based corruption does not directly apply, a parallel line of work has developed *discrete diffusion models* (Sohl-Dickstein et al., 2015; Austin et al., 2021; Hoogeboom et al., 2021; Campbell et al., 2022). These models have demonstrated strong empirical performance across a wide range of applications, including language modeling (Lou et al., 2024; Sahoo et al., 2024), image generation (Campbell et al., 2022), tabular data synthesis (Kotelnikov et al., 2023), molecular generation (Vignac et al., 2023), symbolic music generation (Campbell et al., 2022), genomic sequence design (Schiff et al., 2025), and protein design (Campbell et al., 2024).

Most discrete diffusion models are based on the continuous-time Markov chain (CTMC) framework, either directly or through simplified variants inspired by it. In this framework, the forward noising process is governed by a transition rate matrix. The reverse process is parameterized by a learned approximation of the *discrete score*, defined as the ratios of forward marginal probabilities. This quantity serves as the discrete analog of the Stein score in continuous diffusion (Campbell et al., 2022; Meng et al., 2022; Lou et al., 2024). Two choices of rate matrix dominate practice: the *uniform* rate matrix and the *absorbing* (or masked) rate matrix (Austin et al., 2021; Lou et al., 2024; Shi et al., 2024). The latter, in particular, has emerged as the dominant paradigm for large-scale discrete generative modeling, especially in language tasks (Sahoo et al., 2024; Nie et al., 2026).

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

Despite this empirical progress, existing convergence analyses for score-based discrete diffusion suffer from fundamental limitations. These include divergent bounds for singular priors such as masked diffusion, vacuous scaling with vocabulary size in large-vocabulary regimes, restriction to specific rate matrices or evaluation metrics, and reliance on strong regularity assumptions. In this work, we develop a unified analysis framework based on adjoint equations that simultaneously addresses all of these limitations.

Table 1: Comparison of results for discrete diffusion models.

Reference	Rates/Prior	S-free[1]	Key Conditions	Metrics/ Divergences
This work; Thm 3.1, Lem 3.2 & Cors 3.3-3.4	**General, including Uniform & Masked**	**Yes**[2]	**One standard rate-matrix condition**	**Any IPM**
Zhang et al. (2025); Thm 1	Uniform	No	Fully-supported prior	KL, TV
Liang et al. (2025); Thm 2	Surrogate Masked	No[3]	Fully-supported prior	KL
Ren et al. (2025); Thm 4.7	Uniform[4]	No	Fully-supported prior	KL
Liang et al. (2026b); Thm 2	Uniform	No	Fully-supported prior	KL
Dmitriev et al. (2026); Thms 1& 3	Uniform & Surrogate Masked	No	Fully-supported prior	KL
Liang et al. (2026d); Thm 2	Masked	No[5]	Differentiating TV metric	TV

1: S: vocabulary size. In modern language tasks, $S > 10^5$ (Nie et al., 2026) renders S-dependent bounds loose/vacuous.
2: Our convergence bounds for uniform and masked rates are S-independent.
3: Liang et al. (2025)'s Theorem 1 omitted an S-dependent constant.
4: Ren et al. (2025)'s Theorem C.1 assumes a uniform prior.
5: Liang et al. (2026d)'s Proposition 1 contains an S-dependent coefficient.

We summarize our contributions as follows.

- We establish the first convergence bounds that are **entirely independent of the vocabulary size** S and applicable to non-fully-supported priors such as the singular masked distribution. These properties are jointly essential for language modeling, the most important application of discrete diffusion, where vocabulary sizes are prohibitively large and masked diffusion is the dominant paradigm.
- Our framework yields a single unified derivation that **encompasses all integral probability metrics (IPMs) and all priors**. In contrast, existing analyses are typically restricted to one or two specific metrics or priors. Moreover, our results are based on a single standard rate-matrix regularity assumption. See Table 1 for detailed comparisons. In turn, our results can **extend existing step complexity guarantees to any IPM**, since their analyses can be directly applied to our bounds.
- Finally, our theoretical framework extends beyond convergence bounds. It establishes a versatile mathematical toolkit that paves the way for future, broader theoretical analyses of discrete diffusion models, including principled choices of loss functions for different dynamics and sharper (S-free) step complexity bounds.

These contributions are driven by the following main technical novelties.

- We work with the **adjoint equation** (specifically, the Kolmogorov backward equation) in the **space of observables**, in contrast to prior analyses that operate directly in the space of probability measures. This dual perspective sharply departs from existing pathspace-KL and TV-based estimates. Together with a regularity analysis of the adjoint equation, it is the key enabler of the insensitivity to non-fully-supported priors and the applicability to all IPMs. To the best of our knowledge, this is **the first such application** to discrete diffusion.
- We introduce **novel score-marginal cancellation and exit-rate routing techniques** that removes the S-dependence of convergence bounds for masked transitions. To the best of our knowledge, this is **the first S-free bound for masked transitions.**

- We apply a **coupling argument** that bounds the initialization error under uniform transitions independently of S. To the best of our knowledge, **this is the first such bound**, and the first use of coupling techniques in discrete diffusion.

2 Problem Formulation

We state the setup used throughout; a more detailed description is given in Appendix B.

Notation. Scalars are lowercase (x) and vectors boldface lowercase ($\mathbf{x}$); all vectors are columns. We consider data on the discrete state space $\mathcal{X} = [S]^d$, where $[S] = \{1, \dots, S\}$ and $\mathbf{x} = [x^1 \cdots x^d]$ has entries $x^i \in [S]$. This setup is common to text (Lou et al., 2024), image pixels (Campbell et al., 2022), tabular data (Kotelnikov et al., 2023), molecular graphs (Vignac et al., 2023), musical notes (Campbell et al., 2022), and biological sequences (Schiff et al., 2025; Campbell et al., 2024), among others.

We write $\mathbf{x}_{i \to \hat{x}^i}$ for $\mathbf{x}$ with its i-th entry replaced by $\hat{x}^i$, and $\mathbf{x}^{\setminus i}$ for $\mathbf{x}$ with that entry removed. The Hamming distance $\mathrm{Ham}(\mathbf{x}, \mathbf{y}) = |\{i \in [d] : x^i \neq y^i\}|$ counts differing entries, where $|\cdot|$ is set cardinality. We write $\mathbf{e}_i$ for the i-th standard basis vector, $\mathbf{1}_n$ for the all-ones vector, I_n for the identity, and $\delta\{\cdot, \cdot\}$ for the Kronecker delta. A distribution on $\mathcal{X}$ is a probability mass vector $p \in \Delta(\mathcal{X})$, the simplex over the S^d states.

Forward Noising Process. The forward noising process $X = \{X_t\}_{0 \le t \le T}$ is a continuous-time Markov chain (CTMC) on $\mathcal{X}$ whose marginal $p_t \in \Delta(\mathcal{X})$ is governed by the *Kolmogorov forward equation* (KFE), a linear ordinary differential equation (ODE) given by

$$\frac{dp_t}{dt} = Q_t^\top p_t, \quad p_0 = p_{\mathrm{data}}, \quad \text{for } t \in [0, T], \tag{1}$$

with rate matrix (infinitesimal generator) $Q_t \in \mathbb{R}^{S^d \times S^d}$, where $Q_t(\mathbf{x}, \mathbf{y})$ is the infinitesimal rate of transitioning from states $\mathbf{x}$ to $\mathbf{y}$. Because Q_t is constrained to have nonnegative off-diagonal entries and zero row sums, the dynamics (1) preserve the total mass of p_t, ensuring it remains a valid probability distribution throughout the evolution.

Because the dimensions of Q_t scale exponentially with the sequence length d, direct computation or storage of the generator is intractable. In practice, Q_t is structured to operate on each dimension independently via a token-level rate matrix $Q_t^{\mathrm{tok}} \in \mathbb{R}^{S \times S}$. In standard frameworks (Campbell et al., 2022; Lou et al., 2024; Zhang et al., 2025), the nonzero off-diagonal entries of Q_t are structured as

$$Q_t(\mathbf{x}, \mathbf{x}_{i \to \hat{x}^i}) = Q_t\left([x^1, \cdots, x^i, \cdots, x^d], [x^1, \cdots, \hat{x}^i, \cdots, x^d]\right) = Q_t^{\mathrm{tok}}(x^i, \hat{x}^i), \tag{2}$$

where $x^i \neq \hat{x}^i \in [S]$. This structure implies that Q_t is a Kronecker sum of independent generators, connecting only sequences at Hamming distance 1 (our results also apply to general rate matrices; see Section 3). The two common choices for Q_t^{tok} are the masked (absorbing) and uniform rates,

$$(Q_t^{\mathrm{masked}})^\top = \beta(t)(\mathbf{e}_{[\mathrm{MASK}]}\mathbf{1}_S^\top - I_S) \quad \text{and} \quad (Q_t^{\mathrm{uniform}})^\top = \beta(t)\left(\tfrac{1}{S}\mathbf{1}_S\mathbf{1}_S^\top - I_S\right), \tag{3}$$

respectively. For the masked rate, the S-th token is the mask [MASK]. Consequently, p_t approaches a tractable limiting distribution p_{base} as $t \to \infty$: either a Dirac distribution centered on the all-mask sequence or the uniform distribution on $\mathcal{X}$. The ease of sampling from these distributions makes them ideal priors for the reverse generative process.

Under (2), the forward transition kernel $p_{t|s}(\mathbf{x}_t|\mathbf{x}_s) := \Pr(X_t = \mathbf{x}_t | X_s = \mathbf{x}_s)$ factorizes as $p_{t|s}(\mathbf{x}_t|\mathbf{x}_s) = \prod_{i=1}^d p^i_{t|s}(x^i_t|x^i_s)$ for $0 \le s < t \le T$, where $p^i_{t|s}$ is the forward transition kernel of the i-th token. This factorization is standard both for scaling to high-dimensional state spaces in practice (Austin et al., 2021; Campbell et al., 2022; 2024; Lou et al., 2024) and in theoretical analysis (Zhao et al., 2025; Chen & Ying, 2025; Zhang et al., 2025; Ren et al., 2025).

Reverse Denoising Process. The forward process (1) admits a reverse-time process, which is also a CTMC (Kelly, 1979). The reverse process satisfies the reverse-time Kolmogorov equation

$$-\frac{dp_t}{dt} = (Q_t^{\leftarrow})^\top p_t, \quad \text{for } t \in [0, T], \tag{4}$$

where the reverse rate matrix $Q_t^{\leftarrow}$ is given by

$$Q_t^{\leftarrow}(\mathbf{x}, \mathbf{y}) = \frac{p_t(\mathbf{y})}{p_t(\mathbf{x})} Q_t(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{d} \frac{p_t(\mathbf{y})}{p_t(\mathbf{x})} \delta\{\mathbf{x}^{\setminus i}, \mathbf{y}^{\setminus i}\} Q_t^{\text{tok}}(y^i, x^i) \quad \text{for } \mathbf{x} \neq \mathbf{y}, \tag{5}$$

and $Q_t^{\leftarrow}(\mathbf{x}, \mathbf{x}) = -\sum_{\mathbf{y} \neq \mathbf{x}} Q_t^{\leftarrow}(\mathbf{x}, \mathbf{y})$. Here, the second equality uses (2). The Kronecker delta ensures that at most one term in the summation survives, specifically, the term where $\mathbf{x}$ and $\mathbf{y}$ differ by exactly one token. Thus, $Q_t^{\leftarrow}$ inherits the same sparse structure of the forward rate Q_t, restricting nonzero transition rates to sequences at Hamming distance 1. For the reverse process under a general rate, see Appendix B.

Training Formulation. The reverse process therefore depends on p_t only through the ratios

$$s_t(\mathbf{x})_{\mathbf{y}} := \frac{p_t(\mathbf{y})}{p_t(\mathbf{x})}, \quad \mathbf{x} \neq \mathbf{y} \in \mathcal{X}, \tag{6}$$

known as the discrete/concrete score (Meng et al., 2022), which is analogous to the Stein score $\nabla_x \log p_t$ in continuous diffusion models (Song et al., 2021b). Under (2) only Hamming-distance-1 pairs carry nonzero rate, so the reverse dynamics involve $s_t(\mathbf{x})_{\mathbf{y}}$ only at $\mathbf{y} = \mathbf{x}_{i \to \hat{x}^i}$. In score-based discrete diffusion, the goal is to learn a score estimator $\tilde{s} : [0, T] \times \mathcal{X} \to \mathbb{R}^{d(S-1)}$ approximating (6). To this end, we consider two standard training losses: the weighted score-matching (WSM) loss (Meng et al., 2022)

$$\mathcal{L}_{\text{WSM}}(s, \tilde{s}) = \int_0^T \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p_t} \sum_{\mathbf{y} \neq \mathbf{x}} w_t(\mathbf{x}, \mathbf{y}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right)^2 dt, \tag{7}$$

and the score entropy (SE) loss (Benton et al., 2024; Lou et al., 2024)

$$\mathcal{L}_{\text{SE}}(s, \tilde{s}) = \int_0^T \mathbb{E}_{\mathbf{x} \sim p_t} \sum_{\mathbf{y} \neq \mathbf{x}} w_t(\mathbf{x}, \mathbf{y}) \left(s_t(\mathbf{x})_{\mathbf{y}} \log \left(\frac{s_t(\mathbf{x})_{\mathbf{y}}}{\tilde{s}_t(\mathbf{x})_{\mathbf{y}}} \right) + \tilde{s}_t(\mathbf{x})_{\mathbf{y}} - s_t(\mathbf{x})_{\mathbf{y}} \right) dt, \tag{8}$$

where $w_t(\mathbf{x}, \mathbf{y}) \geq 0$ are weighting coefficients, and $\mathcal{L}_{\text{SE}}$ is the Bregman divergence generated by $a \mapsto a \log a - a$. Following Lou et al. (2024), we use the transition weight $w_t(\mathbf{x}, \mathbf{y}) = Q_t(\mathbf{y}, \mathbf{x})$. This choice parallels the score-matching loss in continuous space, where it is typically weighted by the diffusion coefficient to align with the objective with the evidence lower bound (ELBO) (Song et al., 2021b). Both objectives are intractable, as the ground-truth score depends on the unknown marginal p_t. In practice, one uses denoising or implicit score matching, which are tractable and equivalent to the original loss up to a constant (Lou et al., 2024).

Approximate Reverse Process. The generative process is governed by the approximate reverse-time dynamics

$$-\frac{d\tilde{p}_t}{dt} = (\tilde{Q}_t^{\leftarrow})^{\top} \tilde{p}_t, \quad \tilde{p}_T = p_{\text{base}} \approx p_T, \quad \text{for } t \in [0, T], \tag{9}$$

where $\tilde{p}$ denotes the marginal of the approximate reverse process and

$$\tilde{Q}_t^{\leftarrow}(\mathbf{x}, \mathbf{y}) = \tilde{s}_t(\mathbf{x})_{\mathbf{y}} Q_t(\mathbf{y}, \mathbf{x}) \quad \text{for } \mathbf{x} \neq \mathbf{y}, \quad \text{and} \quad \tilde{Q}_t^{\leftarrow}(\mathbf{x}, \mathbf{x}) = -\sum_{\mathbf{y} \neq \mathbf{x}} \tilde{Q}_t^{\leftarrow}(\mathbf{x}, \mathbf{y}). \tag{10}$$

This differs from the exact reverse process (4) in two ways: the true rate $Q_t^{\leftarrow}$ is replaced by $\tilde{Q}_t^{\leftarrow}$, constructed using the score estimator, and the intractable p_T is replaced by the prior p_{base}.

Integral Probability Metric (IPM). IPMs are a general class of metrics for measuring the discrepancy between two distributions $p, \tilde{p} \in \Delta(\mathcal{X})$. They are defined as

$$\gamma_{\Psi}(p, \tilde{p}) = \sup_{\psi \in \Psi} \left| \mathbb{E}_{\mathbf{x} \sim p}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \tilde{p}}[\psi(\mathbf{x})] \right|, \tag{11}$$

where the admissible function class Ψ determines the metric. Taking Ψ to be $\{\psi : \mathcal{X} \to \mathbb{R}, \|\psi\|_{\infty} \leq 1\}$, $\{\|\psi\|_{\text{Lip}} \leq 1\}$, or $\{\|\psi\|_{\mathcal{H}} \leq 1\}$ gives the total variation distance, the Wasserstein-1 distance, and the maximum mean discrepancy, respectively, where $\|\cdot\|_{\mathcal{H}}$ is the norm of a reproducing kernel Hilbert space $\mathcal{H}$. We refer readers to Sriperumbudur et al. (2009); Peyré & Cuturi (2019); Birrell et al. (2022) for a more detailed overview of IPMs and their properties.

Error and Adjoint Equations. Our analysis relies on computing the mismatch between the exact generative process (4) and its approximation (9) over time. Defining the error $\lambda_t := p_t - \tilde{p}_t$ and subtracting (9) from (4), we obtain the *error equation* for $t \in [0, T]$,

$$\frac{d\lambda_t}{dt} = (-Q_t^{\leftarrow} + \tilde{Q}_t^{\leftarrow})^\top p_t - (\tilde{Q}_t^{\leftarrow})^\top \lambda_t, \quad \text{with} \quad \lambda_T = p_T - p_{\text{base}}. \tag{12}$$

To quantify this discrepancy under any IPM, we extend the duality argument of Mimikos-Stamatopoulos et al. (2024, Section 4.1), originally developed for continuous diffusion, by developing new technical tools specialized for the discrete setting. The core idea is to test (12) against a function $\phi : [0, T] \times \mathcal{X} \to \mathbb{R}$ satisfying the *adjoint equation* associated with the approximate reverse dynamics (9), namely the *Kolmogorov backward equation*

$$\frac{d\phi_t}{dt} = \tilde{Q}_t^{\leftarrow} \phi_t, \quad \phi_0(\mathbf{x}) = \psi(\mathbf{x}), \quad \text{for } t \in [0, T]. \tag{13}$$

By selecting ψ from the admissible function class Ψ of the target metric (11), we can leverage (13) to explicitly compute, in any IPM, how the approximation error propagates back to the initial data distribution.

3 Main Results

Recall that any valid rate matrix of a CTMC satisfies two structural properties by definition: *non-negative off-diagonal entries*, $Q_t(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x} \neq \mathbf{y}$, and *zero row sums*, $\sum_{\mathbf{y}} Q_t(\mathbf{x}, \mathbf{y}) = 0$ for all $\mathbf{x}$, which together conserve probability mass. Beyond these structural properties, we impose only the following mild regularity condition to keep our framework as general as possible.

Assumption 1 (Rate-matrix Integrability)**.** *$t \mapsto Q_t(\mathbf{x}, \mathbf{y})$ is integrable on $[0, T]$ for all $(\mathbf{x}, \mathbf{y})$.*

Assumption 1 ensures that the transition kernel associated with Q_t is well-defined on $[0, T]$. Notably, both the structural properties and Assumption 1 are satisfied by virtually all discrete-diffusion-based generative models used in practice, underscoring the broad applicability of our theory.

3.1 Main Theorem

Under the mild Assumption 1, our main theorem bounds the discrepancy between the data distribution and the generated distribution, simultaneously for all IPMs (11). Due to space constraints, we defer all proofs to the Appendix.

Theorem 3.1 (IPM Bounds for Score-based Discrete Diffusion)**.** *Under Assumption 1, for any function class $\Psi = \{\psi : \mathcal{X} \to \mathbb{R}\}$, the IPM γ_Ψ (defined in* (11)*) between the data distribution p_{data} and the generated distribution $\tilde{p}_0$ satisfies the following. For bounds involving $\mathcal{L}_{\text{SE}}$, we additionally assume $\tilde{s}_t(\mathbf{x})_{\mathbf{y}} \in [s_t(\mathbf{x})_{\mathbf{y}}/2,\, 3s_t(\mathbf{x})_{\mathbf{y}}/2]$ for all $(t, \mathbf{x}, \mathbf{y})$ in the support of $p_t Q_t$, a condition that holds when the score estimator is reasonably accurate.*
Case 1 (Masked rate; $Q_t^{\text{tok}} = Q_t^{\text{masked}}$ (3)):

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \leq 2C_\Psi \left(\underbrace{d e^{-\|\beta\|_1}}_{\textit{initialization error}} + \sqrt{2d} \sqrt{\|\beta(1 - p(\mathbf{m}))\|_1} \underbrace{\sqrt{\mathcal{L}_{\text{SE}}(s, \tilde{s}) + \tfrac{2}{3}\, \mathcal{L}_3(s, \tilde{s})}}_{\textit{training error}} \right), \tag{14}$$

where d is the sequence length, $\|\beta\|_1 := \int_0^T \beta(t)dt$, $\mathbf{m} :=$ [MASK], ..., [MASK] is the sequence of all mask tokens, $\|\beta(1 - p(\mathbf{m}))\|_1 := \int_0^T \beta(t)(1 - p_t(\mathbf{m}))dt$, $p_t(\mathbf{m})$ is the marginal probability of $\mathbf{m}$ at time t, $\mathcal{L}_{\text{SE}}$ is the score entropy loss (8)*, and $\mathcal{L}_3 = \mathcal{O}(|s - \tilde{s}|^3)$ is a cubic correction term with scaling **independent of the vocabulary size** S; its exact formulation is provided in Appendix D.1;*
Case 2 (Uniform rate; $Q_t^{\text{tok}} = Q_t^{\text{uniform}}$ (3)):

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \leq 2C_\Psi \left(\underbrace{d e^{-\|\beta\|_1}}_{\textit{initialization error}} + \sqrt{2d} \sqrt{\|\beta\|_1} \underbrace{\sqrt{\mathcal{L}_{\text{SE}}(s, \tilde{s}) + \tfrac{2}{3}\, \mathcal{L}_3(s, \tilde{s})}}_{\textit{training error}} \right), \textit{ and} \tag{15}$$

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \le 2C_\Psi \left(\underbrace{d e^{-\|\beta\|_1}}_{\textit{initialization error}} + \sqrt{2d}\sqrt{\|\beta\|_1} \underbrace{\sqrt{\mathcal{L}_{\text{WSM}}(s, \tilde{s})}}_{\textit{training error}} \right), \tag{16}$$

where $\mathcal{L}_{\text{WSM}}$ is the weighted score-matching loss defined in (7)*;*
Case 3 (General rate):

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \le 2C_\Psi \left(\underbrace{\text{TV}(p_T, p_{\text{base}})}_{\textit{initialization error}} + \sqrt{2}\sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}[\mu_t^{\text{out}}(\mathbf{x})]\, dt} \underbrace{\sqrt{\mathcal{L}_{\text{SE}}(s, \tilde{s}) + \tfrac{2}{3}\mathcal{L}_3(s, \tilde{s})}}_{\textit{training error}} \right), \tag{17}$$

and

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \le 2C_\Psi \left(\underbrace{\text{TV}(p_T, p_{\text{base}})}_{\textit{initialization error}} + \sqrt{2}\sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\text{in}}(\mathbf{x})\right] dt} \underbrace{\sqrt{\mathcal{L}_{\text{WSM}}(s, \tilde{s})}}_{\textit{training error}} \right), \tag{18}$$

where $\mu_t^{\text{out}}(\mathbf{x}) := \sum_{\mathbf{y}\neq\mathbf{x}} Q_t(\mathbf{x}, \mathbf{y})$ and $\mu_t^{\text{in}}(\mathbf{x}) := \sum_{\mathbf{y}\neq\mathbf{x}} Q_t(\mathbf{y}, \mathbf{x})$ denote the exit rate from and entrance rate to state $\mathbf{x}$ at time t, respectively. The dependence on the IPM enters only through the single constant C_Ψ. The list of common IPMs and their constant C_Ψ are summarized in Table 2; the result extends straightforwardly to all other IPMs.

Table 2: Examples of common IPMs, their corresponding function class, and the associated constants for Theorem 3.1, Corollary 3.3, and Corollary 3.4. Here, W_1 and MMD represent the Wasserstein-1 distance and maximum mean discrepancy, respectively. **Our results extend straightforwardly and universally to every IPM.**

IPM	Function class Ψ	C_Ψ
Total variation (TV)	$\|\psi\|_\infty \le 1$	$1/2$
Per-position TV	$\psi(\mathbf{x}) = f(x^i),\ i \in [d],\ \|f\|_\infty \le 1$	$1/2$
Block / k-gram TV	ψ depends on k positions, $\|\psi\|_\infty \le 1$	$1/2$
Bounded Lipschitz	$\|\psi\|_\infty \le 1,\ \text{Lip}_d(\psi) \le 1$	1
$W_1^{d_H}$ (Hamming distance)	$\text{Lip}_{d_H}(\psi) \le 1$	$d/2$
$W_1^{d_{\text{embed}}}$	$\text{Lip}_{d_{\text{embed}}}(\psi) \le 1$	$\text{diam}_{d_{\text{embed}}}(\mathcal{X})/2$
MMD (kernel $K \le 1$)	$\|\psi\|_{\mathcal{H}_K} \le 1$	1
MMD (general bounded K)	$\|\psi\|_{\mathcal{H}_K} \le 1$	$\sup_{\mathbf{x}} \sqrt{K(\mathbf{x}, \mathbf{x})}$

A key consequence of Theorem 3.1 is that the approximation error, measured by any IPM (11), is bounded by score-matching objectives. Specifically, the training objectives ($\mathcal{L}_{\text{WSM}}$ and $\mathcal{L}_{\text{SE}}$) and the faster-vanishing cubic correction ($\mathcal{L}_3$) all approach zero as the score estimator converges to the true score ($\tilde{s} \to s$). This provides a rigorous theoretical justification for the standard optimization objective: **Model training directly and explicitly minimizes an upper bound on the entire family of IPMs between the generated and the target distributions.** Consequently, a globally optimal model achieves the theoretical minimum of the distributional discrepancy.

