AN ODE-BASED SAMPLER FOR DISCRETE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The marginals of a continuous diffusion model can be reproduced by a deterministic ordinary differential equation (ODE), known as the probability flow ODE, which provides a foundation for accelerated sampling and consistency distillation. In this paper we address the question whether an analogous formulation exists for discrete diffusion models and answer it affirmatively by introducing exponential residual flows (ERFs). Our framework augments the discrete state with a continuous residual variable for each token and position. These residuals evolve deterministically according to piecewise-defined ODEs whose vector fields depend on the current discrete state and the forward and backward rate matrices of the prescribed diffusion model. The discrete state changes whenever a residual reaches zero. We prove that this construction preserves the marginals of the original diffusion. We derive a practical discretization of our sampler and specialize it to masked and uniform diffusion. The deterministic trajectories of ERFs support consistency distillation and yield empirical improvements over previous approaches across a variety of benchmarks including code and text generation tasks.

1 INTRODUCTION

Discrete diffusion models have emerged as a compelling alternative to autoregressive models. By allowing multiple positions to be updated in parallel, they offer a path toward faster decoding (Lou et al., 2024; Sahoo et al., 2024) and potentially more efficient steering and control (Schiff et al., 2024). In continuous time, these models are described by forward and reverse continuous-time Markov chains (Campbell et al., 2022; Lou et al., 2024). Standard simulation of the reverse process is inherently stochastic: Gillespie's algorithm draws fresh randomness at each jump (Gillespie, 1977), while ancestral sampling of discrete-time reverse chains likewise relies on stochastic transitions (Shi et al., 2024; Sahoo et al., 2024; Hoogeboom et al., 2021; Schiff et al., 2024).

This raises a natural question: can one design a discrete diffusion sampler whose only randomness is the initial condition, as in the probability-flow ODE for continuous diffusion models (Song et al., 2021b;a)? A direct analogue on a finite state space is impossible because a deterministic map cannot split probability mass. Continuous auxiliary randomness is therefore required at initialization, particularly for masked diffusion, whose terminal distribution is a Dirac mass. Existing uses of auxiliary randomness serve different purposes. The projected Gaussian trajectories of Sahoo et al. (2025) depend jointly on the clean sample and auxiliary noise for distillation and do not define a deterministic flow driven solely by the current discrete state and an initial random seed. Gumbel Distillation (Zhang et al., 2026) instead uses Gumbel noise to construct deterministic training targets from an autoregressive teacher for parallel decoding, rather than a marginal-preserving deterministic flow for a discrete diffusion process.

To our knowledge, these approaches therefore do not provide a discrete analogue of the probability-flow ODE. In this work, we construct such an analogue. More precisely, our contributions are as follows:

- Our central contribution is a new deterministic sampler for discrete diffusion models, which we call the *exponential residual flow* (ERF). We augment the state with non-negative auxiliary variables, called residuals, initialized as independent exponential random variables, one for each token and position. Their evolution is then governed by an ODE determined by the forward and reverse rates of

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.

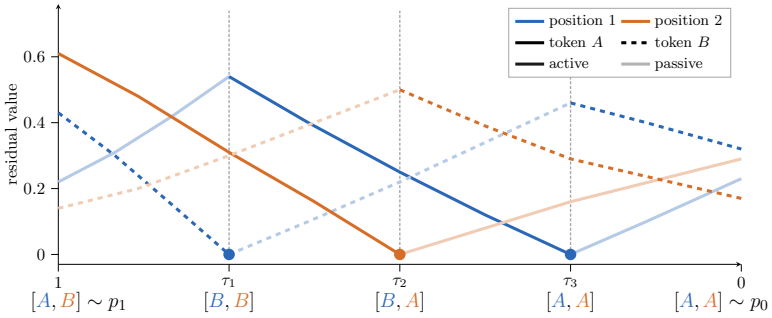

Figure 1: Exponential residual clocks in the deterministic sampler. Each clock triggers a jump when it reaches zero, then becomes passive and begins increasing again.

the discrete dynamics that governed jumps of the discrete state. A token change is triggered whenever a residual reaches zero. In contrast to Gillespie's first-reaction construction, where residual clocks are refreshed after each jump, ERF continues to evolve all remaining residuals deterministically. Consequently, once the initial augmented state has been sampled, no additional randomness is required. We prove that these augmented dynamics preserve the marginal distribution of the original CTMC at every time. At each fixed time, the residuals remain independent standard exponential random variables and are independent of the discrete state. The resulting process generally differs from the reverse Markov chain and need not be Markovian.

- To make this construction applicable to learned diffusion models, we introduce a novel numerical integration scheme that freezes the leave-one-out denoiser over each sampling interval while preserving the known time dependence of the transition rates. We derive explicit formulations for both masked and uniform diffusion models, and show that the resulting scheme is exact for factorized data distributions when the denoiser is exact.

- We evaluate ERF against ancestral sampling under matched model-evaluation budgets and find comparable generation quality. The deterministic nature of ERF also makes it directly amenable to consistency distillation. We exploit this property to train few-step student models and evaluate the resulting speed–quality trade-offs on text generation, code generation, and Sudoku completion.

Notation. We consider integer-valued sequences in $\mathsf{X} = \mathsf{V}^L$, where $\mathsf{V} = [\![1, K]\!]$, and write $\Delta_K = \{p \in \mathbb{R}^K : p^k \geq 0,\ \sum_{k=1}^K p^k = 1\}$ for the probability simplex. For $x \in \mathsf{X}$, x^ℓ denotes its ℓ-th token and $x^{-\ell} \in \mathsf{V}^{L-1}$ the sequence with position ℓ removed. For masked diffusion, the mask token is $\mathsf{m} := K$. For any vector $x = (x^1, \dots, x^d) \in \mathsf{V}^d$, we denote its one-hot matrix by $\tilde{x} \in \{0,1\}^{d \times K}$, with entries $\tilde{x}^{\ell k} := \mathbb{1}_{x^\ell = k}$. For a single token $x \in \mathsf{V}$, $\tilde{x} \in \mathbb{R}^K$ denotes its one-hot vector. For $\pi \in \Delta_K$, $\mathrm{Cat}(\pi)$ denotes the categorical distribution, and $\mathrm{Cat}(x; \pi) := \langle \tilde{x}, \pi \rangle$ is the probability of x under $\mathrm{Cat}(\pi)$. We use the convention $\sup \emptyset = -\infty$.

2 Background on Discrete Diffusion Models

We first briefly introduce discrete diffusion models (Sahoo et al., 2024; Shi et al., 2024; Austin et al., 2021; Hoogeboom et al., 2021; Schiff et al., 2024). This class of generative models constructs a stochastic process that transports a simple reference distribution, $p_1 := \prod_{\ell=1}^L \mathrm{Cat}(\pi^\ell)$, with $\pi \in (\Delta_K)^L$, to a target data distribution p_0. To this end, we use a forward process that progressively corrupts the data distribution, ultimately transforming it into p_1. The generative model is then obtained by learning an approximation of the corresponding reverse process.

We consider as forward process a Continuous Time Markov Chain (CTMC) process $(X_t)_{t \in [0,1]}$ started at p_0 with transitions $q_{t|s}(x_t|x_s) = \prod_{\ell=1}^L q^\ell_{t|s}(x^\ell_t|x^\ell_s)$, where $x_s, x_t \in \mathsf{X}$, $0 \leq s < t \leq 1$ and

$$q^\ell_{t|s}(x^\ell_t|x^\ell_s) = \mathrm{Cat}\big(x^\ell_t; \alpha_{t|s}\tilde{x}^\ell_s + (1 - \alpha_{