Remark 3.1 (Key Improvements)**.** *Our results improve existing theory in the following key aspects (see also Table 1).*

1. ***Independence of Vocabulary Size S:*** *The coefficients of our bounds are independent of the vocabulary size S under masked and uniform rates. This is a significant improvement over existing literature, where bounds typically depend on S and become loose or even vacuous in large-scale applications like language modeling (e.g., $S > 10^5$ in SOTA architectures (Nie et al., 2026)). Concretely, the TV bounds of Liang et al. (2026d;a) carry a prefactor of $\sqrt{S}$, which exceeds 300 when $S > 10^5$ and thus inflates the training-error term by two orders of magnitude. In contrast, our S-independence is a direct consequence of our novel derivations, which circumvent the S-dependency inherent in prior approaches.*

Our key techniques include: adjoint equations (see (13) *and Appendix D.1), together with a regularity analysis (see Appendix E.1), that yield bounds for any IPM; a score–marginal cancellation technique and an exit-routing technique that achieve S-independence under masked transitions (see Remark 3.2 and Appendix H); and a coupling argument that achieves S-independence under uniform transitions (see Remark 3.4 and Appendix F).*

2. ***Unified Derivations and Universal IPM Coverage:*** *Our novel applications of adjoint equations and regularity theory allow us to derive bounds for all IPMs under a unified framework. This stands in stark contrast to existing analyses, which typically focus on only one or two specific metrics; see Table 1.*

3. ***Direct Extension of Existing Guarantees to Any IPM:*** *Our convergence results bound any IPM in terms of the score entropy loss* (8)*. Existing step complexity analyses bound the score entropy loss* (8) *under numerical simulation, and in turn use the bound to derive step complexity required for the approximation error to fall below a given threshold. Since our results are in terms of the same score entropy, existing step complexity analyses can be directly applied to our bounds, extending their guarantees to any IPM.*

4. ***Agnostic to Rates and Priors:*** *Our framework is agnostic to the choice of rate matrix or prior distribution; it applies straightforwardly to any valid transition process satisfying Assumption 1. This sharply departs from existing works, which are inapplicable to masked transition (Zhang et al., 2025; Liang et al., 2025; Ren et al., 2025; Liang et al., 2026b; Dmitriev et al., 2026) or focus on one transition process (Liang et al., 2026d).*

5. ***Mild Assumptions:*** *Our analysis relies solely on Assumption 1. This is in stark contrast to existing theory; for instance, the works of Zhang et al. (2025); Liang et al. (2025); Ren et al. (2025); Liang et al. (2026b); Dmitriev et al. (2026) rely critically on the non-singularity of p_{base}, without which their estimates diverge—a constraint we do not impose.*

Remark 3.2 (Score-Marginal Cancellation and Exit-Routing)**.** *At its core, our* ***score-marginal cancellation technique*** *uses a simple identity $p_t(\mathbf{x})s_t(\mathbf{x})_{\mathbf{y}} = \cancel{p_t(\mathbf{x})}\frac{p_t(\mathbf{y})}{\cancel{p_t(\mathbf{x})}} = p_t(\mathbf{y})$. Via this identity and algebraic manipulation, we can express the bound in Case 3 of Theorem 3.1 using either the entrance rate μ_t^{in}* (18) *or the exit rate μ_t^{out}* (17)*. For more details, see Appendices H.1 and H.2.*

Cases 1 and 2 can both be derived as specializations of Case 3. This allows users to pick the option that gives a tighter bound for a rate matrix at hand. For instance, under the masked rate, the scaling of $\mathbb{E}[\mu_t^{\text{out}}]$ is independent of S whereas $\mathbb{E}[\mu_t^{\text{in}}]$ grows as $\mathcal{O}(S)$. ***Thus, using the exit rate bound in Case 1 yields an S-free guarantee.*** *See Remark D.2 for a more detailed discussion.*

Remark 3.3 (Theoretical Justification for Score Entropy)**.** *From Remark 3.2, the convergence bound under the masked rate is S-free when the training loss $\mathcal{L}_{\text{SE}}$ is used, but grows as $\mathcal{O}(S)$ when guided by $\mathcal{L}_{\text{WSM}}$. Consequently, the $\mathcal{L}_{\text{SE}}$-based bound is substantially tighter for language tasks, where the vocabulary size S is prohibitively large. This offers a principled explanation for the strong empirical performance of models trained using the score entropy loss, complementing existing theory on the score entropy loss Lou et al. (2024); Shi et al. (2024); Ou et al. (2025).*

Remark 3.4 (Coupling Technique for Uniform Rate)**.** *To the best of our knowledge, existing results for the uniform rate (Zhang et al., 2025; Ren et al., 2025; Liang et al., 2026b; Dmitriev et al., 2026) bound the initialization error under the Kullback-Leibler (KL) divergence by using a modified log-Sobolev inequality to derive the exponential mixing of the forward process; this introduces a $\log(S)$ factor. Converting to TV distance (a type of IPM* (11)*) via Pinsker's inequality then yields a $\sqrt{\log(S)}$ factor; see, e.g., the proof of Proposition 2 of Zhang et al. (2025).*

In contrast, our initialization error is S-free. This is achieved by using a novel coupling argument adapted from the classical Markovian coupling theorem (Levin & Peres, 2017, Theorem 5.2). See Section F for a detailed derivation.

3.2 Weak Error Bound

Our main Theorem 3.1 is a specialization of the following lemma, which controls the error against a single fixed test function ψ. For compactness, from here on we state the bounds in a single master form and collect the coefficients in tables.

Lemma 3.2 (Weak Error Bound for Score-based Discrete Diffusion)**.** *Under the same assumptions as Theorem 3.1, the weak error between the true data distribution p_{data} and the generated data distribution $\tilde{p}_0$ satisfies the following.*

$$|\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \;\le\; 2\,\|\psi\|_\infty \Big(\underbrace{\mathcal{I}}_{\textit{init. error}} + \sqrt{2\,\mathcal{C}}\;\underbrace{\sqrt{\mathcal{L}}}_{\textit{train. error}} \Big). \tag{19}$$

Here, $(\mathcal{I}, \mathcal{C}, \mathcal{L})$ are reported in Table 3.

Table 3: Coefficients of the convergence bounds for Lemma 3.2 and Corollary 3.3

Rate	Initialization Error $\mathcal{I}$	Coefficient $\mathcal{C}$	Training Error $\mathcal{L}$
Masked	$d\,e^{-\|\beta\|_1}$	$d\,\|\beta(1-p(\mathbf{m}))\|_1$	$\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \frac{2}{3}\mathcal{L}_3(s,\tilde{s})$
Uniform	$d\,e^{-\|\beta\|_1}$	$d\,\|\beta\|_1$	$\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \frac{2}{3}\mathcal{L}_3(s,\tilde{s})$
Uniform	$d\,e^{-\|\beta\|_1}$	$d\,\|\beta\|_1$	$\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})$
General	$\mathrm{TV}(p_T, p_{\mathrm{base}})$	$\int_0^T \mathbb{E}_{p_t}[\mu_t^{\mathrm{out}}]\,dt$	$\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \frac{2}{3}\mathcal{L}_3(s,\tilde{s})$
General	$\mathrm{TV}(p_T, p_{\mathrm{base}})$	$\int_0^T \mathbb{E}_{p_t}[\mu_t^{\mathrm{in}}]\,dt$	$\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})$

Two features of (19) drive our results. First, it holds for each bounded test function ψ. The derivation considers a fixed ψ, propagates it through the adjoint equation (13), and never references the class Ψ. Second, the test function enters the right-hand side only through the scalar $\|\psi\|_\infty$; every other term is independent of it. This is a direct consequence of our regularity analysis (see Lemma E.1), which explicitly bounds the evolution of ψ under the adjoint equation. Thus, taking the supremum over an admissible function class Ψ in (19) yields Theorem 3.1, where C_Ψ is $\sup_{\psi\in\Psi}\|\psi\|_\infty$ up to a scalar multiple.

3.3 Error Under Early Stopping

In early stopping, the reverse denoising process is simulated until an early stopping time $\delta > 0$ rather than $t = 0$. This is necessary when p_{data} does not have full support on $\mathcal{X}$. In this case, $p_t(\mathbf{x})$ can be close to zero at $t \approx 0$ for some $\mathbf{x}$. This causes the score function (6) to be unbounded, resulting in numerical instability during generation and training.

The cost of early stopping is an additional approximation error, because the output distribution approximates p_δ instead of the data distribution $p_0 = p_{\mathrm{data}}$. In the following, we compute the error caused by early stopping.

Corollary 3.3 (Error Bounds under Early Stopping)**.** *Under the same assumptions as Theorem 3.1, for any function class $\Psi = \{\psi : \mathcal{X} \to \mathbb{R}\}$, the IPM γ_Ψ (defined in* (11)*) between the data distribution p_{data} and the generated distribution $\tilde{p}_\delta$, obtained under early-stopping time $\delta > 0$, satisfies*

$$\gamma_\Psi(p_{\mathrm{data}}, \tilde{p}_\delta) \;\le\; 2\,C_\Psi \Big(\underbrace{\mathcal{E}_\delta}_{\textit{early-stopping error}} + \underbrace{\mathcal{I}}_{\textit{init. error}} + \sqrt{2\,\mathcal{C}}\;\underbrace{\sqrt{\mathcal{L}_\delta}}_{\textit{early-stopping train. error}} \Big). \tag{20}$$

Here, $(\mathcal{I}, \mathcal{C})$ are reported in Table 3. The early stopping error $\mathcal{E}_\delta$ is given by $d\int_0^\delta \beta(t)dt$ for the masked and uniform rates and $\mathrm{TV}(p_0, p_\delta)$ for a general rate. The early-stopping training error $\mathcal{L}_\delta$ is also reported in Table 3, except that the time integration of the training losses start at time δ rather than 0. These δ-truncated losses are used in practice to avoid numerical instability. See Table 2 for a common list of IPM and their constant C_Ψ. The results extend to all other IPMs.

3.4 Step Complexity for Uniformization

We derive the expected number of steps under uniformization (Jensen, 1953; Grassmann, 1977; de Souza e Silva & Gail, 2000); see (Chen & Ying, 2025, Section 2) for an introduction of the uniformization algorithm. Since early stopping is the practically relevant setting, we build on the early-stopping bounds from Corollary 3.3.

Corollary 3.4 (Expected Step Complexities for Uniformization)**.** *Assume that* $\tilde{s}_t(\mathbf{x})_{\mathbf{y}} \asymp s_t(\mathbf{x})_{\mathbf{y}}$ *whenever* $Q_t(\mathbf{y}, \mathbf{x}) > 0$*, and the uniformization Poisson rate is chosen as* $\Lambda_k := C \sup_{\substack{\mathbf{x}\in\mathcal{X} \\ t\in(t_k,t_{k+1}]}} -\tilde{Q}_t^{\leftarrow}(\mathbf{x}, \mathbf{x})$ *for some constant* $C \geq 1$*. Then, for any IPM* γ_Ψ *and tolerance* $\epsilon > 0$*, the approximation error satisfies* $\gamma_\Psi(p_{\mathrm{data}}, \tilde{p}_\delta) < \epsilon$ *if we choose* $\delta \asymp \frac{\epsilon}{dC_\Psi \beta_{\max}}$, $T \asymp \frac{1}{\beta_{\min}} \log\left(\frac{dC_\Psi}{\epsilon}\right)$*, and* $\mathcal{L}_{\mathrm{SE}}^{\delta}(s, \tilde{s}) = \mathcal{O}(\frac{\epsilon^2}{dC_\Psi^2})$*, in which case the expected number of steps are given by*

1. ***Case 1 (Masked rate;*** $Q_t^{\mathrm{tok}} = Q_t^{\mathrm{masked}}$ **(3))*:***
$$\mathbb{E}[N] = \mathcal{O}\left(d\,\frac{\beta_{\max}}{\beta_{\min}}\,\log\left(\frac{\beta_{\max}}{\beta_{\min}}\cdot\frac{dC_\Psi}{\epsilon}\right)\right), \quad \textit{and,}$$

2. ***Case 2 (Uniform rate;*** $Q_t^{\mathrm{tok}} = Q_t^{\mathrm{uniform}}$ **(3))*:*** *if the true score further satisfies* $s_t(\mathbf{x})_{\mathbf{y}} \lesssim \max(1, t^{-1})$ *(following Ren et al., 2025, Assumption 4.4), then*
$$\mathbb{E}[N] = \mathcal{O}\left(\beta_{\max}\, d\,\left(\log\beta_{\max} + \frac{1}{\beta_{\min}}\log\frac{dC_\Psi}{\epsilon}\right)\right).$$

Here, $\beta_{\max} = \max_{0\leq t\leq T}\beta(t)$ *and* $\beta_{\min} = \min_{0\leq t\leq T}\beta(t)$*, and the values of* C_Ψ *are given in Table 2.*

Here, the choice of the Poisson rate is a standard assumption required to ensure the validity of the uniformization method for the CTMC. See, e.g., Grassmann (1977); Chen & Ying (2025).

Remark 3.5 (S-free Training Loss Requirement)**.** *In Corollary 3.4, our S-free convergence bounds yield a distinct advantage on training loss requirement over existing results. In existing works (Liang et al., 2025; 2026a)*[1]*, the S-dependent prefactor of the training error term means the training error needs to be $\mathcal{O}(1/S)$. This imposes an often impractical requirement for large token vocabularies in modern language tasks. Since our training error prefactor is S-free, the required bound $\mathcal{L}_{\mathrm{SE}}^{\delta}(s, \tilde{s}) = \mathcal{O}(\frac{\epsilon^2}{dC_\Psi^2})$ does not tighten as S grows.*

As also noted in Remark 3.1.3, existing step complexity analyses bound the score entropy loss under numerical simulation. The same loss controls the approximation error in Corollary 3.3. Thus, **existing step complexity guarantees can be extended to any IPM** by directly applying existing derivations to our bounds in Corollary 3.3.

In Corollary 3.4, we adapt the step complexity derivations of Liang et al. (2025); Ren et al. (2025) for uniformization to our setting. Their derivation is originally for a time-homogeneous forward process, and we modify it to handle our time-inhomogeneous forward process (3). See Appendix I for the detailed derivations.

Similarly, derivations for other algorithms, e.g., τ-leaping (Liang et al., 2025; Dmitriev et al., 2026), can also be applied to our framework to obtain step complexity. Since our goal in this paper is to propose a unified and foundational convergence framework for all IPMs, we defer the step complexity analysis for other algorithms to future work.

4 Conclusion

We introduced a novel adjoint-equation-based framework for the convergence analysis of discrete diffusion models. Our analysis yields the first vocabulary-size-independent bounds in any integral probability metric, valid under singular priors and across uniform and masked transitions. The framework applies to general rate matrices. Five novel techniques drive these improvements: use of adjoint equations in the space of observables, a regularity analysis, a coupling argument tailored to uniform transitions, and score-marginal cancellation and exit routing techniques. More broadly, our framework opens several directions for further study of discrete diffusion, including sharper step complexity bounds applicable to any IPM, the choice of loss functions, convergence analyses for higher-order and adaptive samplers, and approximation errors from practical implementation.

[1] In (Liang et al., 2025, Assumption 1 & Proposition 1) and (Liang et al., 2026a, Assumption 1 & Proposition A1), it is required that their $\mathcal{L}_{\mathrm{TV}} = \mathcal{O}(\sqrt{S\mathcal{L}_{\mathrm{SE}}})$ to be bounded by a given threshold. This effectively requires $\mathcal{L}_{\mathrm{SE}}$ to be $\mathcal{O}(1/S)$.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for polishing writing. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refine hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, and interpreting results. And generating synthetic data sets, cleaning and reformatting dataset, supporting qualitative and thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools for drafting parts of a research paper and identifying relevant literature. We have reviewed all AI-assisted work. We have checked for the correctness of the AI writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

The authors believe that this paper submission does not raise questions regarding the Code of Ethics.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This is a theory paper, and all derivations are reported.

REFERENCES

Markus Arns, Peter Buchholz, and Andriy Panchenko. On the numerical analysis of inhomogeneous continuous-time markov chains. *INFORMS Journal on Computing*, 22(3):416–432, 2010.

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.

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.

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f,gamma)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022. URL http://jmlr.org/papers/v23/21-0100.html.

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.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=kQwSbv0BR4.

Yang Cao, Linda R Petzold, Muruhan Rathinam, and Daniel T Gillespie. The numerical stability of leaping methods for stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 121(24):12169–12178, 2004.

Yang Cao, Daniel T Gillespie, and Linda R Petzold. Efficient step size selection for the tau-leaping simulation method. *The Journal of chemical physics*, 124(4), 2006.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

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. doi: 10.4208/jml.240812. URL https://www.global-sci.com/index.php/jml/article/view/13211.

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

Zixiang Chen, Huizhuo Yuan, Yongqian Li, Yiwen Kou, Junkai Zhang, and Quanquan Gu. Fast sampling via discrete non-markov diffusion models with predetermined transition time. *Advances in Neural Information Processing Systems*, 37:106870–106905, 2024.

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

Edmundo de Souza e Silva and H Richard Gail. Transient solutions for markov chains. In *Computational probability*, pp. 43–79. Springer, 2000.

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

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

Winfried K Grassmann. Transient solutions in markovian queueing systems. *Computers & Operations Research*, 4(1):47–53, 1977.

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

Asger Hobolth and Eric A Stone. Simulation from endpoint-conditioned, continuous-time markov chains on a finite state space, with applications to molecular evolution. *The annals of applied statistics*, 3(3):1204, 2009.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in neural information processing systems*, 34:12454–12465, 2021.

Arne Jensen. Markoff chains as an aid in the study of markoff processes. *Scandinavian Actuarial Journal*, 1953(sup1):87–91, 1953.

Frank P Kelly. *Reversibility and stochastic networks*. J. Wiley, 1979.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DjJmre5IkP.

Jaeyeon Kim, Jonathan Geuter, David Alvarez-Melis, Sham Kakade, and Sitan Chen. Stop training for the worst: Progressive unmasking accelerates masked diffusion training. In *ICLR 2026 Workshop on MM Intelligence*, 2026. URL https://openreview.net/forum?id=UIcRg65DF0.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pp. 17564–17579. PMLR, 2023.

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

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Gen Li and Changxiao Cai. Breaking ar's sampling bottleneck: Provable acceleration via diffusion language models. *Advances in Neural Information Processing Systems*, 38:11700–11725, 2026.

Jingyuan Li, Xiaoyi Jiang, Fukang Wen, Wei Liu, Renqian Luo, Yi Zhu, Zuoqiang Shi, and Pipi Hu. Neural continuous-time markov chain: Discrete diffusion via decoupled jump timing and direction. *arXiv preprint arXiv:2604.15694*, 2026.

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*, 38:20283–20318, 2025.

Yuchen Liang, Lifeng Lai, Ness Shroff, and Yingbin Liang. Convergence guarantees for time-inhomogeneous uniform-rate discrete diffusion models. *Entropy*, 28(6):675, 2026a.

Yuchen Liang, Yingbin Liang, Lifeng Lai, and Ness Shroff. Discrete diffusion models: Novel analysis and new sampler guarantees. *Advances in Neural Information Processing Systems*, 38: 165511–165548, 2026b.

Yuchen Liang, Ness Shroff, and Yingbin Liang. From scores to Gibbs correctors: Accelerating uniform-rate discrete diffusion models. *arXiv preprint arXiv:2605.27352*, 2026c.

Yuchen Liang, Zhiheng Tan, Ness Shroff, and Yingbin Liang. Sharp convergence rates for masked diffusion models, 2026d. URL https://arxiv.org/abs/2602.22505.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=CNicRIVIPA.

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.

Nikiforos Mimikos-Stamatopoulos, Benjamin J Zhang, and Markos A Katsoulakis. Score-based generative models are provably robust: an uncertainty quantification perspective. *Advances in Neural Information Processing Systems*, 37:63154–63183, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

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 *International Conference on Learning Representations*, volume 2025, pp. 64972–65009, 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 *International Conference on Learning Representations*, volume 2025, pp. 96272–96300, 2025.

Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

Yinuo Ren, Haoxuan Chen, Grant 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*, volume 2025, pp. 42904–42941, 2025.

Yinuo Ren, Haoxuan Chen, Yuchen Zhu, Wei Guo, Yongxin Chen, Grant Rotskoff, Molei Tao, and Lexing Ying. Fast solvers for discrete diffusion models: Theory and applications of high-order algorithms. *Advances in Neural Information Processing Systems*, 38:167228–167282, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

Yair Schiff, Subham Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 43776–43821, 2025.

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*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. Score-based continuous-time discrete diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=BYWWwSY2G5s.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UaAD-Nu86WX.

Jiawei Xu, Minghui Liu, Aakriti Agrawal, Yifan Chen, and Furong Huang. Scheduling thoughts: Learning the order of thought in diffusion language models. *arXiv preprint arXiv:2606.23567*, 2026.

Lixing Zhang, Yuchen Liang, and Liyan Xie. Masked diffusion modeling for anomaly detection. *arXiv preprint arXiv:2605.30046*, 2026.

Zikun Zhang, Zixiang Chen, and Quanquan Gu. Convergence of score-based discrete diffusion models: A discrete-time analysis. In *International Conference on Learning Representations*, volume 2025, pp. 34747–34772, 2025.

Lingxiao Zhao, Xueying Ding, Lijun Yu, and Leman Akoglu. Unified discrete diffusion for categorical data. *Journal of Machine Learning Research*, 26(215):1–49, 2025.

Yixiu Zhao, Jiaxin Shi, Feng Chen, Shaul Druckmann, Lester Mackey, and Scott Linderman. Informed correctors for discrete diffusion models. *Advances in Neural Information Processing Systems*, 38:125510–125538, 2026.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *International Conference on Learning Representations*, volume 2025, pp. 63186–63227, 2025.

Yuchen Zhu, Wei Guo, Jaemoo Choi, Guan-Horng Liu, Yongxin Chen, and Molei Tao. Mdns: Masked diffusion neural sampler via stochastic optimal control. *Advances in Neural Information Processing Systems*, 38:35260–35308, 2026.

APPENDIX

CONTENTS

A Related Work **15**

B Problem Formulation: Full Details **16**

B.1 Background 16

B.2 Problem Formulation 18

C Proof Sketch: Convergence Analysis via Adjoint Equations **20**

D Proof of Main Theorems **21**

D.1 Proof of Lemma 3.2 21

D.2 Proof of Corollary D.1 27

D.3 Proof of Theorem 3.1 28

D.4 Proof of Corollary 3.3 28

E Proof of Auxiliary Lemmas For Main Theorems **30**

E.1 Lemmas for the Adjoint Equation 31

E.2 Lemmas for the Training Loss Term 32

F Bounding the Initialization Error Term under Uniform Rate **41**

G Bounding the Early-Stopping Error Term **43**

H S-Independence of the Bounds **44**

H.1 Scaling of the Bounds 44

H.2 Derivational Details to Achieve S-Independence 45

I Proof of Uniformization Iteration Complexity **46**

I.1 Proof of Auxiliary Lemmas 46

I.2 Proof of Corollary 3.4 49

A RELATED WORK

Discrete Diffusion Models. Discrete diffusion models were first proposed by Sohl-Dickstein et al. (2015). A substantial line of work formulates the forward and reverse processes as discrete-time Markov chains over categorical states (Austin et al., 2021; Hoogeboom et al., 2021; Chen et al., 2024), with the model trained by optimizing the standard variational lower bound. Within this family, multinomial diffusion (Hoogeboom et al., 2021) and D3PM (Austin et al., 2021) mirror

DDPM (Ho et al., 2020) in the continuous setting, while DNDM (Chen et al., 2024) introduces a non-Markovian reverse process that enables accelerated sampling, corresponding to the role of DDIM (Song et al., 2021a).

In a parallel line of work, Campbell et al. (2022) formulate discrete diffusion as a continuous-time Markov chain (CTMC). This formulation allows for natural theoretical convergence analysis and enables more flexible numerical inference that is not constrained by the discretization fixed at training time. Importantly, the formulation admits an exact formula for the reverse process, in which the unknown quantity is the ratio of marginal probabilities. This ratio is the discrete analog of the score in continuous diffusion (Song et al., 2021b), and is commonly referred to as the discrete or concrete score (Meng et al., 2022). A major line of work has since focused on learning the discrete score. Meng et al. (2022) propose concrete score matching with an L^2 objective, Sun et al. (2023) learn it via cross-entropy on conditional marginals, and Lou et al. (2024) introduce the score entropy loss, which improves training stability and sample quality.

Building on the CTMC framework, recent work explores alternative formulations for discrete generative modeling (Zhu et al., 2026; Li et al., 2026), and Park et al. (2025); Zhao et al. (2026); Ren et al. (2026); Liang et al. (2026c) propose higher-order, adaptive, or correction-based numerical solvers to accelerate inference. In parallel, a body of CTMC-inspired but simplified models has emerged as a practical alternative. These models, particularly those based on masked diffusion, have shown promising performance on large-scale language and image generation (Chang et al., 2022; 2023; Shi et al., 2024; Sahoo et al., 2024; Zheng et al., 2025; Kim et al., 2025; 2026; Nie et al., 2026; Xu et al., 2026; Zhang et al., 2026), bridging the gap in accuracy and speed to their autoregressive counterparts.

Convergence Analysis. Campbell et al. (2022) derive a total variation error bound for the tau-leaping sampling algorithm, but their results rely on strong assumptions and are not directly linked to a practical training objective.

Chen & Ying (2025) perform a convergence analysis on the binary hypercube, deriving an approximation error bound in KL divergence, similar results are also available in Le-Tuyet-Nhi et al. (2025). Zhang et al. (2025) extend this result to general categorical state spaces, also in KL. Ren et al. (2025) propose a stochastic integral framework that yields KL-divergence error bounds via path-measure arguments, while Liang et al. (2026b) establish KL bounds through a differential-inequality argument that avoids path-measure techniques. These analyses all *fundamentally rely on KL-based estimates, which diverge when the prior is singular*. As a result, *their guarantees do not directly apply to masked diffusions*—the dominant paradigm for discrete diffusion language models.

Some recent works have attempted to address this gap. Liang et al. (2025); Dmitriev et al. (2026) mix the uniform and singular mask distributions as a surrogate prior, circumventing the divergent KL at the all-mask state. This construction, however, does not apply to true masked diffusions. Conforti et al. (2025) combine KL-based path-measure estimates with a separate TV control of the initialization mismatch at the fully masked state. More recently, Liang et al. (2026d) address the KL issue by switching to TV distance for the approximation error. *Nevertheless, all these results depend on the vocabulary size, and the bounds can become vacuous for modern language tasks.*

A separate line of theory studies the sampling efficiency of discrete diffusion models, often under idealized score assumptions Chen et al. (2025); Li & Cai (2026); Liang et al. (2026a); these questions are complementary to, but outside the scope of the present work.

B Problem Formulation: Full Details

B.1 Background

Notation and Preliminaries. We denote scalars by lowercase letters (x) and vectors by boldface lowercase letters ($\mathbf{x}$). All vectors are assumed to be column vectors unless otherwise specified. We denote the i-th entry of a vector $\mathbf{x}$ by x^i. Specifically, a d-dimensional vector is given by $\mathbf{x} = [x^1 \cdots x^d]$. We use $\mathbf{x}_{i \to \hat{x}^i}$ to represent the vector $\mathbf{x}$ with its i-th entry replaced by $\hat{x}^i$. We denote $\mathbf{x}^{\setminus i}$ as the vector $\mathbf{x}$ with its i-th entry removed.

For $N \in \mathbb{Z}^+$, $[N]$ denotes the set $\{1, 2, \cdots, N\}$. We denote $\mathbf{e}_i$ as the i-th standard basis vector, $\mathbf{1}_N$ as the N-dimensional vector of all ones, and I_N as the $N \times N$ identity matrix. We use $\delta\{\cdot, \cdot\}$ to represent the Kronecker delta function. The Hamming distance between $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ is the number of indices where the entries differ: $\text{Ham}(\mathbf{x}, \mathbf{y}) = |\{i \in [N] : x^i \neq y^i\}|$, where $|\cdot|$ denotes the cardinality of a set.

We consider data defined on a discrete state space $\mathcal{X} = [N]$, where N is the total number of possible states. A probability distribution on this discrete space is characterized by a probability mass vector $p \in \Delta(\mathcal{X})$, representing the probability simplex over the N possible states.

Forward Process. The forward noising process is defined as a continuous-time Markov chain (CTMC) on $\mathcal{X}$ over the time interval $[0, T]$, governed by a rate matrix (or infinitesimal generator) $Q_t \in \mathbb{R}^{N \times N}$. Specifically, the process is initialized at $t = 0$ with the data distribution p_{data}. The evolution of the marginal distribution p_t is governed by the *Kolmogorov forward equation (KFE)*, a linear ordinary differential equation (ODE) given by

$$\frac{dp_t}{dt} = Q_t^\top p_t, \quad p_0 = p_{\text{data}}, \quad \text{for } t \in [0, T]. \tag{21}$$

Here, $Q_t(x, y)$ defines the infinitesimal rate of transitioning from states x to y. Because Q_t is constrained to have nonnegative off-diagonal entries and zero row sums, the dynamics (21) preserve the total mass of p_t, ensuring it remains a valid probability distribution throughout the evolution.

There are two common choices for Q_t. They are the masked (absorbing) and uniform rates and given by

$$(Q_t^{\text{masked}})^\top = \beta(t)(\mathbf{e}_{[\text{MASK}]}\mathbf{1}_N^\top - I_N) \quad \text{and} \quad (Q_t^{\text{uniform}})^\top = \beta(t)\left(\tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top - I_N\right), \tag{22}$$

respectively. For the masked rate, the Nth entry corresponds to the mask token [MASK]. Consequently, p_t approaches a tractable limiting distribution p_{base}: either a Dirac distribution centered on the mask token or the uniform distribution, as $t \to \infty$. The ease of sampling from these distributions makes them ideal priors for the reverse generative process.

The dynamics (21) can be numerically integrated by discretizing the time horizon into intervals with width Δt. Specifically, the forward transition kernel from time t to $t + \Delta t$ is given by

$$p_{t+\Delta t|t}(y|x) = \delta\{x, y\} + Q_t(x, y)\Delta t + o(\Delta t), \tag{23}$$

where $\delta\{\cdot, \cdot\}$ denotes the Kronecker delta, and $Q_t(x, y)$ denotes the (x, y)-th entry of Q_t. Numerical simulation of CTMCs can be broadly divided into exact path simulation such as the Gillespie's algorithm and approximate simulation such as τ-leaping method. We refer to Gillespie (2001); Cao et al. (2004; 2006); Hobolth & Stone (2009); Arns et al. (2010) for a comprehensive review of these methods.

Reverse Process. The forward process (21) admits a reverse-time process, which is also a CTMC (Kelly, 1979). The reverse process satisfies the reverse-time Kolmogorov equation

$$-\frac{dp_t}{dt} = (Q_t^{\leftarrow})^\top p_t, \quad \text{for } t \in [0, T], \tag{4}$$

where $Q_t^{\leftarrow}$ is the reverse-time rate matrix, and

$$Q_t^{\leftarrow}(x, y) = \frac{p_t(y)}{p_t(x)} Q_t(y, x) \quad \text{for } x \neq y, \quad \text{and} \quad Q_t^{\leftarrow}(x, x) = -\sum_{y \neq x} Q_t^{\leftarrow}(x, y). \tag{24}$$

Training Formulation. The ratios $s_t(x) := \left[\frac{p_t(y)}{p_t(x)}\right]_{y \neq x} \in \mathbb{R}_{\geq 0}^{N-1}$ are known as the discrete/-concrete score (Meng et al., 2022), which is analogous to the Stein score $\nabla_x \log p_t$ in continuous diffusion models (Song et al., 2021b). We denote the entry corresponding to the state y as $s_t(x)_y := \frac{p_t(y)}{p_t(x)} \in \mathbb{R}_{\geq 0}$. In score-based discrete diffusion, the goal is to learn a score estimator

$$\tilde{s}_t(x) \approx s_t(x) = \left[\frac{p_t(y)}{p_t(x)}\right]_{y \neq x}, \quad \text{for } t \in [0, T]. \tag{25}$$

To this end, we consider two standard training losses. The first is a weighted score-matching (WSM) loss (Meng et al., 2022)

$$\mathcal{L}_{\mathrm{WSM}}(s, \tilde{s}) = \int_0^T \frac{1}{2} \mathbb{E}_{x \sim p_t} \sum_{y \neq x} w_t(x, y) \left(s_t(x)_y - \tilde{s}_t(x)_y\right)^2 dt, \tag{26}$$

and the second is the score entropy (SE) loss (Benton et al., 2024; Lou et al., 2024)

$$\mathcal{L}_{\mathrm{SE}}(s, \tilde{s}) = \int_0^T \mathbb{E}_{x \sim p_t} \sum_{y \neq x} w_t(x, y) \left(s_t(x)_y \log \left(\frac{s_t(x)_y}{\tilde{s}_t(x)_y} \right) + \tilde{s}_t(x)_y - s_t(x)_y \right) dt, \tag{27}$$

where $w_t(x, y) \geq 0$ are weighting coefficients, and $\mathcal{L}_{\mathrm{SE}}$ is obtained from the Bregman divergence generated by the function $a \mapsto a \log a - a$. Following Lou et al. (2024), we consider the transition weight where $w_t(x, y) = Q_t(y, x)$. This choice parallels the score-matching loss in continuous space, where it is typically weighted by the diffusion coefficient to align with the objective with the evidence lower bound (ELBO) (Song et al., 2021b).

In practice, this training objective is intractable as the ground-truth score depends on the unknown marginal distribution p_t. To resolve this, one typically uses denoising or implicit score matching, which are tractable objectives that are equivalent to the original loss up to a constant independent of the score estimator Lou et al. (2024).

Integral Probability Metric (IPM). IPMs are a general class of metrics for measuring the discrepancy between two probability measures $p, \tilde{p} \in \mathcal{P}(\mathcal{X})$. They are defined as

$$\gamma_\Psi(p, \tilde{p}) = \sup_{\psi \in \Psi} \left| \mathbb{E}_{\mathbf{x} \sim p}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \tilde{p}}[\psi(\mathbf{x})] \right|, \tag{11}$$

where Ψ is a suitable class of admissible functions. Different choices of Ψ induce different types of IPMs. Commonly used IPMs include the total variation (TV) distance, Wasserstein-1 (W_1) distance, and maximum mean discrepancy (MMD). These correspond to choosing Ψ as $\{\psi : \mathcal{X} \to \mathbb{R}, \|\psi\|_\infty \leq 1\}$, $\{\psi : \mathcal{X} \to \mathbb{R}, \|\psi\|_{\mathrm{Lip}} \leq 1\}$, and $\{\psi : \mathcal{X} \to \mathbb{R}, \|\psi\|_{\mathcal{H}} \leq 1\}$, respectively, where $\|\cdot\|_{\mathcal{H}}$ is the norm for a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. We refer readers to Sriperumbudur et al. (2009); Peyré & Cuturi (2019); Birrell et al. (2022) for a more detailed overview of IPMs and their properties.

B.2 Problem Formulation

The discrete state space is given by $\mathcal{X} = [S]^d$, containing sequences that factorize as $\mathbf{x} = [x^1, \cdots, x^d]$. This setup is common among various applications, including text token sequences (Lou et al., 2024), image pixel values (Campbell et al., 2022), tabular data (Kotelnikov et al., 2023), chemical molecule graphs (Vignac et al., 2023), musical notes (Campbell et al., 2022), DNA sequences (Schiff et al., 2025), protein sequences (Campbell et al., 2024), and others.

Forward Noising Process. The forward process $X = \{X_t\}_{0 \leq t \leq T}$ is defined as a continuous-time Markov chain (CTMC) on $\mathcal{X}$. The distribution of X_t is characterized by a probability mass vector $p_t \in \Delta(\mathcal{X})$, representing the probability simplex over the S^d possible states. This distribution is governed by a Kolmogorov forward equation with rate matrix $Q_t \in \mathbb{R}^{S^d \times S^d}$ given by

$$\frac{dp_t}{dt} = Q_t^\top p_t, \quad p_0 = p_{\mathrm{data}}, \quad \text{for } t \in [0, T]. \tag{1}$$

Since the dimensions of Q_t scale exponentially with the sequence length d, direct computation or storage of the generator is intractable. In practice, Q_t is structured to operate on each dimension independently via a token-level rate matrix $Q_t^{\mathrm{tok}} \in \mathbb{R}^{S \times S}$. In standard frameworks (Campbell et al., 2022; Lou et al., 2024; Zhang et al., 2025), the nonzero off-diagonal entries of Q_t are structured as follows (although our results apply to general rate matrices; see Section 3)

$$Q_t(\mathbf{x}, \mathbf{x}_{i \to \hat{x}^i}) = Q_t\left([x^1, \cdots, x^i, \cdots, x^d], [x^1, \cdots, \hat{x}^i, \cdots, x^d]\right) = Q_t^{\mathrm{tok}}(x^i, \hat{x}^i), \tag{2}$$

where $x^i \neq \hat{x}^i \in [S]$, and Q_t^{tok} is either the masked rate matrix or uniform rate matrix (3). Under these two choices, the asymptotic limit p_{base} is the Dirac distribution centered on the mask token and

the uniform distribution over $\mathcal{X}$, respectively. The structure in (2) implies that Q_t is a Kronecker sum of independent generators, modeling transitions between states at Hamming distance 1 (i.e., states differing by exactly one token).

Under this structure of Q_t, the forward transition kernel $p_{t|s}(\mathbf{x}_t|\mathbf{x}_s) := \Pr(X_t = \mathbf{x}_t|X_s = \mathbf{x}_s)$ factorizes as $p_{t|s}(\mathbf{x}_t|\mathbf{x}_s) = \prod_{i=1}^{d} p^i_{t|s}(x^i_t|x^i_s)$ for $0 \le s < t \le T$. Here, $p^i_{t|s}(x^i_t|x^i_s) := \Pr(X^i_t = x^i_t|X^i_s = x^i_s)$ is the forward transition kernel for the ith token governed by rate Q^{tok}_t.

The factorization of transition kernel is standard in the discrete diffusion literature, as it is essential for scaling to high-dimensional state spaces for empirical implementation Austin et al. (2021); Campbell et al. (2022; 2024); Lou et al. (2024). This assumption is also widely adopted in theoretical analysis Zhao et al. (2025); Chen & Ying (2025); Zhang et al. (2025); Ren et al. (2025).

Reverse Denoising Process. The forward process (1) admits a reverse-time (generative) process governed by the following Kolmogorov equation (Kelly, 1979)

$$-\frac{dp_t}{dt} = (Q^{\leftarrow}_t)^\top p_t, \quad \text{for } t \in [0, T], \tag{28}$$

where the reverse rate matrix $Q^{\leftarrow}_t$ is given by

$$Q^{\leftarrow}_t(\mathbf{x}, \mathbf{y}) = \frac{p_t(\mathbf{y})}{p_t(\mathbf{x})} Q_t(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{d} \frac{p_t(\mathbf{y})}{p_t(\mathbf{x})} \delta\{\mathbf{x}^{\setminus i}, \mathbf{y}^{\setminus i}\} Q^{\text{tok}}_t(y^i, x^i) \quad \text{for } \mathbf{x} \neq \mathbf{y}, \tag{5}$$

and $Q^{\leftarrow}_t(\mathbf{x}, \mathbf{x}) = -\sum_{\mathbf{y} \neq \mathbf{x}} Q^{\leftarrow}_t(\mathbf{x}, \mathbf{y})$. In the second step of (5), we used the definition of Q_t (2). The Kronecker delta ensures that at most one term in the summation survives—specifically, the term where $\mathbf{x}$ and $\mathbf{y}$ differ by exactly one token. Thus, the reverse rate matrix $Q^{\leftarrow}_t$ inherits the same sparse structure as the forward rate Q_t, restricting non-zero transition rates to sequences at Hamming distance 1.

Hence, we consider the discrete score s_t defined for these sequences

$$s_t(\mathbf{x})_{i \to \hat{x}^i} := \frac{p_t(\mathbf{x}_{i \to \hat{x}^i})}{p_t(\mathbf{x})} \quad \text{for } \mathbf{x} \in \mathcal{X},\ i \in [d],\ x^i \neq \hat{x}^i \in [S].$$

The score estimator $\tilde{s} : [0, T] \times \mathcal{X} \to \mathbb{R}^{d(S-1)}$ can then be obtained by minimizing $\mathcal{L}_{\text{WSM}}(s, \tilde{s})$ defined in (26) or $\mathcal{L}_{\text{SE}}(s, \tilde{s})$ defined in (27).

Approximate Reverse Process. The generative process is governed by the approximate reverse-time dynamics

$$-\frac{d\tilde{p}_t}{dt} = (\tilde{Q}^{\leftarrow}_t)^\top \tilde{p}_t, \quad \tilde{p}_T = p_{\text{base}} \approx p_T, \quad \text{for } t \in [0, T], \tag{9}$$

where the approximate reverse rate matrix $\tilde{Q}^{\leftarrow}_t$ is defined by

$$\tilde{Q}^{\leftarrow}_t(\mathbf{x}, \mathbf{y}) = \tilde{s}_t(\mathbf{x})_{\mathbf{y}} Q_t(\mathbf{y}, \mathbf{x}) \quad \text{for } \mathbf{x} \neq \mathbf{y}, \quad \text{and} \quad \tilde{Q}^{\leftarrow}_t(\mathbf{x}, \mathbf{x}) = -\sum_{\mathbf{y} \neq \mathbf{x}} \tilde{Q}^{\leftarrow}_t(\mathbf{x}, \mathbf{y}). \tag{10}$$

It is important to note that the approximate reverse process differs from the exact reverse process (28) by two key aspects. First, the true reverse rate $Q^{\leftarrow}_t$ is replaced by the approximate $\tilde{Q}^{\leftarrow}_t$ constructed using the score estimator through (10). Second, since the ground truth terminal distribution p_T is intractable, the prior p_{base} is used as a starting distribution for the reverse process. Here, $\tilde{p}$ denotes the marginal distribution of the approximate reverse process.

Error and Adjoint Equations. Our analysis relies on computing the mismatch between the exact generative process (4) and its approximation (9) over time. Defining the error $\lambda_t := p_t - \tilde{p}_t$ and subtracting (9) from (4), we obtain the *error equation* for $t \in [0, T]$,

$$\frac{d\lambda_t}{dt} = (-Q^{\leftarrow}_t + \tilde{Q}^{\leftarrow}_t)^\top p_t - (\tilde{Q}^{\leftarrow}_t)^\top \lambda_t, \quad \text{with} \quad \lambda_T = p_T - p_{\text{base}}. \tag{12}$$

To quantify this discrepancy under any IPM, we extend the duality argument of Mimikos-Stamatopoulos et al. (2024, Section 4.1) originally developed for continuous diffusion, by developing new technical tools specialized for the discrete setting. The core idea is to test (12) against a function $\phi : [0,T] \times [S]^d \to \mathbb{R}$ that satisfies the *adjoint equation* associated with the approximate reverse dynamics (9), namely the *Kolmogorov backward equation*

$$\frac{d\phi_t}{dt} = \tilde{Q}_t^{\leftarrow}\phi_t, \quad \phi_0(\mathbf{x}) = \psi(\mathbf{x}), \quad \text{for } t \in [0,T]. \tag{13}$$

By selecting ψ from the admissible function class Ψ of the target metric (11), we can leverage (13) to explicitly compute, in any IPM, how the approximation error propagates back to the initial data distribution.

C Proof Sketch: Convergence Analysis via Adjoint Equations

In this section, we highlight the key technical innovation underlying our main Lemma 3.2 and Corollaries D.1-3.1. Our core strategy involves (1) constructing test functions for the IPMs (11) via an adjoint equation, (2) estimating the regularity of the test functions, and (3) bounding the approximation error for any IPM by specializing the admissible function class accordingly. The detailed derivations are given in Appendix D.

Adjoint Equation. We first compute the mismatch between the exact generative process (4) and its approximation (9) over time. To this end, defining the error $\lambda_t := p_t - \tilde{p}_t$ and subtracting (9) from (4), we obtain the *error equation* for $t \in [0,T]$,

$$\frac{d\lambda_t}{dt} = (-Q_t^{\leftarrow} + \tilde{Q}_t^{\leftarrow})^\top p_t - (\tilde{Q}_t^{\leftarrow})^\top \lambda_t, \quad \text{with} \quad \lambda_T = p_T - p_{\text{base}}. \tag{12}$$

To quantify this discrepancy under any IPM, we extend the duality argument of Mimikos-Stamatopoulos et al. (2024, Section 4.1) originally developed for continuous diffusion, by developing new technical tools specialized for the discrete setting. The core idea is to test (12) against a function $\phi : [0,T] \times [S]^d \to \mathbb{R}$ that satisfies the *adjoint equation* associated with the approximate reverse dynamics (9), namely the *Kolmogorov backward equation*

$$\frac{d\phi_t}{dt} = \tilde{Q}_t^{\leftarrow}\phi_t, \quad \phi_0(\mathbf{x}) = \psi(\mathbf{x}), \quad \text{for } t \in [0,T]. \tag{13}$$

In particular, we integrate the error equation (12) against the test function ϕ_t in space, integrate in time and apply integration by parts to move the time derivative to ϕ_t, yielding

$$\int_0^T \langle \frac{d}{dt}\lambda_t, \phi_t \rangle \, dt = \langle \lambda_T, \phi_T \rangle - \langle \lambda_0, \phi_0 \rangle - \int_0^T \langle \lambda_t, \frac{d}{dt}\phi_t \rangle \, dt. \tag{29}$$

Substituting (12) and (13) into (29), the λ_t-dependent contributions cancel via the adjointness identity $\langle (\tilde{Q}_t^{\leftarrow})^\top \lambda_t, \phi_t \rangle = \langle \lambda_t, \tilde{Q}_t^{\leftarrow}\phi_t \rangle$ – which is precisely why $\tilde{Q}_t^{\leftarrow}$ is the choice of generator in (13) – leaving only the p_t-dependent rate-matrix difference:

$$\langle \lambda_0, \psi \rangle = \langle \lambda_0, \phi_0 \rangle = \langle \lambda_T, \phi_T \rangle + \int_0^T \left\langle (Q_t^{\leftarrow} - \tilde{Q}_t^{\leftarrow})^\top p_t, \phi_t \right\rangle dt. \tag{30}$$

Taking absolute values and plugging in $\lambda_t = p_t - \tilde{p}_t$, we obtain an upper bound for the approximation error

$$\begin{aligned}
&|\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \\
&\leq |\langle \lambda_T, \phi_T \rangle| + \int_0^T \left| \left\langle (Q_t^{\leftarrow} - \tilde{Q}_t^{\leftarrow})^\top p_t, \phi_t \right\rangle \right| dt \\
&\leq \|\phi_T\|_\infty \sum_{\mathbf{x}\in\mathcal{X}} |p_T - p_{\text{base}}|(\mathbf{x}) + 2\int_0^T \|\phi_t\|_\infty \sum_{\substack{\mathbf{x},\mathbf{y}\in\mathcal{X} \\ \mathbf{x}\neq\mathbf{y}}} p_t(\mathbf{x}) \left| (Q_t^{\leftarrow} - \tilde{Q}_t^{\leftarrow})(\mathbf{x},\mathbf{y}) \right| dt \\
&= 2\|\phi_T\|_\infty \text{TV}(p_T, p_{\text{base}}) + 2\int_0^T \|\phi_t\|_\infty \sum_{\substack{\mathbf{x},\mathbf{y}\in\mathcal{X} \\ \mathbf{x}\neq\mathbf{y}}} p_t(\mathbf{x}) Q_t(\mathbf{y},\mathbf{x}) \, |s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}| \, dt,
\end{aligned} \tag{31}$$

where $\|\phi_t\|_\infty := \max_{\mathbf{x}\in\mathcal{X}} |\phi_t(\mathbf{x})|$, and in the last step, we used the reverse rates formulations (24) and (10), as well as the definition of TV distance. Importantly, the integration term contains the score error $|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|$, which we represent using the standard training losses (7)–(8) through our novel score-marginal cancellation technique; see Lemmas E.6 and E.9 and their proofs for more details. As a result, our bounds are expressed directly in terms of the standard training losses.

Regularity of Test Functions. To make the bound (31) explicit, it remains to control the test function sup-norm $\|\phi_t\|_\infty$. Recall that ϕ_t is the time evolution of ψ under the adjoint equation (13), with $\phi_0 = \psi$. Importantly, we show in Lemma E.1 that the adjoint dynamics is non-expansive in the sup-norm: the solution ϕ_t to the adjoint equation satisfies $\|\phi_t\|_\infty \le \|\psi\|_\infty$ for all $t \in [0, T]$. Thus, plugging in the contraction result into (31), we obtain

$$\begin{aligned}&|\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \\ &\le 2\|\psi\|_\infty \mathrm{TV}(p_T, p_{\mathrm{base}}) + 2\|\psi\|_\infty \int_0^T \sum_{\substack{\mathbf{x},\mathbf{y}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} p_t(\mathbf{x}) Q_t(\mathbf{y},\mathbf{x})\, |s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|\, dt, \end{aligned} \tag{32}$$

effectively removing the only remaining unknown. Intuitively, the non-expansiveness holds because ϕ_t is an expectation of ψ over trajectories of the CTMC associated with the adjoint equation, and an expectation never exceeds the maximum. In our proof of Lemma E.1, we make this rigorous by using a Feynman-Kac representation.

Specialization to Any IPM. The bound (32) is valid for any admissible function ψ, and its right-hand side depends on ψ only through the multiplicative factor $\|\psi\|_\infty$. Taking the supremum over the admissible class Ψ of the target IPM (11), the left-hand side becomes the IPM $\gamma_\Psi(p_{\mathrm{data}}, \tilde{p}_0)$. Since $\|\psi\|_\infty$ is decoupled from the other terms, those terms are unaffected by the supremum. And we have a prefactor C_Ψ which is a constant multiple of $\sup_{\psi\in\Psi} \|\psi\|_\infty$. Since the state space $\mathcal{X} = [S]^d$ is finite, any admissible function ψ is bounded and C_Ψ is finite for every IPM. A single bound thus specializes to every IPM, with only the constant C_Ψ changing.

D Proof of Main Theorems

We restate and prove the main theorems.

D.1 Proof of Lemma 3.2

Lemma 3.2 (Weak Error Bound for Score-based Discrete Diffusion)**.** *Under the same assumptions as Theorem 3.1, the weak error between the true data distribution p_{data} and the generated data distribution $\tilde{p}_0$ satisfies the following.*

$$|\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \;\le\; 2\,\|\psi\|_\infty \Big(\underbrace{\mathcal{I}}_{\textit{init. error}} + \sqrt{2\,\mathcal{C}}\ \underbrace{\sqrt{\mathcal{L}}}_{\textit{train. error}} \Big). \tag{19}$$

Here, $(\mathcal{I}, \mathcal{C}, \mathcal{L})$ are reported in Table 3.

Table 4: Coefficients of the convergence bounds for Lemma 3.2 and Corollary 3.3

Rate	Initialization Error $\mathcal{I}$	Coefficient $\mathcal{C}$	Training Error $\mathcal{L}$
Masked	$d\, e^{-\|\beta\|_1}$	$d\,\|\beta(1-p(\mathbf{m}))\|_1$	$\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \frac{2}{3}\mathcal{L}_3(s,\tilde{s})$
Uniform	$d\, e^{-\|\beta\|_1}$	$d\,\|\beta\|_1$	$\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \frac{2}{3}\mathcal{L}_3(s,\tilde{s})$
Uniform	$d\, e^{-\|\beta\|_1}$	$d\,\|\beta\|_1$	$\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})$
General	$\mathrm{TV}(p_T, p_{\mathrm{base}})$	$\int_0^T \mathbb{E}_{p_t}[\mu_t^{\mathrm{out}}]\, dt$	$\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \frac{2}{3}\mathcal{L}_3(s,\tilde{s})$
General	$\mathrm{TV}(p_T, p_{\mathrm{base}})$	$\int_0^T \mathbb{E}_{p_t}[\mu_t^{\mathrm{in}}]\, dt$	$\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})$

Here, the cubic correction term is given by

$$\mathcal{L}_3(s,\tilde{s}) = \begin{cases} \displaystyle\int_0^T \beta(t)(1-p_t(\mathbf{m}))\mathbb{E}_{\mathbf{y}\sim p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m})}\left[\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})}\frac{|s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}}|^3}{(s_t(\mathbf{x})_{\mathbf{y}})^3}\right]dt, & \text{(33a)}\\ \text{under the masked rate,} \\ \displaystyle\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\frac{\beta(t)}{S}\sum_{\substack{\mathbf{y}\in\mathcal{X}\\ d_H(\mathbf{y},\mathbf{x})=1}}\left(\frac{|s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}}|^3}{(s_t(\mathbf{x})_{\mathbf{y}})^2}\right)\right]dt, & \text{(33b)}\\ \text{under the uniform rate,} \\ \displaystyle\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} Q_t(\mathbf{y},\mathbf{x})\left(\frac{|s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}}|^3}{(s_t(\mathbf{x})_{\mathbf{y}})^2}\right)\right]dt = \int_0^T \mathbb{E}_{\mathbf{y}\sim p_t}\left[\sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x})\left(\frac{|s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}}|^3}{(s_t(\mathbf{x})_{\mathbf{y}})^3}\right)\right]dt, & \text{(33c)}\\ \text{under a general rate.} \end{cases}$$

where $\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})$ is the successor set of $\mathbf{y}$ (i.e. containing sequences $\mathbf{y}$ can transit to), defined by

$$\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y}) = \left\{\mathbf{x}\in\mathcal{X}\,\middle|\,\exists i\in[d], \mathbf{x}^{\setminus i}=\mathbf{y}^{\setminus i}, x^i=[\text{MASK}], y^i\neq[\text{MASK}]\right\}. \tag{34}$$

Remark D.1 (Observables, not metrics, are the primary objects of our analysis)**.** *The first output of our analysis is a single-observable guarantee (the Lemma above): for every bounded $\psi:\mathcal{X}\to\mathbb{R}$,* (19) *controls $|\mathbb{E}_{p_{\text{data}}}\psi - \mathbb{E}_{\tilde{p}_0}\psi|$ by $\|\psi\|_\infty$ times the approximation-error terms bounded in Theorem 3.1, with ψ entering only through $\|\psi\|_\infty$. The IPM bound of Theorem 3.1 is the corollary obtained by taking a supremum over a class Ψ. In language modeling Lou et al. (2024) one wants this supremum; in most other applications the success criterion is itself a fixed bounded observable: motif and k-mer frequencies for DNA Schiff et al. (2025), the designable fraction and mean pLDDT for proteins Campbell et al. (2024), low-order marginals and moments for tabular data Kotelnikov et al. (2023), and maximum-mean-discrepancy metrics—IPMs with a bounded kernel—for images Campbell et al. (2022) and molecular graphs Vignac et al. (2023). Each has $\|\psi\|_\infty = O(1)$, so our bound applies directly, without ever forming a full metric.*

A subtlety worth stating: the Fréchet-type scores common in these fields (FID, FCD, and sequence analogues) are not IPMs—they are W_2 distances between Gaussians fitted in a feature space—so they do not arise from the supremum in Theorem 3.1 directly. They are, however, smooth functions of finitely many bounded *observables: the first two moments $\mathbb{E}\phi$, $\mathbb{E}[\phi\phi^\top]$ of a fixed feature map ϕ. Our single-observable estimate controls each such moment S-freely, so FID and its analogues are governed by our guarantee up to the local Lipschitz modulus of the Fréchet functional and the (S-independent) feature dimension, degrading only near covariance degeneracy. The observable Lemma thus reaches **beyond IPMs**: an IPM is a single* linear *observable, a Fréchet score a* nonlinear *functional of moment-observables, and both are controlled.*

Remark D.2 (Details of Exit-Rate Routing Technique)**.** *Writing the Kolmogorov forward equation* (1) *entry-wise at state $\mathbf{x}$ yields the* flow balance

$$\frac{dp_t(\mathbf{x})}{dt} \;=\; \underbrace{\sum_{\mathbf{y}\neq\mathbf{x}} Q_t(\mathbf{y},\mathbf{x})\,p_t(\mathbf{y})}_{\textit{actual inflow}} \;-\; \underbrace{\mu_t^{\text{out}}(\mathbf{x})\,p_t(\mathbf{x})}_{\textit{actual outflow}},$$

in which the jump rate $\mu_t^{\text{out}}(\mathbf{x}) = -Q_t(\mathbf{x},\mathbf{x})$ governs the outflow and the entrance rate $\mu_t^{\text{in}}(\mathbf{x}) = \sum_{\mathbf{y}\neq\mathbf{x}} Q_t(\mathbf{y},\mathbf{x})$ governs the inflow. Both are intrinsic to the generator Q_t: μ_t^{out} is the negative diagonal, μ_t^{in} is the column-off-diagonal sum.

Under the uniform rate (3)*, Q_t^{tok} is symmetric and*

$$\mu_t^{\text{out,tok}}(x) \;=\; \mu_t^{\text{in,tok}}(x) \;=\; \beta(t)\big(1-\tfrac{1}{S}\big) \quad \textit{for every } x\in[S],$$

so the two lemmas give the same scaling.

Under the masked rate (3)*, Q_t^{tok} is asymmetric and the two rates differ:*

$$\mathbb{E}_{p_t^{\text{tok}}}\big[\mu_t^{\text{out,tok}}(x)\big] \;=\; \beta(t)\big(1-p_t^{\text{tok}}(\textit{[MASK]})\big)$$

is S-independent, whereas

$$\mathbb{E}_{p_t^{\mathrm{tok}}}\big[\mu_t^{\mathrm{in,tok}}(x)\big] \;=\; (S-1)\,\beta(t)\,p_t^{\mathrm{tok}}(\textit{[MASK]})$$

carries a factor of $S-1$ — the per-token vocabulary size.

Accordingly, in the proof of Lemma 3.2 we use the SE route for the masked case to obtain an S-independent prefactor, and the simpler WSM route (Lemma E.9) for the uniform case, where its prefactor coincides with the SE one. Of course, either bound is mathematically valid, but only the SE route yields an S-independent prefactor under masked dynamics.

Remark D.3 (S-free scaling of $\mathcal{L}_3$)**.** *We note that $|\mathcal{N}^{\leftarrow}_{\textit{masked}}(\mathbf{y})| \le d$ in* (34)*, where d is the sequence length. Overall, in* (33a)*, the outer expectation $\mathbb{E}_{\mathbf{y}\sim p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m})}[\,\cdot\,]$ is a probability average; its support has size S^d-1, but the integrand averaged is the at-most-d-term inner sum of cubic errors, with no count of states entering per term. Thus, the scaling of $\mathcal{L}_3$ is independent of the vocabulary size S under the masked rate.*

Under the uniform rate, the outer expectation $\mathbb{E}_{\mathbf{y}\sim p_t}$ of (33b) *is again a probability average. For each $\mathbf{y}$, the inner summation $\sum_{d_H(\mathbf{x},\mathbf{y})=1}$ runs over the $d(S-1)$ neighbors satisfying $d_H(\mathbf{y},\mathbf{x})=1$; together with the factor $\frac{1}{S}$, its magnitude is on the order of $\frac{d(S-1)}{S} < d$. Thus the scaling of $\mathcal{L}_3$ is again independent of S.*

Proof of Lemma 3.2.

$$\begin{aligned}
\langle\lambda_0,\psi\rangle - \langle\lambda_T,\phi_T\rangle &= \langle\lambda_0,\phi_0\rangle - \langle\lambda_T,\phi_T\rangle \\
&= -\int_0^T \frac{d}{dt}\langle\lambda_t,\phi_t\rangle dt \\
&= -\int_0^T \langle\frac{d\lambda_t}{dt},\phi_t\rangle + \langle\lambda_t,\frac{d\phi_t}{dt}\rangle dt \\
&= \int_0^T \Big\langle (Q_t^{\leftarrow}-\tilde{Q}_t^{\leftarrow})^\top p_t + (\tilde{Q}_t^{\leftarrow})^\top\lambda_t,\phi_t\Big\rangle - \Big\langle\lambda_t,\tilde{Q}_t^{\leftarrow}\phi_t\Big\rangle\, dt \quad \text{by (12) and (13),} \\
&= \int_0^T \Big\langle (Q_t^{\leftarrow}-\tilde{Q}_t^{\leftarrow})^\top p_t,\phi_t\Big\rangle\, dt \\
&= \int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \phi(t,\mathbf{y})\left[(Q_t^{\leftarrow}-\tilde{Q}_t^{\leftarrow})^\top p_t\right](\mathbf{y}) dt \\
&= \int_0^T \sum_{\mathbf{y}\in\mathcal{X}}\sum_{\mathbf{x}\in\mathcal{X}} \phi(t,\mathbf{y})\left[(Q_t^{\leftarrow}-\tilde{Q}_t^{\leftarrow})\right](\mathbf{x},\mathbf{y}) p_t(\mathbf{x}) dt \\
&= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}}\sum_{\mathbf{y}\in\mathcal{X}} \phi(t,\mathbf{y})\left[(Q_t^{\leftarrow}-\tilde{Q}_t^{\leftarrow})\right](\mathbf{x},\mathbf{y}) p_t(\mathbf{x}) dt, \quad \text{by Fubini's theorem,} \\
&= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} p_t(\mathbf{x}) \sum_{\mathbf{y}\in\mathcal{X}} \phi(t,\mathbf{y})\left[(Q_t^{\leftarrow}-\tilde{Q}_t^{\leftarrow})\right](\mathbf{x},\mathbf{y}) dt \\
&= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} p_t(\mathbf{x}) \sum_{\mathbf{y}\in\mathcal{X}} (\phi(t,\mathbf{y})-\phi(t,\mathbf{x}))\left[(Q_t^{\leftarrow}-\tilde{Q}_t^{\leftarrow})\right](\mathbf{x},\mathbf{y}) dt,
\end{aligned}$$

where we used that $Q_t^{\leftarrow}$ and $\tilde{Q}_t^{\leftarrow}$'s rows sum to zero,

$$= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} p_t(\mathbf{x}) \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} (\phi(t,\mathbf{y})-\phi(t,\mathbf{x}))\left[(Q_t^{\leftarrow}-\tilde{Q}_t^{\leftarrow})\right](\mathbf{x},\mathbf{y}) dt,$$

as $\phi(t,\mathbf{y}=\mathbf{x})-\phi(t,\mathbf{x})=0$,

$$= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}}\sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})\,(\phi(t,\mathbf{y})-\phi(t,\mathbf{x}))\left[(Q_t^{\leftarrow}-\tilde{Q}_t^{\leftarrow})\right](\mathbf{x},\mathbf{y}) dt$$

$$= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t,\mathbf{y}) - \phi(t,\mathbf{x})\right) Q_t(\mathbf{y},\mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt, \quad \text{by (24) and (10).}$$

Thus, we have

$$\begin{aligned}
\langle \lambda_0, \psi\rangle - \langle \lambda_T, \phi_T\rangle &= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t,\mathbf{y}) - \phi(t,\mathbf{x})\right) Q_t(\mathbf{y},\mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt \\
\langle \lambda_0, \psi\rangle &= \underbrace{\langle \lambda_T, \phi_T\rangle}_{①} + \underbrace{\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t,\mathbf{y}) - \phi(t,\mathbf{x})\right) Q_t(\mathbf{y},\mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt}_{②} .
\end{aligned} \tag{35}$$

For the first term ①, it can be bounded as

$$\begin{aligned}
\langle \lambda_T, \phi_T\rangle &= \langle p_T - p_{\text{base}}, \phi_T\rangle \\
&= \sum_{\mathbf{x}\in\mathcal{X}} \left[p_T - p_{\text{base}}\right](\mathbf{x}) \cdot \phi_T(\mathbf{x}) \\
&\leq \|\phi_T\|_\infty \sum_{\mathbf{x}} \left|p_T - p_{\text{base}}\right|(\mathbf{x}) \\
&\leq \|\psi\|_\infty \sum_{\mathbf{x}\in\mathcal{X}} \left|p_T - p_{\text{base}}\right|(\mathbf{x}), \quad \text{by Lemma E.1,} \\
&= 2\|\psi\|_\infty \mathrm{TV}(p_T, p_{\text{base}}).
\end{aligned}$$

For the second term ②, we bound it differently, depending on the choice of the rate matrix Q_t^{tok} in (5).

Case 1 (Bound under masked rate using $\mathcal{L}_{\text{SE}}$):

$$\begin{aligned}
&\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t,\mathbf{y}) - \phi(t,\mathbf{x})\right) Q_t(\mathbf{y},\mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt \\
&\leq 2\|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) p_t(\mathbf{y}) dt\right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}}, \quad \text{by Lemma E.6,} \\
&= 2\|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}:\mathbf{x}\neq\mathbf{y}\\ Q_t(\mathbf{y},\mathbf{x})\neq 0}} Q_t(\mathbf{y},\mathbf{x}) p_t(\mathbf{y}) dt\right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}} \\
&= 2\|\psi\|_\infty \left(\int_0^T \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{m}}} \sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})} Q_t^{\text{tok}}(y^i, [\text{MASK}]) p_t(\mathbf{y}) dt\right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}},
\end{aligned}$$

where we used the definition of Q_t (2) under masked transition (3), $\mathbf{m} = [[\text{MASK}], ..., [\text{MASK}]]$ is the sequence of all mask tokens,

$$\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y}) = \left\{ \mathbf{x} \in \mathcal{X} \;\middle|\; \exists i \in [d], \mathbf{x}^{\setminus i} = \mathbf{y}^{\setminus i}, x^i = [\text{MASK}], y^i \neq [\text{MASK}] \right\}$$

is the successor set of $\mathbf{y}$, and i denotes the unique index where $\mathbf{x}$ and $\mathbf{y}$ differ,

$$\begin{aligned}
&\leq 2\|\psi\|_\infty \left(\int_0^T \sum_{\substack{\mathbf{y}\in\mathcal{X} \\ \mathbf{y}\neq\mathbf{m}} } \underbrace{\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})}}_{\leq d \text{ terms}} \beta(t) p_t(\mathbf{y}) dt \right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s}) \right)^{\frac{1}{2}} \\
&\leq 2\|\psi\|_\infty \sqrt{2d} \left(\int_0^T \sum_{\substack{\mathbf{y}\in\mathcal{X} \\ \mathbf{y}\neq\mathbf{m}}} \beta(t) p_t(\mathbf{y}) dt \right)^{\frac{1}{2}} \left(\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s}) \right)^{\frac{1}{2}} \\
&\leq 2\|\psi\|_\infty \sqrt{2d} \left(\int_0^T \beta(t)(1 - p_t(\mathbf{m})) dt \right)^{\frac{1}{2}} \left(\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s}) \right)^{\frac{1}{2}} \\
&= 2\|\psi\|_\infty \sqrt{2d} \sqrt{\|\beta(1 - p(\mathbf{m}))\|_1} \sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})},
\end{aligned}$$

where in the last step, we used $\left(\int_0^T \beta(t)(1 - p_t(\mathbf{m}))dt \right)^{\frac{1}{2}} := \sqrt{\|\beta(1 - p(\mathbf{m}))\|_1}$. Thus, using the bounds, (35) can be evaluated as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p_0}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})] = \langle \lambda_0, \psi \rangle = ① + ② \\
&|\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(\text{TV}(p_T, p_{\text{base}}) + \sqrt{2d} \sqrt{\|\beta(1 - p(\mathbf{m}))\|_1} \sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})} \right) \\
&|\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\|\beta(1 - p(\mathbf{m}))\|_1} \sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})} \right),
\end{aligned}$$

where we used Lemma E.14 in the last step.

Case 2(a) (Bound under uniform rate using $\mathcal{L}_{\text{SE}}$):

$$\begin{aligned}
&\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X} \\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t,\mathbf{y}) - \phi(t,\mathbf{x}) \right) Q_t(\mathbf{y},\mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}} \right) dt \\
&\leq 2\|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X} \\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{y}) Q_t(\mathbf{y},\mathbf{x}) dt \right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s}) \right)^{\frac{1}{2}}, \quad \text{by Lemma E.6,} \\
&\leq 2\|\psi\|_\infty \sqrt{d} \sqrt{\|\beta\|_1} \sqrt{2} \sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})}, \quad \text{by Lemma E.2.} \\
&= 2\|\psi\|_\infty \sqrt{2d} \sqrt{\|\beta\|_1} \sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})}.
\end{aligned}$$

Thus, using the bounds, (35) can be evaluated as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p_0}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})] = \langle \lambda_0, \psi \rangle = ① + ② \\
&|\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(\text{TV}(p_T, p_{\text{base}}) + \sqrt{2d} \sqrt{\|\beta\|_1} \sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})} \right) \\
&|\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\|\beta\|_1} \sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})} \right),
\end{aligned}$$

where we used Theorem F.4 in the last step.

Case 2(b) (Bound under uniform rate using $\mathcal{L}_{\mathrm{WSM}}$):

$$\begin{aligned}
&\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})\left(\phi(t,\mathbf{y})-\phi(t,\mathbf{x})\right) Q_t(\mathbf{y},\mathbf{x})\left(s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt \\
&\leq 2\|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y},\mathbf{x}) dt\right)^{\frac{1}{2}} \cdot \sqrt{2}\sqrt{\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})}, \quad \text{by Lemma E.9,} \\
&\leq 2\|\psi\|_\infty \sqrt{d}\sqrt{\|\beta\|_1}\sqrt{2}\sqrt{\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})}, \quad \text{by Lemma E.2.} \\
&= 2\|\psi\|_\infty \sqrt{2d}\sqrt{\|\beta\|_1}\sqrt{\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})}.
\end{aligned}$$

Thus, using the bounds, (35) can be evaluated as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p_0}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})] = \langle\lambda_0,\psi\rangle = ① + ② \\
&|\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(\mathrm{TV}(p_T,p_{\mathrm{base}}) + \sqrt{2d}\sqrt{\|\beta\|_1}\sqrt{\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})}\right) \\
&|\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\|\beta\|_1}\sqrt{\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})}\right),
\end{aligned}$$

where we used Theorem F.4 in the last step.

Case 3(a) (Bound under general rate using $\mathcal{L}_{\mathrm{SE}}$):

$$\begin{aligned}
&\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})\left(\phi(t,\mathbf{y})-\phi(t,\mathbf{x})\right) Q_t(\mathbf{y},\mathbf{x})\left(s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt \\
&\leq 2\|\psi\|_\infty \left(\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\mathrm{out}}(\mathbf{x})\right] dt\right)^{\frac{1}{2}} \sqrt{2}\left(\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}}, \quad \text{by Corollary E.7.}
\end{aligned}$$

Thus, using the bounds, (35) can be evaluated as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p_0}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})] = \langle\lambda_0,\psi\rangle = ① + ② \\
&|\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(\mathrm{TV}(p_T,p_{\mathrm{base}}) + \sqrt{2}\sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}[\mu_t^{\mathrm{out}}(\mathbf{x})]\, dt}\sqrt{\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\mathcal{L}_3(s,\tilde{s})}\right).
\end{aligned}$$

Case 3(b) (Bound under general rate using $\mathcal{L}_{\mathrm{WSM}}$):

$$\begin{aligned}
&\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})\left(\phi(t,\mathbf{y})-\phi(t,\mathbf{x})\right) Q_t(\mathbf{y},\mathbf{x})\left(s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt \\
&\quad\leq\ 2\sqrt{2}\,\|\psi\|_\infty \left(\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\mathrm{in}}(\mathbf{x})\right] dt\right)^{\frac{1}{2}} \sqrt{\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})}, \quad \text{by Corollary E.10.}
\end{aligned}$$

Thus, using the bounds, (35) can be evaluated as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p_0}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})] = \langle\lambda_0,\psi\rangle = ① + ② \\
&|\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(\mathrm{TV}(p_T,p_{\mathrm{base}}) + \sqrt{2}\sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\mathrm{in}}(\mathbf{x})\right] dt}\sqrt{\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})}\right).
\end{aligned}$$

□

D.2 Proof of Corollary D.1

Corollary D.1 (TV Bounds for Score-based Discrete Diffusion)**.** *Under the same assumptions as Lemma 3.2, the total variation (TV) distance between the true data distribution p_{data} and the generated data distribution $\tilde{p}_0$ satisfies*

Case 1 (Masked rate; $Q_t^{\text{tok}} = Q_t^{\text{masked}}$ (3)):

$$\text{TV}(p_{\text{data}}, \tilde{p}_0) \leq \underbrace{de^{-\|\beta\|_1}}_{\textit{initialization error}} + \sqrt{2d}\sqrt{\|\beta(1-p(\mathbf{m}))\|_1} \underbrace{\sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})}}_{\textit{training error}}, \tag{36}$$

Case 2 (Uniform rate; $Q_t^{\text{tok}} = Q_t^{\text{uniform}}$ (3)):

$$\text{TV}(p_{\text{data}}, \tilde{p}_0) \leq \underbrace{de^{-\|\beta\|_1}}_{\textit{initialization error}} + \sqrt{2d}\sqrt{\|\beta\|_1} \underbrace{\sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})}}_{\textit{training error}}, \textit{ and} \tag{37}$$

$$\text{TV}(p_{\text{data}}, \tilde{p}_0) \leq \underbrace{de^{-\|\beta\|_1}}_{\textit{initialization error}} + \sqrt{2d}\sqrt{\|\beta\|_1} \underbrace{\sqrt{\mathcal{L}_{\text{WSM}}(s,\tilde{s})}}_{\textit{training error}}, \tag{38}$$

Case 3 (General rate):

$$\text{TV}(p_{\text{data}}, \tilde{p}_0) \leq \underbrace{\text{TV}(p_T, p_{\text{base}})}_{\textit{initialization error}} + \sqrt{2}\sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}[\mu_t^{\text{out}}(\mathbf{x})]\, dt} \underbrace{\sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})}}_{\textit{training error}}, \textit{ and} \tag{39}$$

$$\text{TV}(p_{\text{data}}, \tilde{p}_0) \leq \underbrace{\text{TV}(p_T, p_{\text{base}})}_{\textit{initialization error}} + \sqrt{2}\sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\text{in}}(\mathbf{x})\right]\, dt} \underbrace{\sqrt{\mathcal{L}_{\text{WSM}}(s,\tilde{s})}}_{\textit{training error}}. \tag{40}$$

Proof of Corollary D.1. We demonstrate the derivations for the bounds (36) and (38). The other bounds can be derived similarly. From Lemma 3.2, we have

$$|\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\|\beta(1-p(\mathbf{m}))\|_1}\sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})} \right),$$

and

$$|\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\|\psi\|_\infty \left(de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\|\beta\|_1}\sqrt{\mathcal{L}_{\text{WSM}}(s,\tilde{s})} \right),$$

respectively, for masked and uniform rates. The TV distance can be obtained by taking supremum on the left hand side of the equations over the set of ψ satisfying $\|\psi\|_\infty \leq 1$. Specifically, the supremum is attained at $\psi = \text{sign}(p_{\text{data}} - \tilde{p}_0)$ (satisfying $\|\psi\|_\infty = 1$). Thus, we have

$$\begin{aligned} 2\text{TV}(p_{\text{data}}, \tilde{p}_0) &= \sup_{\psi:\|\psi\|_\infty \leq 1} |\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \\ &\leq 2\underbrace{\|\psi\|_\infty}_{=1} \left(de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\|\beta(1-p(\mathbf{m}))\|_1}\sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})} \right), \end{aligned}$$

and

$$2\text{TV}(p_{\text{data}}, \tilde{p}_0) = \sup_{\psi:\|\psi\|_\infty \leq 1} |\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\psi(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim \tilde{p}_0}[\psi(\mathbf{x})]| \leq 2\underbrace{\|\psi\|_\infty}_{=1} \left(de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\|\beta\|_1}\sqrt{\mathcal{L}_{\text{WSM}}(s,\tilde{s})} \right),$$

respectively. Dividing both sides by 2, we obtain,

$$\text{TV}(p_{\text{data}}, \tilde{p}_0) \leq de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\|\beta(1-p(\mathbf{m}))\|_1}\sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})},$$

and

$$\text{TV}(p_{\text{data}}, \tilde{p}_0) \leq de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\|\beta\|_1}\sqrt{\mathcal{L}_{\text{WSM}}(s,\tilde{s})},$$

respectively, as desired. □

D.3 Proof of Theorem 3.1

Theorem 3.1 (IPM Bounds for Score-based Discrete Diffusion)**.** *Under Assumption 1, for any function class $\Psi = \{\psi : \mathcal{X} \to \mathbb{R}\}$, the IPM γ_Ψ (defined in* (11)*) between the data distribution p_{data} and the generated distribution $\tilde{p}_0$ satisfies the following. For bounds involving $\mathcal{L}_{\text{SE}}$, we additionally assume $\tilde{s}_t(\mathbf{x})_{\mathbf{y}} \in [s_t(\mathbf{x})_{\mathbf{y}}/2,\, 3s_t(\mathbf{x})_{\mathbf{y}}/2]$ for all $(t, \mathbf{x}, \mathbf{y})$ in the support of $p_t Q_t$, a condition that holds when the score estimator is reasonably accurate.*
Case 1 (Masked rate; $Q_t^{\text{tok}} = Q_t^{\text{masked}}$ (3)):

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \le 2C_\Psi \left(\underbrace{d e^{-\|\beta\|_1}}_{\textit{initialization error}} + \sqrt{2d}\sqrt{\|\beta(1-p(\mathbf{m}))\|_1} \underbrace{\sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})}}_{\textit{training error}} \right), \tag{14}$$

where d is the sequence length, $\|\beta\|_1 := \int_0^T \beta(t)dt$, $\mathbf{m} :=$ [MASK], ..., [MASK] is the sequence of all mask tokens, $\|\beta(1-p(\mathbf{m}))\|_1 := \int_0^T \beta(t)(1-p_t(\mathbf{m}))dt$, $p_t(\mathbf{m})$ is the marginal probability of $\mathbf{m}$ at time t, $\mathcal{L}_{\text{SE}}$ is the score entropy loss (8)*, and $\mathcal{L}_3 = \mathcal{O}(|s-\tilde{s}|^3)$ is a cubic correction term with scaling **independent of the vocabulary size** S; its exact formulation is provided in Appendix D.1;*
Case 2 (Uniform rate; $Q_t^{\text{tok}} = Q_t^{\text{uniform}}$ (3)):

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \le 2C_\Psi \left(\underbrace{d e^{-\|\beta\|_1}}_{\textit{initialization error}} + \sqrt{2d}\sqrt{\|\beta\|_1} \underbrace{\sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})}}_{\textit{training error}} \right), \textit{ and} \tag{15}$$

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \le 2C_\Psi \left(\underbrace{d e^{-\|\beta\|_1}}_{\textit{initialization error}} + \sqrt{2d}\sqrt{\|\beta\|_1} \underbrace{\sqrt{\mathcal{L}_{\text{WSM}}(s,\tilde{s})}}_{\textit{training error}} \right), \tag{16}$$

where $\mathcal{L}_{\text{WSM}}$ is the weighted score-matching loss defined in (7)*;*
Case 3 (General rate):

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \le 2C_\Psi \left(\underbrace{\text{TV}(p_T, p_{\text{base}})}_{\textit{initialization error}} + \sqrt{2}\sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}[\mu_t^{\text{out}}(\mathbf{x})]\, dt} \underbrace{\sqrt{\mathcal{L}_{\text{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}_3(s,\tilde{s})}}_{\textit{training error}} \right), \tag{17}$$

and

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_0) \le 2C_\Psi \left(\underbrace{\text{TV}(p_T, p_{\text{base}})}_{\textit{initialization error}} + \sqrt{2}\sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\text{in}}(\mathbf{x})\right] dt} \underbrace{\sqrt{\mathcal{L}_{\text{WSM}}(s,\tilde{s})}}_{\textit{training error}} \right), \tag{18}$$

where $\mu_t^{\text{out}}(\mathbf{x}) := \sum_{\mathbf{y}\neq\mathbf{x}} Q_t(\mathbf{x},\mathbf{y})$ and $\mu_t^{\text{in}}(\mathbf{x}) := \sum_{\mathbf{y}\neq\mathbf{x}} Q_t(\mathbf{y},\mathbf{x})$ denote the exit rate from and entrance rate to state $\mathbf{x}$ at time t, respectively. The dependence on the IPM enters only through the single constant C_Ψ. The list of common IPMs and their constant C_Ψ are summarized in Table 2; the result extends straightforwardly to all other IPMs.

Proof. The bounds are derived similarly to Corollary D.1. Specifically, we take the supremum in Lemma 3.2 over a corresponding admissible function class Ψ. The left-hand side of the bounds will become integral probability metrics (for total variation, it equals $2 \times$ TV). And the term $\|\psi\|_\infty$ becomes the coefficient C_Ψ; see Table 2 for a list of common IPMs and their coefficients C_Ψ. □

D.4 Proof of Corollary 3.3

Corollary 3.3 (Error Bounds under Early Stopping)**.** *Under the same assumptions as Theorem 3.1, for any function class $\Psi = \{\psi : \mathcal{X} \to \mathbb{R}\}$, the IPM γ_Ψ (defined in* (11)*) between the data distribution p_{data} and the generated distribution $\tilde{p}_\delta$, obtained under early-stopping time $\delta > 0$, satisfies*

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_\delta) \;\le\; 2\,C_\Psi \Big(\underbrace{\mathcal{E}_\delta}_{\textit{early-stopping error}} + \underbrace{\mathcal{I}}_{\textit{init. error}} + \sqrt{2\,\mathcal{C}} \underbrace{\sqrt{\mathcal{L}_\delta}}_{\textit{early-stopping train. error}} \Big). \tag{20}$$

Table 5: Examples of common IPMs, their corresponding function class, and the associated constants for Theorem 3.1, Corollary 3.3, and Corollary 3.4. Here, W_1 and MMD represent the Wasserstein-1 distance and maximum mean discrepancy, respectively. **Our results extend straightforwardly and universally to every IPM.**

IPM	Function class Ψ	C_Ψ
Total variation (TV)	$\|\psi\|_\infty \le 1$	$1/2$
Per-position TV	$\psi(\mathbf{x}) = f(x^i),\ i \in [d],\ \|f\|_\infty \le 1$	$1/2$
Block / k-gram TV	ψ depends on k positions, $\|\psi\|_\infty \le 1$	$1/2$
Bounded Lipschitz	$\|\psi\|_\infty \le 1,\ \mathrm{Lip}_d(\psi) \le 1$	1
$W_1^{d_H}$ (Hamming distance)	$\mathrm{Lip}_{d_H}(\psi) \le 1$	$d/2$
$W_1^{d_{\mathrm{embed}}}$	$\mathrm{Lip}_{d_{\mathrm{embed}}}(\psi) \le 1$	$\mathrm{diam}_{d_{\mathrm{embed}}}(\mathcal{X})/2$
MMD (kernel $K \le 1$)	$\|\psi\|_{\mathcal{H}_K} \le 1$	1
MMD (general bounded K)	$\|\psi\|_{\mathcal{H}_K} \le 1$	$\sup_{\mathbf{x}} \sqrt{K(\mathbf{x},\mathbf{x})}$

Here, $(\mathcal{I}, \mathcal{C})$ are reported in Table 3. The early stopping error $\mathcal{E}_\delta$ is given by $d\int_0^\delta \beta(t)dt$ for the masked and uniform rates and $\mathrm{TV}(p_0, p_\delta)$ for a general rate. The early-stopping training error $\mathcal{L}_\delta$ is also reported in Table 3, except that the time integration of the training losses start at time δ rather than 0. These δ-truncated losses are used in practice to avoid numerical instability. See Table 2 for a common list of IPM and their constant C_Ψ. The results extend to all other IPMs.

Proof. We have, for any IPM γ_Ψ

$$\gamma_\Psi(p_{\mathrm{data}}, \tilde{p}_\delta) = \gamma_\Psi(p_0, \tilde{p}_\delta) \le \gamma_\Psi(p_0, p_\delta) + \gamma_\Psi(p_\delta, \tilde{p}_\delta), \quad \text{by triangle inequality.} \tag{41}$$

The first term in the right-hand-side of (41) is **the early-stopping error**. For both masked and uniform rates, we have

$$\begin{aligned}
\gamma_\Psi(p_0, p_\delta) &= \sup_{\psi\in\Psi} \left|\mathbb{E}_{\mathbf{x}\sim p_0}\left[\psi(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim p_\delta}\left[\psi(\mathbf{x})\right]\right| \\
&= \sup_{\psi\in\Psi} \left|\sum_{\mathbf{x}\in\mathcal{X}} \left(p_0(\mathbf{x}) - p_\delta(\mathbf{x})\right)\psi(\mathbf{x})\right| \\
&\le \sup_{\psi\in\Psi} \left[\|\psi\|_\infty \sum_{\mathbf{x}\in\mathcal{X}} |p_0(\mathbf{x}) - p_\delta(\mathbf{x})|\right] \\
&= 2C_\Psi \mathrm{TV}(p_0, p_\delta), \quad \text{where } C_\Psi \text{ is given in Table 2,} \\
&\le 2C_\Psi \left[d\int_0^\delta \beta(t)\, dt\right],
\end{aligned}$$

where in the last step, we used Lemma G.1 for the uniform rate and Lemma G.2 for the masked rate.

The second term on the right-hand side of (41) compares the true and generated marginals at time δ. Since the generative reverse process is run only on $[\delta, T]$, the proofs of Lemma 3.2 and Theorem 3.1 apply verbatim with the time interval $[0, T]$ replaced by $[\delta, T]$, yielding the same bound in terms of the δ-truncated losses $\mathcal{L}^\delta_{\mathrm{SE}}$, $\mathcal{L}^\delta_{\mathrm{WSM}}$, $\mathcal{L}^\delta_3$. Similarly, the time integrals of β in their coefficients are also taken over $[\delta, T]$.

Thus, we have, for the masked rate

$$\begin{aligned}
\gamma_\Psi(p_{\mathrm{data}}, \tilde{p}_\delta) &\le \gamma_\Psi(p_0, p_\delta) + \gamma_\Psi(p_\delta, \tilde{p}_\delta) \\
&\le 2C_\Psi \left[d\int_0^\delta \beta(t)\, dt + de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\int_\delta^T \beta(t)(1 - p_t(\mathbf{m}))dt}\sqrt{\mathcal{L}^\delta_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\mathcal{L}^\delta_3(s,\tilde{s})}\right] \\
&\le 2C_\Psi \left[d\int_0^\delta \beta(t)\, dt + de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\int_0^T \beta(t)(1 - p_t(\mathbf{m}))dt}\sqrt{\mathcal{L}^\delta_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\mathcal{L}^\delta_3(s,\tilde{s})}\right]
\end{aligned}$$

$$= 2C_\Psi \left[d \int_0^\delta \beta(t)\, dt + d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\|\beta(1-p(\mathbf{m}))\|_1} \sqrt{\mathcal{L}^\delta_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}^\delta_3(s,\tilde{s})} \right].$$

Here, in the second last step, we upper bound the time integral over $[\delta, T]$ by $[0, T]$ (since the integrand $\beta(t)(1 - p_t(\mathbf{m}))$ is nonnegative), allowing us to represent the bound using established notation concisely in the last step.

The bounds for the uniform rate follow similarly as

$$\begin{aligned}
\gamma_\Psi(p_{\mathrm{data}}, \tilde{p}_\delta) &\le \gamma_\Psi(p_0, p_\delta) + \gamma_\Psi(p_\delta, \tilde{p}_\delta) \\
&\le 2C_\Psi \left[d \int_0^\delta \beta(t)\, dt + d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\int_\delta^T \beta(t) dt} \sqrt{\mathcal{L}^\delta_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}^\delta_3(s,\tilde{s})} \right] \\
&\le 2C_\Psi \left[d \int_0^\delta \beta(t)\, dt + d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\int_0^T \beta(t) dt} \sqrt{\mathcal{L}^\delta_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}^\delta_3(s,\tilde{s})} \right] \\
&= 2C_\Psi \left[d \int_0^\delta \beta(t)\, dt + d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\|\beta\|_1} \sqrt{\mathcal{L}^\delta_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}^\delta_3(s,\tilde{s})} \right],
\end{aligned}$$

and

$$\begin{aligned}
\gamma_\Psi(p_{\mathrm{data}}, \tilde{p}_\delta) &\le \gamma_\Psi(p_0, p_\delta) + \gamma_\Psi(p_\delta, \tilde{p}_\delta) \\
&\le 2C_\Psi \left[d \int_0^\delta \beta(t)\, dt + d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\int_\delta^T \beta(t) dt} \sqrt{\mathcal{L}^\delta_{\mathrm{WSM}}(s,\tilde{s})} \right] \\
&\le 2C_\Psi \left[d \int_0^\delta \beta(t)\, dt + d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\int_0^T \beta(t) dt} \sqrt{\mathcal{L}^\delta_{\mathrm{WSM}}(s,\tilde{s})} \right] \\
&= 2C_\Psi \left[d \int_0^\delta \beta(t)\, dt + d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\|\beta\|_1} \sqrt{\mathcal{L}^\delta_{\mathrm{WSM}}(s,\tilde{s})} \right].
\end{aligned}$$

The bounds for a general rate also follow similarly as

$$\begin{aligned}
\gamma_\Psi(p_{\mathrm{data}}, \tilde{p}_\delta) &\le \gamma_\Psi(p_0, p_\delta) + \gamma_\Psi(p_\delta, \tilde{p}_\delta) \\
&\le 2C_\Psi \left[\mathrm{TV}(p_0, p_\delta) + \mathrm{TV}(p_T, p_{\mathrm{base}}) + \sqrt{2} \sqrt{\int_\delta^T \mathbb{E}_{\mathbf{x}\sim p_t}[\mu_t^{\mathrm{out}}(\mathbf{x})]\, dt} \sqrt{\mathcal{L}^\delta_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}^\delta_3(s,\tilde{s})} \right] \\
&\le 2C_\Psi \left[\mathrm{TV}(p_0, p_\delta) + \mathrm{TV}(p_T, p_{\mathrm{base}}) + \sqrt{2} \sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}[\mu_t^{\mathrm{out}}(\mathbf{x})]\, dt} \sqrt{\mathcal{L}^\delta_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{2}{3}\,\mathcal{L}^\delta_3(s,\tilde{s})} \right],
\end{aligned}$$

and

$$\begin{aligned}
\gamma_\Psi(p_{\mathrm{data}}, \tilde{p}_\delta) &\le \gamma_\Psi(p_0, p_\delta) + \gamma_\Psi(p_\delta, \tilde{p}_\delta) \\
&\le 2C_\Psi \left[\mathrm{TV}(p_0, p_\delta) + \mathrm{TV}(p_T, p_{\mathrm{base}}) + \sqrt{2} \sqrt{\int_\delta^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\mathrm{in}}(\mathbf{x})\right]\, dt} \sqrt{\mathcal{L}^\delta_{\mathrm{WSM}}(s,\tilde{s})} \right] \\
&\le 2C_\Psi \left[\mathrm{TV}(p_0, p_\delta) + \mathrm{TV}(p_T, p_{\mathrm{base}}) + \sqrt{2} \sqrt{\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\mathrm{in}}(\mathbf{x})\right]\, dt} \sqrt{\mathcal{L}^\delta_{\mathrm{WSM}}(s,\tilde{s})} \right].
\end{aligned}$$

□

E Proof of Auxiliary Lemmas For Main Theorems

We state and prove the auxiliary lemmas.

E.1 Lemmas for the Adjoint Equation

Lemma E.1 (Sup-norm Non-expansiveness Under CTMC Generators)**.** *Suppose* $\|\psi\|_\infty < \infty$*, then the solution to the Kolmogorov backward equation* (13) ϕ_t *satisfies*

$$\|\phi_t\|_\infty \le \|\psi\|_\infty \quad \textit{for all } t \in [0, T]. \tag{42}$$

Proof. We provide two proofs here for completeness. The first direction is using a Feynman-Kac representation, which is short and clean. The second direction is to mimic the maximum principle used for parabolic partial differential equations (PDEs), which is our original intuition in deriving this result.

First Direction (Feynman-Kac representation): The Feynman–Kac representation gives $\phi_t(\mathbf{x}) = \mathbb{E}^{\mathbf{x}}[\psi(\mathbf{Y}_t)]$, where $(\mathbf{Y}_s)$ is the CTMC with generator $\tilde{Q}_t^{\leftarrow}$ started at $\mathbf{Y}_0 = \mathbf{x}$. Hence $|\phi_t(\mathbf{x})| \le \mathbb{E}^{\mathbf{x}}[|\psi(\mathbf{Y}_t)|] \le \|\psi\|_\infty$.

Second Direction (Maximum principle): Recall that the Kolmogorov backward equation is given by

$$\frac{d\phi_t}{dt} = \tilde{Q}_t^{\leftarrow}\phi_t, \quad \phi_0(\mathbf{x}) = \psi(\mathbf{x}), \quad \text{for } t \in [0, T].$$

For any $\epsilon > 0$, construct the function

$$\phi_t^\epsilon(\mathbf{x}) := \phi_t(\mathbf{x}) - \epsilon t; \quad \text{in vector notation: } \phi_t^\epsilon := \phi_t - \epsilon t\mathbf{1}.$$

Evaluating its time derivative gives

$$\begin{aligned}\frac{d\phi_t^\epsilon}{dt} &= \frac{d}{dt}(\phi_t - \epsilon t\mathbf{1}) \\ &= \tilde{Q}_t^{\leftarrow}\phi_t - \epsilon\mathbf{1} \\ &= \tilde{Q}_t^{\leftarrow}(\phi_t^\epsilon + \epsilon t\mathbf{1}) - \epsilon\mathbf{1} \\ &= \tilde{Q}_t^{\leftarrow}\phi_t^\epsilon - \epsilon\mathbf{1},\end{aligned} \tag{43}$$

as $\tilde{Q}_t^{\leftarrow}$'s rows sum to 0.

Suppose that ϕ_t^ϵ attains its maximum at $t_0 \in (0, T]$ and $\mathbf{x}_0$, then we have the following two properties:

1. $$\frac{d}{dt}\phi_{t_0}^\epsilon(\mathbf{x}_0) \ge 0, \tag{44}$$

   as $\frac{d}{dt}\phi_{t_0}^\epsilon(\mathbf{x}_0) = 0$ if $t_0 < T$ and $\frac{d}{dt}\phi_{t_0}^\epsilon(\mathbf{x}_0) \ge 0$ if $t_0 = T$,

2. $$[\tilde{Q}_{t_0}^{\leftarrow}\phi_{t_0}^\epsilon](\mathbf{x}_0) \le 0, \tag{45}$$

   because

$$\begin{aligned}[\tilde{Q}_{t_0}^{\leftarrow}\phi_{t_0}^\epsilon](\mathbf{x}_0) &= \sum_{\mathbf{y}} \tilde{Q}_{t_0}^{\leftarrow}(\mathbf{x}_0, \mathbf{y})\phi_{t_0}^\epsilon(\mathbf{y}) \\ &= \sum_{\mathbf{y}\neq\mathbf{x}_0} \left[\tilde{Q}_{t_0}^{\leftarrow}(\mathbf{x}_0, \mathbf{y})\phi_{t_0}^\epsilon(\mathbf{y})\right] + \tilde{Q}_{t_0}^{\leftarrow}(\mathbf{x}_0, \mathbf{x}_0)\phi_{t_0}^\epsilon(\mathbf{x}_0) \\ &= \sum_{\mathbf{y}\neq\mathbf{x}_0} \underbrace{\tilde{Q}_{t_0}^{\leftarrow}(\mathbf{x}_0, \mathbf{y})}_{\ge 0} \underbrace{\left(\phi_{t_0}^\epsilon(\mathbf{y}) - \phi_{t_0}^\epsilon(\mathbf{x}_0)\right)}_{\le 0, \text{ as } \phi_{t_0}^\epsilon(\mathbf{x}_0) \text{ is max}} \le 0.\end{aligned}$$

The two properties together give us

$$0 \le \frac{d}{dt}\phi_{t_0}^\epsilon(\mathbf{x}_0), \quad \text{by (44)}$$

$$= \underbrace{\left[\tilde{Q}_t^{\leftarrow} \phi_t^{\epsilon}\right](\mathbf{x}_0)}_{\leq 0, \text{ by (45)}} - \epsilon, \quad \text{by (43)},$$
$$\leq -\epsilon$$
$$< 0,$$

leading to a contradiction. This implies ϕ_t^{ϵ} cannot attain its maximum at $t_0 \in (0, T]$. Thus, ϕ_t^{ϵ} must attain its maximum at $t = 0$, and we have, for all $t \in [0, T], \mathbf{x}$,

$$\phi_t(\mathbf{x}) - \epsilon t = \phi_t^{\epsilon}(\mathbf{x}) \leq \max_{\mathbf{y}} \phi_0^{\epsilon}(\mathbf{y}) = \max_{\mathbf{y}} \phi_0(\mathbf{y}) = \max_{\mathbf{y}} \psi(\mathbf{y}).$$

Since $\epsilon > 0$ is arbitrary, taking $\epsilon \to 0^+$ gives

$$\phi_t(\mathbf{x}) \leq \max_{\mathbf{y}} \psi(\mathbf{y}), \quad \text{for all } t \in [0, T], \mathbf{x}.$$

This implies

$$\|\phi_t\|_\infty \leq \|\psi\|_\infty, \quad \text{for all } t \in [0, T],$$

as desired. □

E.2 Lemmas for the Training Loss Term

Lemma E.2. *Suppose the forward process* (1) *is governed by the factorizing transition* (2) *with uniform rate* (3)*, then*

$$\left(\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{y}) Q_t(\mathbf{y}, \mathbf{x}) dt \right)^{\frac{1}{2}} = \left(\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) dt \right)^{\frac{1}{2}} \leq \sqrt{d} \sqrt{\|\beta\|_1},$$

where $\|\beta\|_1 := \int_0^T |\beta(t)| dt = \int_0^T \beta(t) dt$.

Proof. First, we derive the equality.

$$\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{y}) Q_t(\mathbf{y}, \mathbf{x}) dt = \int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{y}) \frac{\beta(t)}{S} dt, \quad \text{by (3) and (2)},$$
$$= \int_0^T \sum_{\mathbf{y} \in \mathcal{X}} p_t(\mathbf{y}) \sum_{\substack{\mathbf{x} \in \mathcal{X} \\ \mathbf{x} \neq \mathbf{y}}} \frac{\beta(t)}{S} dt, \quad \text{by Fubini's Theorem},$$
$$= \int_0^T \sum_{\mathbf{x} \in \mathcal{X}} p_t(\mathbf{x}) \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} \frac{\beta(t)}{S} dt, \quad \text{where we swapped the dummy variable } \mathbf{x} \text{ and } \mathbf{y},$$
$$= \int_0^T \sum_{\mathbf{x} \in \mathcal{X}} p_t(\mathbf{x}) \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} Q_t(\mathbf{y}, \mathbf{x}) dt, \quad \text{by (3) and (2) again},$$
$$= \int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) dt.$$

Next, we derive the upper bound.

$$\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) dt$$
$$= \int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ d_H(\mathbf{y}, \mathbf{x}) = 1}} p_t(\mathbf{x}) \frac{\beta(t)}{S} dt, \quad \text{by (3) and (2)},$$

$$= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} p_t(\mathbf{x}) \frac{\beta(t)d(S-1)}{S} dt, \quad \text{as each } \mathbf{x} \text{ has } d(S-1) \text{ neighbors } \mathbf{y} \text{ satisfying } d_H(\mathbf{y},\mathbf{x}) = 1,$$
$$= \int_0^T \frac{\beta(t)d(S-1)}{S} dt$$
$$\leq d \int_0^T \beta(t) dt$$
$$= d\|\beta\|_1,$$

taking square root on both sides yields the desired result. □

Lemma E.3. *Under the factorizing transition* (2) *with the masked transition rate* (3)*, for* $0 < t \leq T$*, and* $g_t : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$*,*

$$\sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{m}}} \sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\textit{masked}}(\mathbf{y})} p_t(\mathbf{y}) g_t(\mathbf{x},\mathbf{y}) = (1-p_t(\mathbf{m}))\mathbb{E}_{\mathbf{y}\sim p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m})}\left[\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\textit{masked}}(\mathbf{y})} g_t(\mathbf{x},\mathbf{y})\right], \tag{46}$$

where $\mathbf{m} = [\textit{MASK}], ..., [\textit{MASK}]$ *is the sequence of all mask tokens, and*

$$\mathcal{N}^{\leftarrow}_{\textit{masked}}(\mathbf{y}) = \left\{\mathbf{x}\in\mathcal{X} \,\middle|\, \exists i \in [d], \mathbf{x}^{\setminus i} = \mathbf{y}^{\setminus i}, x^i = \textit{[MASK]}, y^i \neq \textit{[MASK]}\right\}$$

is the successor set of $\mathbf{y}$*.*

Proof. We have

$$p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m}) = \frac{p_t(\mathbf{y})}{1-p_t(\mathbf{m})}, \qquad \text{for } \mathbf{y}\neq\mathbf{m},$$
$$(1-p_t(\mathbf{m}))p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m}) = p_t(\mathbf{y}), \qquad \text{for } \mathbf{y}\neq\mathbf{m},$$
$$(1-p_t(\mathbf{m}))p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m})g_t(\mathbf{x},\mathbf{y}) = p_t(\mathbf{y})g_t(\mathbf{x},\mathbf{y}), \qquad \text{for } \mathbf{y}\neq\mathbf{m},\ \mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y}),$$
$$(1-p_t(\mathbf{m}))\sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{m}}}\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})} p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m})g_t(\mathbf{x},\mathbf{y}) = \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{m}}}\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})} p_t(\mathbf{y})g_t(\mathbf{x},\mathbf{y})$$
$$(1-p_t(\mathbf{m}))\mathbb{E}_{\mathbf{y}\sim p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m})}\left[\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})} g_t(\mathbf{x},\mathbf{y})\right] = \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{m}}}\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})} p_t(\mathbf{y})g_t(\mathbf{x},\mathbf{y}),$$

swapping left-hand-side and right-hand-side, we get the desired result. □

Lemma E.4. *For* $|u| \leq 1/2$*,*

$$\frac{u^2}{2} + u + \log(1-u) \;\leq\; \frac{2|u|^3}{3}. \tag{47}$$

Proof. Define $f(u) := \frac{2|u|^3}{3} - \frac{u^2}{2} - u - \log(1-u)$ on $[-1/2, 1/2]$. We show $f(u) \geq 0$ by case analysis on the sign of u.

Case $u \in [0, 1/2]$. Here $|u|^3 = u^3$, so $f(u) = \frac{2u^3}{3} - \frac{u^2}{2} - u - \log(1-u)$ and

$$f(0) = 0,$$
$$f'(u) = 2u^2 - u - 1 + \tfrac{1}{1-u}, \qquad f'(0) = 0,$$
$$f''(u) = 4u - 1 + \tfrac{1}{(1-u)^2}, \qquad f''(0) = 0,$$
$$f'''(u) = 4 + \tfrac{2}{(1-u)^3} \;>\; 0 \quad \text{on } [0, 1/2].$$

Since $f''' > 0$ on $[0, 1/2]$, f'' is strictly increasing; combined with $f''(0) = 0$, this gives $f''(u) > 0$ for $u \in (0, 1/2]$. Hence f' is strictly increasing on $[0, 1/2]$, and $f'(0) = 0$ gives $f'(u) > 0$

for $u \in (0, 1/2]$. Hence f is strictly increasing on $[0, 1/2]$, and $f(0) = 0$ gives $f(u) > 0$ for $u \in (0, 1/2]$.

Case $u \in [-1/2, 0]$. Here $|u|^3 = -u^3$, so $f(u) = -\frac{2u^3}{3} - \frac{u^2}{2} - u - \log(1-u)$ and

$$\begin{aligned} f(0) &= 0, \\ f'(u) &= -2u^2 - u - 1 + \tfrac{1}{1-u}, & f'(0) &= 0, \\ f''(u) &= -4u - 1 + \tfrac{1}{(1-u)^2}, & f''(0) &= 0, \\ f'''(u) &= -4 + \tfrac{2}{(1-u)^3}. \end{aligned}$$

On $[-1/2, 0]$, $1 - u \in [1, 3/2]$, so $\frac{2}{(1-u)^3} \in [\frac{16}{27}, 2]$, giving $f'''(u) \le 2 - 4 = -2 < 0$. Hence f'' is strictly decreasing on $[-1/2, 0]$; combined with $f''(0) = 0$, this gives $f''(u) > 0$ for $u \in [-1/2, 0)$. Hence f' is strictly increasing on $[-1/2, 0]$, and $f'(0) = 0$ gives $f'(u) < 0$ for $u \in [-1/2, 0)$. Hence f is strictly decreasing on $[-1/2, 0]$, and $f(0) = 0$ gives $f(u) > 0$ for $u \in [-1/2, 0)$.

Combining both cases, $f(u) \ge 0$ on $[-1/2, 1/2]$, with equality only at $u = 0$. Rearranging gives (47). □

Lemma E.5 (Bregman with positive cubic remainder)**.** *Define the Bregman divergence with generating function* $\varphi(a) = a \log a - a$ *as*

$$D_\varphi(s, \tilde{s}) = \varphi(s) - \varphi(\tilde{s}) - \varphi'(\tilde{s})(s - \tilde{s}) = s \log\left(\frac{s}{\tilde{s}}\right) + \tilde{s} - s. \tag{48}$$

For $s > 0$ *and* $\tilde{s} \in [s/2, 3s/2]$,

$$\frac{s}{2}\left(1 - \frac{\tilde{s}}{s}\right)^2 \;\le\; D_\varphi(s, \tilde{s}) + \frac{2\,|s - \tilde{s}|^3}{3s^2}. \tag{49}$$

Proof. Let $u := (s - \tilde{s})/s$. The hypothesis $\tilde{s} \in [s/2, 3s/2]$ gives $|u| \le 1/2$. Since $\tilde{s}/s = 1 - u$,

$$D_\varphi(s, \tilde{s}) = -s \log(1 - u) - su.$$

Dividing (49) by $s > 0$ and rearranging gives the equivalent statement

$$\frac{u^2}{2} + u + \log(1 - u) \;\le\; \frac{2|u|^3}{3},$$

which is Lemma E.4. □

Lemma E.6 (Bounding Error by Score Entropy)**.** *Under a general transition rate* Q_t *satisfying Assumption 1, we have*

$$\begin{aligned} &\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t, \mathbf{y}) - \phi(t, \mathbf{x})\right) Q_t(\mathbf{y}, \mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt \\ &\le 2\|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{y} \in \mathcal{X}} \sum_{\substack{\mathbf{x} \in \mathcal{X} \\ \mathbf{x} \neq \mathbf{y}}} Q_t(\mathbf{y}, \mathbf{x}) p_t(\mathbf{y}) dt \right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\mathrm{SE}}(s, \tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s, \tilde{s})\right)^{\frac{1}{2}}. \end{aligned}$$

Proof. We have

$$\begin{aligned} &\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t, \mathbf{y}) - \phi(t, \mathbf{x})\right) Q_t(\mathbf{y}, \mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt \\ &= \int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t, \mathbf{y}) - \phi(t, \mathbf{x})\right) Q_t(\mathbf{y}, \mathbf{x}) s_t(\mathbf{x})_{\mathbf{y}} \left(1 - \frac{\tilde{s}_t(\mathbf{x})_{\mathbf{y}}}{s_t(\mathbf{x})_{\mathbf{y}}}\right) dt \end{aligned}$$

$$\leq \left(\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y},\mathbf{x}) s_t(\mathbf{x})_{\mathbf{y}} \left(\phi(t,\mathbf{y}) - \phi(t,\mathbf{x})\right)^2 dt \right)^{\frac{1}{2}}$$
$$\left(\underbrace{\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y},\mathbf{x}) s_t(\mathbf{x})_{\mathbf{y}} \left(1 - \frac{\tilde{s}_t(\mathbf{x})_{\mathbf{y}}}{s_t(\mathbf{x})_{\mathbf{y}}}\right)^2 dt}_{=\mathcal{L}_{\mathrm{WRSM}},\ \text{defined in (54)}} \right)^{\frac{1}{2}},$$

by Cauchy-Schwarz inequality,

$$\leq \left(\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} \cancel{p_t(\mathbf{x})} Q_t(\mathbf{y},\mathbf{x}) \frac{p_t(\mathbf{y})}{\cancel{p_t(\mathbf{x})}} \left(\phi(t,\mathbf{y}) - \phi(t,\mathbf{x})\right)^2 dt \right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}}, \quad \text{by Lemma E.12.}$$

And critically, we applied the score-cancellation technique that changes the probability mass term from $p_t(\mathbf{x})$ to $p_t(\mathbf{y})$. This removes the S-dependence of the resulting bound for masked rates. See Sections H.1 and H.2 for more details.

$$= \left(\int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) p_t(\mathbf{y}) \left(\phi(t,\mathbf{y}) - \phi(t,\mathbf{x})\right)^2 dt \right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}},$$

where in the first term, we used Fubini's Theorem,

$$\leq 2 \max_{0\leq t\leq T} \left(\|\phi_t\|_\infty\right) \left(\int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) p_t(\mathbf{y}) dt \right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}}$$

$$= 2\|\phi\|_\infty \left(\int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) p_t(\mathbf{y}) dt \right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}}$$

$$\leq 2\|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) p_t(\mathbf{y}) dt \right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}}, \quad \text{by Lemma E.1.}$$

□

Corollary E.7 (Bounding Error by Score Entropy)**.** *Under a general transition rate Q_t satisfying Assumption 1, let*

$$\mu_t^{\mathrm{out}}(\mathbf{x}) \;:=\; -Q_t(\mathbf{x},\mathbf{x}) \;=\; \sum_{\mathbf{y}\neq\mathbf{x}} Q_t(\mathbf{x},\mathbf{y})$$

*denote the **exit rate** from state $\mathbf{x}$ at time t. Then*

$$\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t,\mathbf{y}) - \phi(t,\mathbf{x})\right) Q_t(\mathbf{y},\mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right) dt$$
$$\leq\; 2\|\psi\|_\infty \left(\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\mathrm{out}}(\mathbf{x})\right] dt \right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}}.$$

Proof. By Lemma E.6, we have

$$\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})\,(\phi(t,\mathbf{y})-\phi(t,\mathbf{x}))\,Q_t(\mathbf{y},\mathbf{x})\,(s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}})\,dt$$
$$\leq 2\|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x})p_t(\mathbf{y})dt\right)^{\frac{1}{2}} \left(2\,\mathcal{L}_{\mathrm{SE}}(s,\tilde{s})+\tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s})\right)^{\frac{1}{2}}. \tag{50}$$

We first bound the integrand on the right-hand-side. Fubini's theorem gives

$$\sum_{\mathbf{x}}\sum_{\mathbf{y}\neq\mathbf{x}} Q_t(\mathbf{y},\mathbf{x})\,p_t(\mathbf{y}) \;=\; \sum_{\mathbf{y}} p_t(\mathbf{y}) \sum_{\mathbf{x}\neq\mathbf{y}} Q_t(\mathbf{y},\mathbf{x}) \;=\; \sum_{\mathbf{y}} p_t(\mathbf{y})\,\mu_t^{\mathrm{out}}(\mathbf{y}) \;=\; \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\mathrm{out}}(\mathbf{x})\right].$$

Thus, we obtain

$$\left(\int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x})p_t(\mathbf{y})dt\right)^{\frac{1}{2}} \;=\; \left(\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\mathrm{out}}(\mathbf{x})\right] dt\right)^{\frac{1}{2}}. \tag{51}$$

Combining (50) and (51) yields the claim. □

Remark E.8 (Interpretation of the prefactor)**.** *The quantity*

$$\mathcal{N}_T \;:=\; \int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\mu_t^{\mathrm{out}}(\mathbf{x})\right] dt$$

appearing in the bound is the expected total number of jumps *the forward CTMC makes on* $[0,T]$ *when initialized at* $\mathbf{X}_0 \sim p_{\mathrm{data}}$*. Indeed, the jump-counting process* $J_t := \#\{$*jumps on* $[0,t]\}$ *has compensator* $A_t = \int_0^t \mu_s^{\mathrm{out}}(\mathbf{X}_s)\,ds$*, so by the Doob–Meyer decomposition* $\mathbb{E}[J_T] = \mathbb{E}[A_T] = \mathcal{N}_T$*. The bound thus takes the form*

$$|\,\textit{approximation error}\,| \leq 2\|\psi\|_\infty \sqrt{\mathcal{N}_T}\,\sqrt{2\mathcal{L}_{\mathrm{SE}} + (4/3)\mathcal{L}_3}\,,$$

i.e. the score-matching loss is amplified by the square root of the expected dynamical activity of the forward process.

Lemma E.9 (Bounding Error by Score Matching Loss)**.** *Under a general transition rate* Q_t *satisfying Assumption 1, we have*

$$\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})\,(\phi(t,\mathbf{y})-\phi(t,\mathbf{x}))\,Q_t(\mathbf{y},\mathbf{x})\,(s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}})\,dt$$
$$\leq 2\|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x})dt\right)^{\frac{1}{2}} \cdot \sqrt{2}\sqrt{\mathcal{L}_{\mathrm{WSM}}(s,\tilde{s})}$$

Proof.

$$\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})\,(\phi(t,\mathbf{y})-\phi(t,\mathbf{x}))\,Q_t(\mathbf{y},\mathbf{x})\,(s_t(\mathbf{x})_{\mathbf{y}}-\tilde{s}_t(\mathbf{x})_{\mathbf{y}})\,dt$$
$$\leq \left(\int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x})\,(\phi(t,\mathbf{y})-\phi(t,\mathbf{x}))^2\,dt\right)^{\frac{1}{2}}$$

$$\left(\underbrace{2 \cdot \int_0^T \frac{1}{2} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}} \right)^2 dt}_{= \mathcal{L}_{\mathrm{WSM}}(s, \tilde{s}) \text{ defined in (7)}} \right)^{\frac{1}{2}},$$

by Cauchy-Schwarz inequality,

$$\leq 2 \max_{0 \leq t \leq T} \left(\|\phi_t\|_\infty \right) \left(\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) dt \right)^{\frac{1}{2}} \cdot \sqrt{2} \sqrt{\mathcal{L}_{\mathrm{WSM}}(s, \tilde{s})}$$

$$\leq 2 \|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) dt \right)^{\frac{1}{2}} \cdot \sqrt{2} \sqrt{\mathcal{L}_{\mathrm{WSM}}(s, \tilde{s})}, \quad \text{by Lemma E.1.}$$

□

Corollary E.10 (Bounding Error by Score Matching Loss)**.** *Under a general transition rate Q_t satisfying Assumption 1, let*

$$\mu_t^{\mathrm{in}}(\mathbf{x}) \; := \; \sum_{\mathbf{y} \neq \mathbf{x}} Q_t(\mathbf{y}, \mathbf{x})$$

*denote the **entrance rate** at state $\mathbf{x}$ at time t (the total instantaneous rate at which other states feed into $\mathbf{x}$). Then*

$$\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t, \mathbf{y}) - \phi(t, \mathbf{x}) \right) Q_t(\mathbf{y}, \mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}} \right) dt$$

$$\leq \; 2\sqrt{2} \, \|\psi\|_\infty \left(\int_0^T \mathbb{E}_{\mathbf{x} \sim p_t} \left[\mu_t^{\mathrm{in}}(\mathbf{x}) \right] dt \right)^{\frac{1}{2}} \sqrt{\mathcal{L}_{\mathrm{WSM}}(s, \tilde{s})}.$$

Proof. By Lemma E.9, we have

$$\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) \left(\phi(t, \mathbf{y}) - \phi(t, \mathbf{x}) \right) Q_t(\mathbf{y}, \mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}} \right) dt$$

$$\leq 2 \|\psi\|_\infty \left(\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) dt \right)^{\frac{1}{2}} \cdot \sqrt{2} \sqrt{\mathcal{L}_{\mathrm{WSM}}(s, \tilde{s})} \tag{52}$$

On the other hand, we have

$$\sum_{\mathbf{x}} \sum_{\mathbf{y} \neq \mathbf{x}} p_t(\mathbf{x}) \, Q_t(\mathbf{y}, \mathbf{x}) \; = \; \sum_{\mathbf{x}} p_t(\mathbf{x}) \sum_{\mathbf{y} \neq \mathbf{x}} Q_t(\mathbf{y}, \mathbf{x}) \; = \; \sum_{\mathbf{x}} p_t(\mathbf{x}) \, \mu_t^{\mathrm{in}}(\mathbf{x}) \; = \; \mathbb{E}_{\mathbf{x} \sim p_t} \left[\mu_t^{\mathrm{in}}(\mathbf{x}) \right].$$

Thus, we obtain

$$\int_0^T \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\substack{\mathbf{y} \in \mathcal{X} \\ \mathbf{y} \neq \mathbf{x}}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) dt \; = \; \int_0^T \mathbb{E}_{\mathbf{x} \sim p_t} \left[\mu_t^{\mathrm{in}}(\mathbf{x}) \right] dt. \tag{53}$$

Combining (52) and (53) yields the claim. □

Remark E.11 (Exit rate vs. entrance rate under uniform and masked)**.** *The exit rate μ_t^{out} and entrance rate μ_t^{in} behave very differently under the two transition rates in this work.*

Under the uniform rate (3)*, Q_t^{tok} is symmetric and*

$$\mu_t^{\mathrm{out,tok}}(x) \;=\; \mu_t^{\mathrm{in,tok}}(x) \;=\; \beta(t)\big(1-\tfrac{1}{S}\big) \quad \textit{for every } x \in [S],$$

so the two lemmas give the same scaling.

Under the masked rate (3)*, Q_t^{tok} is asymmetric and the two rates differ:*

$$\mathbb{E}_{p_t^{\mathrm{tok}}}\big[\mu_t^{\mathrm{out,tok}}(x)\big] \;=\; \beta(t)\big(1-p_t^{\mathrm{tok}}(\textit{[MASK]})\big)$$

is S-independent, whereas

$$\mathbb{E}_{p_t^{\mathrm{tok}}}\big[\mu_t^{\mathrm{in,tok}}(x)\big] \;=\; (S-1)\,\beta(t)\,p_t^{\mathrm{tok}}(\textit{[MASK]})$$

carries a factor of $S-1$ — the vocabulary size.

This is why our masked-case analysis (Lemma 3.2, Case 1) uses Lemma E.6 to expose the exit rate, while the uniform case (Case 2) can use either lemma.

Lemma E.12 (Bounding $\mathcal{L}_{\mathrm{WRSM}}$ by $\mathcal{L}_{\mathrm{SE}}$)**.** *Let*

$$\mathcal{L}_{\mathrm{WRSM}}(s,\tilde{s}) = \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x})s_t(\mathbf{x})_{\mathbf{y}} \left(1-\frac{\tilde{s}_t(\mathbf{x})_{\mathbf{y}}}{s_t(\mathbf{x})_{\mathbf{y}}}\right)^2 dt \tag{54}$$

be the weighted relative score-matching loss (WRSM), and recall the score entropy (SE) loss is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) &= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x}) \left[s_t(\mathbf{x})_{\mathbf{y}} \log\left(\frac{s_t(\mathbf{x})_{\mathbf{y}}}{\tilde{s}_t(\mathbf{x})_{\mathbf{y}}}\right) + \tilde{s}_t(\mathbf{x})_{\mathbf{y}} - s_t(\mathbf{x})_{\mathbf{y}}\right] dt \\ &= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x}) D_\varphi(s_t(\mathbf{x})_{\mathbf{y}}), \tilde{s}_t(\mathbf{x})_{\mathbf{y}}) dt.\end{aligned}$$

Suppose that the score estimator $\tilde{s}_t(\mathbf{x})_{\mathbf{y}} \in [s_t(\mathbf{x})_{\mathbf{y}}/2,\, 3s_t(\mathbf{x})_{\mathbf{y}}/2]$ for all $(t,\mathbf{x},\mathbf{y})$ in the support of $p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x})$, then under the masked transition rate ((2) and (3)),

$$\mathcal{L}_{\mathrm{WRSM}}(s,\tilde{s}) \;\le\; 2\,\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \tfrac{4}{3}\,\mathcal{L}_3(s,\tilde{s}), \tag{55}$$

where

$$\mathcal{L}_3(s,\tilde{s}) = \begin{cases} \displaystyle\int_0^T \beta(t)(1-p_t(\mathbf{m}))\mathbb{E}_{\mathbf{y}\sim p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m})}\left[\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\textit{masked}}(\mathbf{y})} \frac{|s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}|^3}{(s_t(\mathbf{x})_{\mathbf{y}})^3}\right] dt, & \text{(33a)} \\ \textit{under the masked rate,} \\ \displaystyle\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\frac{\beta(t)}{S} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ d_H(\mathbf{y},\mathbf{x})=1}} \left(\frac{|s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}|^3}{(s_t(\mathbf{x})_{\mathbf{y}})^2}\right)\right] dt, & \text{(33b)} \\ \textit{under the uniform rate,} \\ \displaystyle\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t}\left[\sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} Q_t(\mathbf{y},\mathbf{x}) \left(\frac{|s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}|^3}{(s_t(\mathbf{x})_{\mathbf{y}})^2}\right)\right] dt = \int_0^T \mathbb{E}_{\mathbf{y}\sim p_t}\left[\sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) \left(\frac{|s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}|^3}{(s_t(\mathbf{x})_{\mathbf{y}})^3}\right)\right] dt, & \text{(33c)} \\ \textit{under a general rate.} \end{cases}$$

Proof.

$$\mathcal{L}_{\mathrm{WRSM}}(s,\tilde{s}) = \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x})s_t(\mathbf{x})_{\mathbf{y}} \left(1-\frac{\tilde{s}_t(\mathbf{x})_{\mathbf{y}}}{s_t(\mathbf{x})_{\mathbf{y}}}\right)^2 dt$$

$$\leq \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x}) \cdot 2 \cdot \left(D_\varphi(s_t(\mathbf{x})_\mathbf{y}, \tilde{s}_t(\mathbf{x})_\mathbf{y}) + \frac{2\,|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{3(s_t(\mathbf{x})_\mathbf{y})^2} \right) dt, \quad \text{by Lemma E.5,}$$

$$= 2\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x}) \left(\frac{4\,|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{3(s_t(\mathbf{x})_\mathbf{y})^2} \right) dt$$

$$= 2\mathcal{L}_{\mathrm{SE}}(s,\tilde{s}) + \frac{4}{3}\mathcal{L}_3(s,\tilde{s}),$$

where we used the definition of $\mathcal{L}_3$, and its formulations under different rates are given in Lemma E.13. □

Lemma E.13 (Formulation of $\mathcal{L}_3$ under masked and uniform rates)**.** *The cubic correction term*

$$\mathcal{L}_3(s,\tilde{s}) := \int_0^T \mathbb{E}_{\mathbf{x}\sim p_t} \left[\sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} Q_t(\mathbf{y},\mathbf{x}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^2} \right) \right] dt$$

can be reformulated as:

Case 1 (Masked rate; $Q_t^{\mathrm{tok}} = Q_t^{\mathrm{masked}}$ **(3))*:***

$$\mathcal{L}_3(s,\tilde{s}) = \int_0^T \beta(t)(1-p_t(\mathbf{m}))\mathbb{E}_{\mathbf{y}\sim p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m})} \left[\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{masked}(\mathbf{y})} \frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right] dt, \tag{57}$$

Case 2 (Uniform rate; $Q_t^{\mathrm{tok}} = Q_t^{\mathrm{uniform}}$ **(3))*:***

$$\mathcal{L}_3(s,\tilde{s}) = \int_0^T \mathbb{E}_{\mathbf{x}\sim p_t} \left[\frac{\beta(t)}{S} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ d_H(\mathbf{y},\mathbf{x})=1}} \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^2} \right) \right] dt, \tag{58}$$

Case 3 (General rate):

$$\mathcal{L}_3(s,\tilde{s}) = \int_0^T \mathbb{E}_{\mathbf{y}\sim p_t} \left[\sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right) \right] dt. \tag{59}$$

Proof. **Case 3 (General rate):**
We first derive the equivalent formulation under a general rate.

$$\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t} \left[\sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} Q_t(\mathbf{y},\mathbf{x}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^2} \right) \right] dt$$

$$= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^2} \right) dt$$

$$= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} p_t(\mathbf{x})Q_t(\mathbf{y},\mathbf{x})s_t(\mathbf{x})_\mathbf{y} \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right) dt$$

$$= \int_0^T \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\substack{\mathbf{y}\in\mathcal{X}\\ \mathbf{y}\neq\mathbf{x}}} \cancel{p_t(\mathbf{x})}Q_t(\mathbf{y},\mathbf{x}) \frac{p_t(\mathbf{y})}{\cancel{p_t(\mathbf{x})}} \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right) dt$$

$$= \int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}\in\mathcal{X}\\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x})p_t(\mathbf{y}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right) dt, \quad \text{by Fubini's Theorem,}$$

$$= \int_0^T \mathbb{E}_{\mathbf{y}\sim p_t} \left[\sum_{\substack{\mathbf{x}\in\mathcal{X} \\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right) \right] dt.$$

Case 1 (Masked rate; $Q_t^{\text{tok}} = Q_t^{\text{masked}}$ (3)):
We start with the equivalent formulation under a generate rate, which we have just derived,

$$\int_0^T \mathbb{E}_{\mathbf{y}\sim p_t} \left[\sum_{\substack{\mathbf{x}\in\mathcal{X} \\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right) \right] dt$$

$$= \int_0^T \sum_{\mathbf{y}\in\mathcal{X}} \sum_{\substack{\mathbf{x}\in\mathcal{X} \\ \mathbf{x}\neq\mathbf{y}}} Q_t(\mathbf{y},\mathbf{x}) p_t(\mathbf{y}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right) dt$$

$$= \int_0^T \sum_{\substack{\mathbf{y}\in\mathcal{X} \\ \mathbf{y}\neq\mathbf{m}}} \sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})} Q_t^{\text{tok}}(y^i, m) p_t(\mathbf{y}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right) dt,$$

where we used the definition of Q_t (2) under masked transition (3), $\mathbf{m} = [\text{[MASK]}, ..., \text{[MASK]}]$ is the sequence of all mask tokens,

$$\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y}) = \left\{ \mathbf{x} \in \mathcal{X} \;\middle|\; \exists i \in [d], \mathbf{x}^{\setminus i} = \mathbf{y}^{\setminus i}, x^i = \text{[MASK]}, y^i \neq \text{[MASK]} \right\}$$

is the successor set of $\mathbf{y}$, and i denotes the unique index where $\mathbf{x}$ and $\mathbf{y}$ differ,

$$= \int_0^T \sum_{\substack{\mathbf{y}\in\mathcal{X} \\ \mathbf{y}\neq\mathbf{m}}} \sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})} \beta(t) p_t(\mathbf{y}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right) dt$$

$$= \int_0^T \beta(t)(1 - p_t(\mathbf{m})) \mathbb{E}_{\mathbf{y}\sim p_t(\mathbf{y}|\mathbf{y}\neq\mathbf{m})} \left[\sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})} \frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^3} \right] dt.$$

Case 2 (Uniform rate; $Q_t^{\text{tok}} = Q_t^{\text{uniform}}$ (3)):
In addition, we have

$$\int_0^T \mathbb{E}_{\mathbf{x}\sim p_t} \left[\sum_{\substack{\mathbf{y}\in\mathcal{X} \\ \mathbf{y}\neq\mathbf{x}}} Q_t(\mathbf{y},\mathbf{x}) \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^2} \right) \right] dt$$

$$= \int_0^T \mathbb{E}_{\mathbf{x}\sim p_t} \left[\frac{\beta(t)}{S} \sum_{\substack{\mathbf{y}\in\mathcal{X} \\ d_H(\mathbf{y},\mathbf{x})=1}} \left(\frac{|s_t(\mathbf{x})_\mathbf{y} - \tilde{s}_t(\mathbf{x})_\mathbf{y}|^3}{(s_t(\mathbf{x})_\mathbf{y})^2} \right) \right] dt, \quad \text{by the definition of the uniform rate (2) and (3),}$$

□

Lemma E.14 (Initialization Error Bound under Masked Diffusion)**.** *Under masked diffusion, the initialization error term is bounded by*

$$\text{TV}(p_T, p_{\text{base}}) \leq d e^{-\int_0^T \beta(\tau) d\tau} = d e^{-\|\beta\|_1}. \tag{60}$$

Proof. The all-mask probability $p_t(\mathbf{m}) = \alpha_t^d$ under the Kronecker-sum factorization (2) with single-token mask probability α_t. Here, α_t satisfies the ODE

$$\partial_t \alpha_t = \beta(t)(1 - \alpha_t) \quad \text{subject to} \quad \alpha_0 = 0, \tag{61}$$

derived from the masked rate (3) and with solution

$$\alpha_t = 1 - e^{-\int_0^t \beta(\tau)d\tau}, \quad \text{for } t \in [0, T]. \tag{62}$$

Thus, the initialization error satisfies

$$\mathrm{TV}(p_T, p_{\mathrm{base}}) = \mathrm{TV}(p_T, \delta_{\mathbf{m}}) = 1 - p_T(\mathbf{m}) = 1 - \alpha_T^d = 1 - \Big(1 - e^{-\int_0^T \beta(\tau)d\tau}\Big)^d \le d \cdot e^{-\int_0^T \beta(\tau)d\tau} = d e^{-\|\beta\|_1}.$$

□

F Bounding the Initialization Error Term under Uniform Rate

In this section, we derive, to the best of our knowledge, the first S-independent initialization error bound under the uniform transition rate in the context of discrete diffusion generative models. Our key novelty is a coupling technique. For better readability, we begin by restating the setup.

Setup. We consider a forward CTMC on $[S]^d$. The dynamics are *non-interacting*: each coordinate (token) $i \in [d]$ evolves independently according to the per-site generator

$$(Q_t^{\mathrm{uniform}})^\top \;=\; \beta(t)\Big(\tfrac{1}{S}\mathbf{1}_S\mathbf{1}_S^\top - I_S\Big) \;=\; \beta(t)\,(\Pi - I), \tag{63}$$

where $\Pi := \frac{1}{S}\mathbf{1}_S\mathbf{1}_S^\top$ is the rank-one projection onto the uniform distribution $\pi = \mathrm{Unif}([S])$. The joint generator is

$$\mathcal{L}_t = \sum_{i=1}^{d} Q_t^{(\mathrm{uniform}),\,i}, \tag{64}$$

with $Q_t^{(\mathrm{uniform}),\,i}$ acting on coordinate i. The initial law is $p_0 = p_{\mathrm{data}}$, the reference law is $\pi^{\otimes d}$. Set

$$\rho_t \;:=\; \exp\Big(-\int_0^t \beta(\tau)\,d\tau\Big); \quad \rho_T \;:=\; \exp\Big(-\int_0^T \beta(\tau)\,d\tau\Big). \tag{65}$$

Bounding $\mathrm{TV}(p_T,\, p_{\mathrm{base}}) = \mathrm{TV}(p_T,\, \pi^{\otimes d})$ reduces to exhibiting *any* coupling with controlled disagreement probability, by the following classical characterization of total variation.

Lemma F.1 (Coupling characterization of TV; (Levin & Peres, 2017, Proposition 4.7))**.** *For any probability measures μ, ν on a common Polish space,*

$$\mathrm{TV}(\mu, \nu) \;=\; \min_{\Gamma \in \Pi(\mu,\nu)} \Pr_{(X,Y)\sim\Gamma}\big(X \neq Y\big),$$

where $\Pi(\mu, \nu)$ is the set of couplings of μ and ν (joint laws with the prescribed marginals), and the minimum is attained (by the maximal coupling). Consequently, for any specific *coupling $\Gamma \in \Pi(\mu, \nu)$,*

$$\mathrm{TV}(\mu, \nu) \;\le\; \Pr_{\Gamma}\big(X \neq Y\big). \tag{66}$$

Per-site uniformization. Next, we consider a per-site uniformization representation of Q_t^{uniform}. The entries of the per-site generator (63) are

$$Q_t^{\mathrm{uniform}}(v, w) \;=\; \beta(t)\Big[\tfrac{1}{S} - \mathbf{1}[v = w]\Big] \;=\; \begin{cases} \beta(t)/S & w \neq v, \\ -(S-1)\,\beta(t)/S & w = v, \end{cases} \tag{67}$$

so the total exit rate from any state $v \in [S]$ is the same constant

$$|Q_t^{\mathrm{uniform}}(v, v)| \;=\; \sum_{w \neq v} Q_t^{\mathrm{uniform}}(v, w) \;=\; \frac{(S-1)\,\beta(t)}{S} \quad \text{for every } v \in [S]. \tag{68}$$

Uniformization requires only $\Lambda(t) \ge \max_v |Q_t^{\mathrm{uniform}}(v, v)| = (S-1)\beta(t)/S$. We choose

$$\Lambda(t) \;=\; \beta(t), \tag{69}$$

which is valid since $\beta(t) \geq (S-1)\beta(t)/S$ for all $S \geq 1$, with slack $\beta(t)/S$ corresponding to "self-jumps" $v \to v$ that are invisible at the generator level. With this choice the uniformization transition matrix is

$$P_\beta(t) \;:=\; I + \frac{Q_t^{\text{uniform}}}{\beta(t)} \;=\; I + (\Pi - I) \;=\; \Pi, \qquad P_\beta(v,w) = \frac{1}{S} \text{ for all } v, w \in [S]. \tag{70}$$

The transition rule is the rank-one projection: *at each Poisson ring the chain is reset to a fresh uniform sample $\xi \sim \pi$, independently of its current state*, with $1/S$ probability of an invisible self-jump.

Lemma F.2 (Uniformized construction)**.** *Let N be a Poisson process with rate $\beta(t)$, and let $(\xi_k)_{k\geq 1}$ be i.i.d.* $\mathrm{Unif}([S])$*, independent of N. Define $(X_t)_{t\geq 0}$ by $X_0 = v_0$ and, at each event time t_k of N, set $X_{t_k} = \xi_k$. Then (X_t) is the per-site CTMC with generator* (63) *started at v_0.*

Proof. Conditional on N and (ξ_k),

$$X_T \;=\; \begin{cases} v_0 & \text{if } N([0,T]) = 0, \\ \xi_{N([0,T])} & \text{otherwise.} \end{cases}$$

The first case has probability $\rho_T = e^{-\int_0^T \beta(s)ds}$; the second has probability $1 - \rho_T$ and yields an independent uniform sample. Hence the T-step kernel is

$$\Pr(X_T = w \mid X_0 = v) \;=\; \rho_T\, \mathbf{1}[v = w] + (1 - \rho_T)/S. \tag{71}$$

This matches the kernel obtained directly from (63): using $(\Pi - I)^2 = -(\Pi - I)$,

$$\exp\Big(\int_0^T Q_s^{\text{uniform}}\, ds\Big) \;=\; \exp\big(-\log \rho_T \cdot (\Pi - I)\big) \;=\; \rho_T\, I + (1 - \rho_T)\, \Pi.$$

The Markov property of (X_t) follows from the strong Markov property of N and the independence of (ξ_k). □

Remark F.3 (Joint Construction)**.** *By non-interaction, the joint CTMC on $[S]^d$ is realized by d* independent *per-site uniformized chains: independent Poisson processes $N_1, \ldots, N_d$ each at rate $\beta(t)$; for each i, an i.i.d. sequence $(\xi_k^{(i)})_{k\geq 1}$ in $[S]$, independent of everything else; at each event time of N_i, position i is reset to the next sample $\xi_k^{(i)}$. The marginal law on position i is then* (71)*; the joint law has generator $\mathcal{L}_t$ defined in* (64)*.*

Initialization Error Bound. We are now ready to state and prove, to the best of our knowledge, the first S-independent initialization error bound under uniform transition in the context of discrete diffusion generative models.

Theorem F.4. *For all $T \geq 0$,*

$$\mathrm{TV}(p_T,\, \pi^{\otimes d}) \;\leq\; d\, \exp\Big(-\int_0^T \beta(t)\, dt\Big) \;=\; d e^{-\|\beta\|_1}.$$

In particular, $\mathrm{TV}(p_T, \pi^{\otimes d}) \leq \varepsilon$ whenever $\|\beta\|_1 \geq \log(d/\varepsilon)$, with no dependence on S.

Proof. We construct an explicit synchronous coupling of two copies of the forward chain, using the joint uniformized representation of Remark F.3, and apply Lemma F.1.

Coupling. Draw $X_0 \sim p_{\text{data}}$ and $Y_0 \sim \pi^{\otimes d}$ independently. Let $N_1, \ldots, N_d$ be i.i.d. Poisson processes with rate $\beta(t)$, and let $(\xi_k^{(i)})_{i\in[d],\, k\geq 1}$ be i.i.d. $\mathrm{Unif}([S])$, independent of N. Evolve both X and Y by the joint uniformized construction of Remark F.3 driven by the *same* clocks (N_i) and *same* samples $(\xi_k^{(i)})$.

Marginal verification. By Lemma F.2 applied per coordinate, $X_t \sim p_t$ (the forward chain started at p_{data}) and $Y_t \sim \pi^{\otimes d}$ for all $t \geq 0$ (the forward chain started at its stationary law, which it preserves).

Merging per coordinate. Let $\tau_i := \inf\{t \geq 0 : N_i([0,t]) \geq 1\}$ be the first ring time at position i. At $t = \tau_i$, the shared sample $\xi_1^{(i)}$ resets both $X^{(i)}$ and $Y^{(i)}$ to the same value (this is property (70) of the uniformized representation: the reset distribution is state-independent, so the same sample is valid for both chains). The shared subsequent samples then keep them equal: $X_t^{(i)} = Y_t^{(i)}$ for all $t \geq \tau_i$. Hence

$$\{X_T \neq Y_T\} \subseteq \bigcup_{i=1}^{d} \{\tau_i > T\}.$$

Union bound. Each τ_i is the first event of a Poisson process with rate $\beta(t)$, so $\Pr(\tau_i > T) = \rho_T$. By the union bound,

$$\Pr(X_T \neq Y_T) \leq \sum_{i=1}^{d} \Pr(\tau_i > T) = d\,\rho_T.$$

Conclusion. The construction is a specific coupling $\Gamma \in \Pi(p_T, \pi^{\otimes d})$. By (66) of Lemma F.1,

$$\mathrm{TV}(p_T, \pi^{\otimes d}) \leq \Pr_{\Gamma}(X_T \neq Y_T) \leq d\,\rho_T = d\,\exp\Big(-\int_0^T \beta(t)\,dt\Big) = de^{-\|\beta\|_1},$$

by (65). The closure $d\,\rho_T \leq \varepsilon$ holds iff $\|\beta\|_1 \geq \log(d/\varepsilon)$. □

Remark (S-independence)**.** *The vocabulary size S enters the bound only through the choice of reference π, never through any rate. The synchronous coupling construction is agnostic to the reset distribution; the analysis depends only on π being preserved by the dynamics, so that $Y_t \sim \pi^{\otimes d}$ for all t.*

Remark. *This argument is the continuous-time, general state-space analog of the classical coupling analysis of the lazy random walk on the hypercube $\{0,1\}^n$ (Levin & Peres, 2017, Chapter 5); the underlying workhorse is the Markovian coupling theorem (Levin & Peres, 2017, Theorem 5.2). An equivalent analysis through the strong stationary time framework (Levin & Peres, 2017, Chapter 6) yields the slightly tighter bound $\mathrm{TV}(p_T, \pi^{\otimes d}) \leq 1 - (1-\rho_T)^d$.*

G Bounding the Early-Stopping Error Term

When the reverse process is stopped at a small time $\delta > 0$ rather than at 0, the law p_δ differs from the true data law p_0, contributing an early-stopping error $\mathrm{TV}(p_0, p_\delta)$. The same coupling argument used to bound the initialization error in Section F applies here, run on the time interval $[0, \delta]$ instead of $[0, T]$. We state the lemmas and proofs below.

Lemma G.1 (Early-Stopping Error under Uniform Transition)**.** *For all $\delta \geq 0$,*

$$\mathrm{TV}(p_0, p_\delta) \leq 1 - \rho_\delta^d = 1 - e^{-d\int_0^\delta \beta(t)\,dt} \leq d\int_0^\delta \beta(t)\,dt.$$

Proof. We couple $X_0 \sim p_0 = p_{\mathrm{data}}$ with $X_\delta \sim p_\delta$ by running a single copy of the forward chain via the joint uniformized construction of Remark F.3, with i.i.d. per-coordinate Poisson clocks $N_1, \ldots, N_d$ of rate $\beta(t)$.

No-ring event. Let $\tau_i := \inf\{t \geq 0 : N_i([0,t]) \geq 1\}$ be the first ring time at coordinate i. On $\{\tau_i > \delta\}$ coordinate i is never reset, so $X_\delta^{(i)} = X_0^{(i)}$. Hence

$$\{X_0 = X_\delta\} \supseteq \bigcap_{i=1}^{d} \{\tau_i > \delta\}.$$

Probability bound. The Poisson clocks are independent and $\Pr(\tau_i > \delta) = e^{-\int_0^\delta \beta(t)\,dt} =: \rho_\delta$, so

$$\Pr(X_0 = X_\delta) \geq \prod_{i=1}^{d} \Pr(\tau_i > \delta) = \rho_\delta^d.$$

Conclusion. The construction is a coupling $\Gamma \in \Pi(p_0,\, p_\delta)$. By (66) of Lemma F.1,

$$\mathrm{TV}(p_0,\, p_\delta) \;\le\; \Pr_{\Gamma}(X_0 \neq X_\delta) \;\le\; 1 - \rho_\delta^d \;=\; 1 - e^{-d\int_0^\delta \beta(t)\,dt},$$

and $1 - e^{-x} \le x$ yields $\mathrm{TV}(p_0,\, p_\delta) \le d\int_0^\delta \beta(t)\,dt$. □

Lemma G.2 (Early-Stopping Bound under Masked Diffusion)**.** *Under masked diffusion with early stopping at $\delta > 0$, the early-stopping mismatch term is bounded by*

$$\mathrm{TV}(p_0, p_\delta) \;\le\; 1 - e^{-d\int_0^\delta \beta(t)\,dt}. \tag{72}$$

Proof. Under the Kronecker-sum factorization (2), the forward process masks each of the d tokens *independently*, each with single-token mask probability

$$\alpha_\delta = 1 - e^{-\int_0^\delta \beta(t)\,dt}, \tag{73}$$

so the probability that a given token remains unmasked is $1 - \alpha_\delta = e^{-\int_0^\delta \beta(t)\,dt}$.

By independence across tokens, the probability that none of the d tokens are selected for masking by time δ is $(1-\alpha_\delta)^d$. Conditional on this event, $\mathbf{x}_\delta = \mathbf{x}_0$ exactly, so there exists a coupling of $(\mathbf{x}_0, \mathbf{x}_\delta) \sim (p_0, p_\delta)$ with

$$\mathbb{P}(x_0 = x_\delta) \;=\; \underbrace{(1-\alpha_\delta)^d}_{\text{prob. of all } d \text{ tokens remain unmasked}}.$$

The total variation is bounded by the coupling disagreement probability (Lemma F.1), hence

$$\mathrm{TV}(p_0, p_\delta) \;\le\; \mathbb{P}(x_0 \neq x_\delta) \;\le\; 1 - (1-\alpha_\delta)^d \;=\; 1 - e^{-d\int_0^\delta \beta(t)\,dt},$$

and $1 - e^{-x} \le x$ yields $\mathrm{TV}(p_0,\, p_\delta) \le d\int_0^\delta \beta(t)\,dt$. □

H S-INDEPENDENCE OF THE BOUNDS

We provide the detailed discussion on the scaling of the bounds in our main theory, Lemma 3.2. A crucial feature of our bound is that, the coefficient of the training loss term is independent of the vocabulary size S. This independence is practically significant: in applications such as language modeling, S corresponds to the vocabulary size, which is prohibitively large and renders any S-dependent bound vacuous.

H.1 SCALING OF THE BOUNDS

Case 1 (Masked). In Case 1, the bound is completely independent of S.

The prefactor $\sqrt{d}\sqrt{\|\beta(1-p(\mathbf{m}))\|_1} = \sqrt{d}\left(\int_0^T \beta(t)(1-p_t(\mathbf{m}))dt\right)^{\frac{1}{2}}$ is obtained using the exit-routing technique; see Remarks 3.2 and D.2. The prefactor depends on the schedule $\beta(t)$ and the all-mask probability $p_t(\mathbf{m}) = \alpha_t^d$ under the Kronecker-sum factorization (2) with single-token mask probability α_t. Here, α_t satisfies the ODE

$$\partial_t \alpha_t = \beta(t)(1-\alpha_t) \quad \text{subject to} \quad \alpha_0 = 0, \tag{74}$$

derived from the masked rate (3)) and with solution

$$\alpha_t = 1 - e^{-\int_0^t \beta(s)ds}, \quad \text{for } t \in [0, T]. \tag{75}$$

Thus, neither of $\beta(t)$ and $p_t(\mathbf{m})$ involves S. So the prefactor is indepedent of S.

The loss $\mathcal{L}_{\mathrm{SE}}$ defined in (8) under masked transition is structurally S-independent: at each $(t, \mathbf{y})$ the integrand is summed by $\sum_{\mathbf{y}\neq\mathbf{x}} = \sum_{\mathbf{x}\in\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})}$, where $\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})$ is the successor set of $\mathbf{y}$. It is defined in (34) and contains at most d terms (one successor per position, obtained by masking that position, and there are d positions).

Case 1 (Masked): Score-marginal Cancellation. The S-independence depends critically on our **score-marginal cancellation** $p_t(\mathbf{x})s_t(\mathbf{x})_{\mathbf{y}} = p_t(\mathbf{y})$ applied in the proof of Lemma 3.2. The unsimplified integrand of (35) sums over the $m(\mathbf{x})(S-1)$ predecessors $\mathbf{y}$ of each $\mathbf{x}$, with the factor $S-1$ counting vocabulary values. The cancellation re-indexes the sum to range over $\mathbf{x} \in \mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})$ ($\leq d$ terms, no vocabulary factor), absorbing the $S-1$; Lemma E.3 then converts the resulting sum into the conditional-expectation form appearing in $\mathcal{L}_{\text{SE}}$.

The boundary is also S-independent, by Lemma E.14.

Case 2 (Uniform). In Case 2, the bound is completely independent of S.

The prefactor $\sqrt{d}\sqrt{\|\beta\|_1}$ and the loss $\mathcal{L}_{\text{WSM}}$ defined in (7) are both S-independent. The boundary term $\text{TV}(p_T, p_{\text{base}})$ is also independent of S, by Theorem F.4.

Full step-by-step derivations of the claims above – the Kronecker-sum factorization $p_t(\mathbf{m}) = \alpha_t^d$, the successor-set count $|\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})| \leq d$, the re-indexing via the score-marginal cancellation, the Case 2 boundary computation, and the Case 1 spectral-gap bound – are given in Appendix H.2.

H.2 Derivational Details to Achieve S-Independence

In this subsection, we detail the techniques used in our derivation to achieve independence in the vocabulary size S.

Case 1: Prefactor of training loss. The prefactor outside $\sqrt{\mathcal{L}_{\text{SE}}(s, \tilde{s}) + \mathcal{O}(|s - \tilde{s}|^3)}$ in Case 1 is

$$\sqrt{2d}\Big(\int_0^T \beta(t)(1 - p_t(\mathbf{m}))\, dt\Big)^{1/2}, \tag{76}$$

which is obtained using the exit-routing technique; see Remarks 3.2 and D.2. The factor $\sqrt{d}$ is the sequence length. Inside the integral, $\beta(t)$ is the schedule and $p_t(\mathbf{m})$ is the joint probability of the all-mask state. Under the Kronecker-sum dynamics (2), the conditional forward kernel factorizes as $p_{t|0}(\mathbf{x} \mid \mathbf{x}_0) = \prod_i p^{\text{tok}}_{t|0}(x^i \mid x^i_0)$, where $p^{\text{tok}}_{t|0}([\text{MASK}] \mid x^i_0) = \alpha_t$ for any non-mask x^i_0 and the single-token mask probability α_t satisfies $\partial_t \alpha_t = \beta(t)(1 - \alpha_t)$. Marginalizing over $\mathbf{x}_0 \sim p_{\text{data}}$ (supported on non-mask sequences),

$$p_t(\mathbf{m}) = \sum_{\mathbf{x}_0} p_{\text{data}}(\mathbf{x}_0) \prod_i p^{\text{tok}}_{t|0}([\text{MASK}] \mid x^i_0) = \alpha_t^d. \tag{77}$$

The integral in (76) depends only on β and d (no S), and the whole prefactor is S-independent.

Case 1: Score-marginal Cancellation $p_t(\mathbf{x})s_t(\mathbf{x})_{\mathbf{y}} = p_t(\mathbf{y})$. The S-independent of the bound (14) results from a re-indexing in the proof. The integrant in the derivation in Section D.1 reads

$$\sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \neq \mathbf{x}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) \left(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}\right)[\,\cdot\,]. \tag{78}$$

For the masked kernel, $Q_t(\mathbf{y}, \mathbf{x}) > 0$ when $\mathbf{y} \to \mathbf{x}$ in the forward chain, i.e., $\mathbf{x}$ is obtained from $\mathbf{y}$ by masking one non-mask position; equivalently, $\mathbf{x} \in \mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})$. With $\mathbf{x}$ fixed, $\mathbf{y}$ is obtained from $\mathbf{x}$ by replacing one of $\mathbf{x}$'s $m(\mathbf{x})$ [MASK] tokens with one of $S-1$ non-mask values:

$$\#\{\mathbf{y} : Q_t(\mathbf{y}, \mathbf{x}) > 0\} = m(\mathbf{x})(S-1), \tag{79}$$

so the factor $S-1$ enters the inner sum.

The substitution $s_t(\mathbf{x})_{\mathbf{y}} = p_t(\mathbf{y})/p_t(\mathbf{x})$ gives $p_t(\mathbf{x})(s_t(\mathbf{x})_{\mathbf{y}} - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}) = p_t(\mathbf{y})(1 - \tilde{s}_t(\mathbf{x})_{\mathbf{y}}/s_t(\mathbf{x})_{\mathbf{y}})$. The sum re-indexes by holding $\mathbf{y}$ fixed and summing over $\mathbf{x} \in \mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})$:

$$\sum_{\mathbf{x}, \mathbf{y} \neq \mathbf{x}} p_t(\mathbf{x}) Q_t(\mathbf{y}, \mathbf{x}) \left(s_t - \tilde{s}_t\right)[\,\cdot\,] = \sum_{\mathbf{y} \neq \mathbf{m}} p_t(\mathbf{y}) \sum_{\mathbf{x} \in \mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})} Q_t(\mathbf{y}, \mathbf{x}) \left(1 - \tfrac{\tilde{s}_t(\mathbf{x})_{\mathbf{y}}}{s_t(\mathbf{x})_{\mathbf{y}}}\right)[\,\cdot\,]. \tag{80}$$

The outer sum is restricted to $\mathbf{y} \neq \mathbf{m}$ since $\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{m}) = \emptyset$, and is weighted by $p_t(\mathbf{y})$ with total mass $\sum_{\mathbf{y} \neq \mathbf{m}} p_t(\mathbf{y}) = 1 - p_t(\mathbf{m}) \leq 1$. The inner sum has $|\mathcal{N}^{\leftarrow}_{\text{masked}}(\mathbf{y})| \leq d$ terms. The re-indexing converts the $m(\mathbf{x})(S-1)$ choices in the original sum into $\leq d$ choices in the re-indexed sum; the factor of $S-1$ is absorbed by the score-marginal cancellation.

Case 1: boundary term. For $p_{\text{base}} = \delta_{\mathbf{m}}$,

$$\begin{aligned}\text{TV}(p_T, \delta_{\mathbf{m}}) &= \tfrac{1}{2}\sum_{\mathbf{x}} |p_T(\mathbf{x}) - \delta_{\mathbf{m}}(\mathbf{x})| \\ &= \tfrac{1}{2}\big[\, |p_T(\mathbf{m}) - 1| + \sum_{\mathbf{x}\neq\mathbf{m}} p_T(\mathbf{x})\,\big] \\ &= \tfrac{1}{2}\big[(1 - p_T(\mathbf{m})) + (1 - p_T(\mathbf{m}))\big] \;=\; 1 - p_T(\mathbf{m}),\end{aligned}$$

using $p_T(\mathbf{m}) \le 1$ and $\sum_{\mathbf{x}\neq\mathbf{m}} p_T(\mathbf{x}) = 1 - p_T(\mathbf{m})$.

By construction of the forward process, conditional on the initial sequence $\mathbf{x}_0$, the d positions evolve as **independent** CTMCs, each with the same time-dependent rate β_t and [MASK] absorbing. Hence for any non-mask token v the per-position mask probability

$$\alpha_T \;:=\; \Pr(\mathbf{x}_T^{(i)} = \text{[MASK]} \mid \mathbf{x}_0^{(i)} = v)$$

depends on neither i (positions share dynamics) nor v (the absorption rate is the same for every non-mask token), so α_T is indeed a (pure) scalar. The data distribution p_{data} is supported on non-mask sequences, so for p_{data}-a.e. $\mathbf{x}_0$ every coordinate $\mathbf{x}_0^{(i)}$ is non-mask, and independence of the per-position CTMCs gives

$$\Pr(\mathbf{x}_T = \mathbf{m} \mid \mathbf{x}_0) \;=\; \prod_{i=1}^{d} \Pr(\mathbf{x}_T^{(i)} = \text{[MASK]} \mid \mathbf{x}_0^{(i)}) \;=\; \alpha_T^d.$$

Marginalizing over $\mathbf{x}_0$,

$$p_T(\mathbf{m}) \;=\; \mathbb{E}_{\mathbf{x}_0\sim p_{\text{data}}}\big[\,\Pr(\mathbf{x}_T = \mathbf{m} \mid \mathbf{x}_0)\big] \;=\; \alpha_T^d,$$

since the conditional probability is the constant α_T^d p_{data}-a.s.

The factorization $1 - \alpha_T^d = (1 - \alpha_T)\sum_{k=0}^{d-1} \alpha_T^k$ with $\alpha_T \in [0, 1]$ then gives

$$\text{TV}(p_T, \delta_{\mathbf{m}}) \;=\; 1 - \alpha_T^d \;\le\; d(1 - \alpha_T),$$

independent of vocabulary size S: absorption sends every non-mask token to a single destination per position.

I Proof of Uniformization Iteration Complexity

I.1 Proof of Auxiliary Lemmas

We state and proof the lemmas for the iteration complexity of uniformization.

Lemma I.1 (Conditional Representation of Score)**.** *For any* $\mathbf{x}$*,* $\mathbf{y}$*, and* $t > 0$ *satisfying* $p_t(\mathbf{x}) \neq 0$ *and* $\mathbf{x}_{i\to y^i} = \mathbf{y}$*, we have*

$$s_t(\mathbf{x})_{\mathbf{y}} = \mathbb{E}_{\mathbf{x}_0\sim p_{0|t}(\cdot|\mathbf{x})}\left[\frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)}\right]. \tag{81}$$

Proof. This lemma directly follows (Chen & Ying, 2025, Lemma 2), we show the proof here for completeness.

$$\begin{aligned}s_t(\mathbf{x})_{\mathbf{y}} = \frac{p_t(\mathbf{y})}{p_t(\mathbf{x})} &= \frac{p_t(\mathbf{x}_{i\to y^i})}{p_t(\mathbf{x})} \\ &= \frac{1}{p_t(\mathbf{x})}\sum_{\mathbf{x}_0\in\mathcal{X}} p_0(\mathbf{x}_0)p_{t|0}(\mathbf{x}_{i\to y^i}|\mathbf{x}_0) \\ &= \frac{1}{p_t(\mathbf{x})}\sum_{\mathbf{x}_0\in\mathcal{X}} p_0(\mathbf{x}_0)\prod_{k\neq i} p^k_{t|0}(x^k|x^k_0)p^i_{t|0}(y^i|x^i_0)\end{aligned}$$

$$= \frac{1}{p_t(\mathbf{x})} \sum_{\mathbf{x}_0 \in \mathcal{X}} p_0(\mathbf{x}_0) \prod_{k=1}^{d} p^k_{t|0}(x^k|x^k_0) \frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)}$$

$$= \sum_{\mathbf{x}_0 \in \mathcal{X}} \frac{p_0(\mathbf{x}_0) p_{t|0}(\mathbf{x}|\mathbf{x}_0)}{p_t(\mathbf{x})} \frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)}$$

$$= \sum_{\mathbf{x}_0 \in \mathcal{X}} p_{0|t}(\mathbf{x}_0|\mathbf{x}) \frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)},$$

by Bayes' rule,

$$= \mathbb{E}_{\mathbf{x}_0 \sim p_{0|t}(\cdot|\mathbf{x})} \left[\frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)} \right].$$

□

Lemma I.2 (Score Representation under Masked Rate)**.** *Under the masked rate* (3) *and* (2)*, when* $Q_t(\mathbf{y}, \mathbf{x}) > 0$*, the score is given by*

$$s_t(\mathbf{x})_{\mathbf{y}} = \frac{e^{-\int_0^t \beta(\tau) d\tau}}{1 - e^{-\int_0^t \beta(\tau) d\tau}} p^i_{0|t}(y^i|\mathbf{x}).$$

Proof. This proof is modified from (Liang et al., 2025, Lemma 1) and (Ou et al., 2025, Theorem 1). Under the masked rate (3) and (2), by Lemma I.1, we have

$$s_t(\mathbf{x})_{\mathbf{y}} = \mathbb{E}_{\mathbf{x}_0 \sim p_{0|t}(\cdot|\mathbf{x})} \left[\frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)} \right]. \tag{82}$$

On the other hand, the single-token mask probability α_t satisfies the ODE

$$\partial_t \alpha_t = \beta(t)(1 - \alpha_t) \quad \text{subject to} \quad \alpha_0 = 0,$$

derived from the masked rate (3) and with solution

$$\alpha_t = 1 - e^{-\int_0^t \beta(\tau) d\tau}, \quad \text{for } t \in [0, T]. \tag{83}$$

Consider the case where $Q_t(\mathbf{y}, \mathbf{x}) > 0$, which means $\mathbf{x} = \mathbf{y}_{i \to [\text{MASK}]}$ and $y^i \neq [\text{MASK}]$ for an $i \in [d]$. And the likelihood $p^i_{t|0}(y^i|x^i_0) \neq 0$ only when $x^i_0 = y^i$. In this case, by (83), the likelihood ratio in (82) is given by

$$\frac{p^i_{t|0}(y^i|y^i)}{p^i_{t|0}(x^i|y^i)} = \frac{p^i_{t|0}(y^i|y^i)}{p^i_{t|0}([\text{MASK}]|y^i)} = \frac{1 - \alpha_t}{\alpha_t} = \frac{e^{-\int_0^t \beta(\tau) d\tau}}{1 - e^{-\int_0^t \beta(\tau) d\tau}}. \tag{84}$$

Thus, when $Q_t(\mathbf{y}, \mathbf{x}) > 0$, (82) can be evaluated as

$$\begin{aligned}
s_t(\mathbf{x})_{\mathbf{y}} &= \mathbb{E}_{\mathbf{x}_0 \sim p_{0|t}(\cdot|\mathbf{x})} \left[\frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)} \right] \\
&= \mathbb{E}_{x^i_0 \sim p^i_{0|t}(\cdot|\mathbf{x})} \left[\frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)} \right], \quad \text{since the expression inside the expectation only depends on } x^i_0, \\
&= \sum_{x^i_0 \in [S]} \frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)} p^i_{0|t}(x^i_0|\mathbf{x}) \\
&= \sum_{\substack{x^i_0 \in [S] \\ x^i_0 = y^i}} \frac{p^i_{t|0}(y^i|x^i_0)}{p^i_{t|0}(x^i|x^i_0)} p^i_{0|t}(x^i_0|\mathbf{x}), \quad \text{as } p^i_{t|0}(y^i|x^i_0) = 0 \text{ when } x^i_0 \neq y^i \text{ under the masked rate,}
\end{aligned}$$

$$= \frac{p^i_{t|0}(y^i|y^i)}{p^i_{t|0}(x^i|y^i)} p^i_{0|t}(y^i|\mathbf{x})$$
$$= \frac{e^{-\int_0^t \beta(\tau)d\tau}}{1-e^{-\int_0^t \beta(\tau)d\tau}} p^i_{0|t}(y^i|\mathbf{x}), \quad \text{by (84).}$$

□

Lemma I.3 (Score Upper Bound under Masked Rate)**.** *Under the masked rate* (3) *and* (2)*,*

$$\sum_{\substack{\mathbf{y}\in[S]^d \\ \mathbf{y}\neq\mathbf{x}}} s_t(\mathbf{x})_{\mathbf{y}} Q_t(\mathbf{y},\mathbf{x}) \leq \frac{e^{-\int_0^t \beta(\tau)d\tau}}{1-e^{-\int_0^t \beta(\tau)d\tau}} \beta(t) \underbrace{\left[\sum_{i\in[d]} \delta\{x^i, \textit{[MASK]}\}\right]}_{\textit{number of masks in } \mathbf{x}}.$$

Proof. This proof is modified from (Liang et al., 2025, Lemma 5).

Under the masked rate (3) and (2), by Lemma I.2, when $Q_t(\mathbf{y},\mathbf{x}) > 0$, we have

$$s_t(\mathbf{x})_{\mathbf{y}} = \frac{e^{-\int_0^t \beta(\tau)d\tau}}{1-e^{-\int_0^t \beta(\tau)d\tau}} p^i_{0|t}(y^i|\mathbf{x}). \tag{85}$$

Here, $\mathbf{x} = \mathbf{y}_{i\to[\text{MASK}]}$ and $y^i \neq [\text{MASK}]$. We evaluate

$$\begin{aligned}
\sum_{\substack{\mathbf{y}\in[S]^d \\ \mathbf{y}\neq\mathbf{x}}} s_t(\mathbf{x})_{\mathbf{y}} Q_t(\mathbf{y},\mathbf{x}) &= \sum_{\substack{\mathbf{y}\in[S]^d \\ Q_t(\mathbf{y},\mathbf{x})>0}} s_t(\mathbf{x})_{\mathbf{y}} Q_t(\mathbf{y},\mathbf{x}) \\
&= \sum_{\substack{i\in[d] \\ x^i=[\text{MASK}]}} \sum_{\substack{y^i\in[S] \\ y^i\neq[\text{MASK}]}} s_t(\mathbf{x})_{\mathbf{y}} \beta(t) \\
&= \frac{e^{-\int_0^t \beta(\tau)d\tau}}{1-e^{-\int_0^t \beta(\tau)d\tau}} \beta(t) \sum_{\substack{i\in[d] \\ x^i=[\text{MASK}]}} \underbrace{\sum_{\substack{y^i\in[S] \\ y^i\neq[\text{MASK}]}} p^i_{0|t}(y^i|\mathbf{x})}_{\leq 1},
\end{aligned}$$

by (85),

$$\leq \frac{e^{-\int_0^t \beta(\tau)d\tau}}{1-e^{-\int_0^t \beta(\tau)d\tau}} \beta(t) \underbrace{\left[\sum_{i\in[d]} \delta\{x^i, [\text{MASK}]\}\right]}_{\text{number of masks in } \mathbf{x}}.$$

□

Lemma I.4. *Given a constant $R > 1$, if r satisfies $1 \leq r \leq R$, then*

$$r - 1 \leq \frac{R-1}{\log R} \log r. \tag{86}$$

Proof. If $r = 1$, then $\log r = 0$ and (86) trivially holds. Consider the case when $r > 1$, , we can equivalently show that

$$\frac{r-1}{\log r} \leq \frac{R-1}{\log R}.$$

To this end, it suffices to show that the function

$$h(r) = \frac{r-1}{\log r}$$

is increasing for $r > 1$. It derivative is given by

$$h'(r) = \frac{\log r - \frac{r-1}{r}}{(\log r)^2} = \frac{\log r - 1 + \frac{1}{r}}{(\log r)^2}. \tag{87}$$

Here, let the numerator be $g(r) = \log r - 1 + \frac{1}{r}$. Remark that $g(1) = 0$ and $g'(r) = \frac{1}{r} - \frac{1}{r^2} > 0$ for $r > 1$. Thus, $g(r) > 0$ for $r > 1$. This implies $h'(r) > 0$ and thus $h(r)$ is increasing for $r > 1$. □

Lemma I.5. *For a time discretization $\{t_k\}_{k=0}^{N_B}$, where $\delta = t_0 < t_1 < \cdots < t_{N_B} = T$, and satisfying $\max_k \frac{t_{k+1}}{t_k} = R$ for some constant $R > 1$, we have*

$$\frac{t_{k+1} - t_k}{t_k} \leq \frac{R-1}{\log R} \log \frac{t_{k+1}}{t_k}. \tag{88}$$

Proof. Note that

$$\frac{t_{k+1} - t_k}{t_k} = \frac{t_{k+1}}{t_k} - 1.$$

Letting $r = \frac{t_{k+1}}{t_k}$ and applying Lemma I.4 yields the desired result. □

I.2 Proof of Corollary 3.4

Corollary 3.4 (Expected Step Complexities for Uniformization)**.** *Assume that $\tilde{s}_t(\mathbf{x})_{\mathbf{y}} \asymp s_t(\mathbf{x})_{\mathbf{y}}$ whenever $Q_t(\mathbf{y}, \mathbf{x}) > 0$, and the uniformization Poisson rate is chosen as $\Lambda_k := C \sup_{\substack{\mathbf{x}\in\mathcal{X} \\ t\in(t_k, t_{k+1}]}} -\tilde{Q}_t^{\leftarrow}(\mathbf{x}, \mathbf{x})$ for some constant $C \geq 1$. Then, for any IPM γ_Ψ and tolerance $\epsilon > 0$, the approximation error satisfies $\gamma_\Psi(p_{\text{data}}, \tilde{p}_\delta) < \epsilon$ if we choose $\delta \asymp \frac{\epsilon}{dC_\Psi \beta_{\max}}$, $T \asymp \frac{1}{\beta_{\min}} \log\left(\frac{dC_\Psi}{\epsilon}\right)$, and $\mathcal{L}_{\text{SE}}^{\delta}(s, \tilde{s}) = \mathcal{O}(\frac{\epsilon^2}{dC_\Psi^2})$, in which case the expected number of steps are given by*

1. ***Case 1 (Masked rate;** $Q_t^{\text{tok}} = Q_t^{\text{masked}}$ **(3))**:*

$$\mathbb{E}[N] = \mathcal{O}\left(d \frac{\beta_{\max}}{\beta_{\min}} \log\left(\frac{\beta_{\max}}{\beta_{\min}} \cdot \frac{dC_\Psi}{\epsilon}\right)\right), \quad \textit{and,}$$

2. ***Case 2 (Uniform rate;** $Q_t^{\text{tok}} = Q_t^{\text{uniform}}$ **(3))**: if the true score further satisfies $s_t(\mathbf{x})_{\mathbf{y}} \lesssim \max(1, t^{-1})$ (following Ren et al., 2025, Assumption 4.4), then*

$$\mathbb{E}[N] = \mathcal{O}\left(\beta_{\max} d \left(\log \beta_{\max} + \frac{1}{\beta_{\min}} \log \frac{dC_\Psi}{\epsilon}\right)\right).$$

Here, $\beta_{\max} = \max_{0\leq t\leq T} \beta(t)$ and $\beta_{\min} = \min_{0\leq t\leq T} \beta(t)$, and the values of C_Ψ are given in Table 2.

Proof of Corollary 3.4 for the Masked Rate. We first prove the expected step complexity for the masked rate. The proof is modified from (Liang et al., 2025, Theorem 3).

By Lemma I.3, we have

$$\sum_{\substack{\mathbf{y}\in[S]^d \\ \mathbf{y}\neq\mathbf{x}}} s_t(\mathbf{x})_{\mathbf{y}} Q_t(\mathbf{y}, \mathbf{x}) \leq \underbrace{\frac{e^{-\int_0^t \beta(\tau)d\tau}}{1 - e^{-\int_0^t \beta(\tau)d\tau}}}_{\leq 1/\int_0^t \beta(\tau)d\tau} \beta(t) \underbrace{\left[\sum_{i\in[d]} \delta\{x^i, [\text{MASK}]\}\right]}_{\text{number of masks in } \mathbf{x},\ \leq d} \leq \frac{d\beta(t)}{\int_0^t \beta(\tau)d\tau}.$$

Thus, under the assumption that $\tilde{s}_t(\mathbf{x})_{\mathbf{y}} \asymp s_t(\mathbf{x})_{\mathbf{y}}$ when $Q_t(\mathbf{y}, \mathbf{x}) > 0$, we have

$$-\tilde{Q}_t^{\leftarrow}(\mathbf{x}, \mathbf{x}) = \sum_{\substack{\mathbf{y}\in[S]^d \\ \mathbf{y}\neq\mathbf{x}}} \tilde{Q}_t^{\leftarrow}(\mathbf{x}, \mathbf{y}) = \sum_{\substack{\mathbf{y}\in[S]^d \\ \mathbf{y}\neq\mathbf{x}}} Q_t(\mathbf{y}, \mathbf{x}) \tilde{s}_t(\mathbf{x})_{\mathbf{y}} \lesssim \sum_{\substack{\mathbf{y}\in[S]^d \\ \mathbf{y}\neq\mathbf{x}}} Q_t(\mathbf{y}, \mathbf{x}) s_t(\mathbf{x})_{\mathbf{y}} \leq \frac{d\beta(t)}{\int_0^t \beta(\tau)d\tau}. \tag{89}$$

To simulate the approximate reverse process, in each discretized time interval $(t_k, t_{k+1}]$, uniformization requires that the Poisson rate Λ_k satisfies

$$\Lambda_k \geq \sup_{\substack{\mathbf{x}\in\mathcal{X} \\ t\in(t_k, t_{k+1}]}} -\tilde{Q}_t^{\leftarrow}(\mathbf{x}, \mathbf{x}).$$

Thus, suppose we choose the Poisson rate to be

$$\Lambda_k = C \cdot \sup_{\substack{\mathbf{x}\in\mathcal{X} \\ t\in(t_k,t_{k+1}]}} -\tilde{Q}_t^{\leftarrow}(\mathbf{x},\mathbf{x}), \tag{90}$$

for some constant $C \geq 1$. Remark that the sum of independent Poisson(Λ_k) random variables is distributed as Poisson$(\sum_k \Lambda_k)$, the expected total number of steps are given by

$$\mathbb{E}[N] = \sum_{k=0}^{N_B-1} \Lambda_k (t_{k+1} - t_k),$$

where $\delta = t_0 < t_1 < \cdots < t_{N_B} = T$ is the time discretization,

$$\begin{aligned}
&\lesssim \sum_{k=0}^{N_B-1} \frac{d \max_{0\leq t\leq T}(\beta(t))}{\int_0^{t_k}\beta(\tau)d\tau}(t_{k+1}-t_k), \quad \text{by (89) and (90)},\\
&= d \max_{0\leq t\leq T}(\beta(t)) \sum_{k=0}^{N_B-1} \frac{1}{\int_0^{t_k}\beta(\tau)d\tau}(t_{k+1}-t_k)\\
&\leq d \max_{0\leq t\leq T}(\beta(t)) \sum_{k=0}^{N_B-1} \frac{1}{t_k \min_{0\leq t\leq t_k}(\beta(t))}(t_{k+1}-t_k)\\
&\leq d \frac{\max_{0\leq t\leq T}(\beta(t))}{\min_{0\leq t\leq T}(\beta(t))} \sum_{k=0}^{N_B-1} \frac{t_{k+1}-t_k}{t_k}\\
&= d \frac{\beta_{\max}}{\beta_{\min}} \sum_{k=0}^{N_B-1} \frac{t_{k+1}-t_k}{t_k}\\
&\leq d \frac{\beta_{\max}}{\beta_{\min}} \sum_{k=0}^{N_B-1} \frac{R-1}{\log R} \log\frac{t_{k+1}}{t_k}, \quad \text{by Lemma I.5, where } R = \max_k \frac{t_{k+1}}{t_k},\\
&\lesssim d \frac{\beta_{\max}}{\beta_{\min}} \sum_{k=0}^{N_B-1} \log\frac{t_{k+1}}{t_k}\\
&= d \frac{\beta_{\max}}{\beta_{\min}} \sum_{k=0}^{N_B-1} \int_{t_k}^{t_{k+1}} \frac{1}{t} dt\\
&= d \frac{\beta_{\max}}{\beta_{\min}} \int_{\delta}^{T} \frac{1}{t} dt\\
&= d \frac{\beta_{\max}}{\beta_{\min}} (\log(T) + \log(\delta^{-1})).
\end{aligned} \tag{91}$$

On the other hand, by Corollary 3.3, the approximation error under any IPM γ_Φ satisfies

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_\delta) \leq 2C_\Psi \left(d\int_0^\delta \beta(t)\, dt + de^{-\|\beta\|_1} + \sqrt{2d}\sqrt{\|\beta(1-p(\mathbf{m}))\|_1}\sqrt{\mathcal{L}_{\text{SE}}^\delta(s,\tilde{s}) + \tfrac{2}{3}\mathcal{L}_3^\delta(s,\tilde{s})} \right).$$

For a given threshold $\epsilon > 0$ for the approximation error, it suffices to bound each of the three terms on the right-hand-side by $\epsilon/3$. This is achieved by

$$\delta < \frac{\epsilon}{6C_\Psi d\beta_{\max}} \;\Rightarrow\; \delta^{-1} > \frac{6C_\Psi d\beta_{\max}}{\epsilon}, \quad \text{thus we choose } \delta^{-1} \asymp \frac{dC_\Psi\beta_{\max}}{\epsilon}, \tag{92}$$

$$T > \frac{1}{\beta_{\min}} \log\left(\frac{6C_\Psi d}{\epsilon}\right), \quad \text{thus we choose } T \asymp \frac{1}{\beta_{\min}} \log\left(\frac{C_\Psi d}{\epsilon}\right), \tag{93}$$

$$\mathcal{L}_{\text{SE}}^\delta(s,\tilde{s}) + \tfrac{2}{3}\mathcal{L}_3^\delta(s,\tilde{s}) = \frac{\epsilon^2}{72dC_\Psi^2\|\beta(1-p(\mathbf{m}))\|_1} \;\Rightarrow\; \mathcal{L}_{\text{SE}}^\delta(s,\tilde{s}) \lesssim \frac{\epsilon^2}{dC_\Psi^2\|\beta(1-p(\mathbf{m}))\|_1}. \tag{94}$$

Plugging (92) and (93) into (91) yields

$$\begin{aligned}\mathbb{E}[N] &\lesssim d\,\frac{\beta_{\max}}{\beta_{\min}}(\log(T)+\log(\delta^{-1}))\\ &= \mathcal{O}\left(d\,\frac{\beta_{\max}}{\beta_{\min}}\left(\log\left(\frac{1}{\beta_{\min}}\log\frac{dC_\Psi}{\epsilon}\right)+\log\frac{dC_\Psi\beta_{\max}}{\epsilon}\right)\right)\\ &= \mathcal{O}\left(d\,\frac{\beta_{\max}}{\beta_{\min}}\left(\log\frac{1}{\beta_{\min}}+\log\frac{dC_\Psi\beta_{\max}}{\epsilon}\right)\right)\\ &= \mathcal{O}\left(d\,\frac{\beta_{\max}}{\beta_{\min}}\log\left(\frac{\beta_{\max}}{\beta_{\min}}\cdot\frac{dC_\Psi}{\epsilon}\right)\right).\end{aligned}$$

□

Proof of Corollary 3.4 for the Uniform Rate. Next, we prove the expected step complexity for the uniform rate. This proof is modified from the proof of (Ren et al., 2025, Theorem 4.9).

To simulate the approximate reverse process, in each discretized time interval $(t_k, t_{k+1}]$, uniformization requires that the Poisson rate Λ_k satisfies

$$\Lambda_k \geq \sup_{\substack{\mathbf{x}\in\mathcal{X}\\ t\in(t_k,t_{k+1}]}} -\tilde{Q}_t^{\leftarrow}(\mathbf{x},\mathbf{x}).$$

Thus, suppose we choose the Poisson rate to be

$$\Lambda_k = C\cdot \sup_{\substack{\mathbf{x}\in\mathcal{X}\\ t\in(t_k,t_{k+1}]}} -\tilde{Q}_t^{\leftarrow}(\mathbf{x},\mathbf{x}), \tag{95}$$

for some constant $C \geq 1$.

Thus, under the assumption that $\tilde{s}_t(\mathbf{x})_{\mathbf{y}} \asymp s_t(\mathbf{x})_{\mathbf{y}}$ when $Q_t(\mathbf{y},\mathbf{x}) > 0$ and $s_t(\mathbf{x})_{\mathbf{y}} \lesssim \max(1, \frac{1}{t})$, we have

$$-\tilde{Q}_t^{\leftarrow}(\mathbf{x},\mathbf{x}) = \sum_{\substack{\mathbf{y}\in[S]^d\\ \mathbf{y}\neq\mathbf{x}}} \tilde{Q}_t^{\leftarrow}(\mathbf{x},\mathbf{y}) = \sum_{\substack{\mathbf{y}\in[S]^d\\ \mathbf{y}\neq\mathbf{x}}} Q_t(\mathbf{y},\mathbf{x})\tilde{s}_t(\mathbf{x})_{\mathbf{y}} \lesssim \underbrace{\sum_{\substack{\mathbf{y}\in[S]^d\\ \mathbf{y}\neq\mathbf{x}}} Q_t(\mathbf{y},\mathbf{x})}_{\substack{d(S-1)\text{ nonzero terms}\\ \text{each term}=\frac{\beta(t)}{S}}} \underbrace{s_t(\mathbf{x})_{\mathbf{y}}}_{\substack{\lesssim\max\left(1,\frac{1}{t}\right),\\ \text{by assumption}}} \lesssim d\beta(t)\max\left(1,\frac{1}{t}\right). \tag{96}$$

Remark that the sum of independent Poisson(Λ_k) random variables is distributed as Poisson$(\sum_k \Lambda_k)$, the expected total number of steps are given by

$$\mathbb{E}[N] = \sum_{k=0}^{N_B-1} \Lambda_k(t_{k+1}-t_k),$$

where $\delta = t_0 < t_1 < \cdots < t_{N_B} = T$ is the time discretization,

$$\begin{aligned}&\lesssim \beta_{\max}\, d \sum_{k=0}^{N_B-1} \max\left(1,\frac{1}{t_k}\right)(t_{k+1}-t_k), \quad \text{by (95) and (96)},\\ &\leq \beta_{\max}\, d \left(\sum_{k=0}^{N_B'-1} \underbrace{\frac{t_{k+1}-t_k}{t_k}}_{\substack{\leq\frac{R-1}{\log R}\log\frac{t_{k+1}}{t_k},\\ \text{by Lemma I.5}}} + \underbrace{\sum_{k=N_B'}^{N_B-1}(t_{k+1}-t_k)}_{\leq T}\right),\end{aligned}$$

where N_B' is the unique index satisfying $t_{N_b'-1} \leq 1 < t_{N_b'}$, and $R = \max_k \frac{t_{k+1}}{t_k}$,

$$\begin{aligned}
&\lesssim \beta_{\max} d \left(\left[\sum_{k=0}^{N_B'-1} \log \frac{t_{k+1}}{t_k} \right] + T \right) \\
&\leq \beta_{\max} d \left(\left[\sum_{k=0}^{N_B-1} \log \frac{t_{k+1}}{t_k} \right] + T \right), \quad \text{as } N_B \geq N_b', \\
&= \beta_{\max} d \left(\int_{\delta}^{T} \frac{1}{t} \, dt + T \right) \\
&= \beta_{\max} d \left(\log(T) + \log\left(\delta^{-1}\right) + T \right). \qquad (97)
\end{aligned}$$

On the other hand, by Corollary 3.3, the approximation error under any IPM γ_Φ satisfies

$$\gamma_\Psi(p_{\text{data}}, \tilde{p}_\delta) \leq 2C_\Psi \left(d \int_0^\delta \beta(t) \, dt + d e^{-\|\beta\|_1} + \sqrt{2d} \sqrt{\|\beta\|_1} \sqrt{\mathcal{L}_{\text{SE}}^\delta(s, \tilde{s}) + \tfrac{2}{3} \mathcal{L}_3^\delta(s, \tilde{s})} \right).$$

For a given threshold $\epsilon > 0$ for the approximation error, it suffices to bound each of the three terms on the right-hand-side by $\epsilon/3$. This is achieved by

$$\delta < \frac{\epsilon}{6C_\Psi d \beta_{\max}} \;\Rightarrow\; \delta^{-1} > \frac{6C_\Psi d \beta_{\max}}{\epsilon}, \quad \text{thus we choose } \delta^{-1} \asymp \frac{dC_\Psi \beta_{\max}}{\epsilon}, \qquad (98)$$

$$T > \frac{1}{\beta_{\min}} \log\left(\frac{6C_\Psi d}{\epsilon}\right), \quad \text{thus we choose } T \asymp \frac{1}{\beta_{\min}} \log\left(\frac{C_\Psi d}{\epsilon}\right), \qquad (99)$$

$$\mathcal{L}_{\text{SE}}^\delta(s, \tilde{s}) + \tfrac{2}{3} \mathcal{L}_3^\delta(s, \tilde{s}) = \frac{\epsilon^2}{72 d C_\Psi^2 \|\beta\|_1} \;\Rightarrow\; \mathcal{L}_{\text{SE}}^\delta(s, \tilde{s}) \lesssim \frac{\epsilon^2}{d C_\Psi^2 \|\beta\|_1}. \qquad (100)$$

Plugging (98) and (99) into (97) yields

$$\begin{aligned}
\mathbb{E}[N] &\lesssim \beta_{\max} d \left(\log(T) + \log\left(\delta^{-1}\right) + T \right) \\
&= \mathcal{O}\left(\beta_{\max} d \left(\log\left(\delta^{-1}\right) + T \right) \right) \\
&= \mathcal{O}\left(\beta_{\max} d \left(\log \frac{dC_\Psi \beta_{\max}}{\epsilon} + \frac{1}{\beta_{\min}} \log \frac{dC_\Psi}{\epsilon} \right) \right) \\
&= \mathcal{O}\left(\beta_{\max} d \left(\log \beta_{\max} + \frac{1}{\beta_{\min}} \log \frac{dC_\Psi}{\epsilon} \right) \right).
\end{aligned}$$

□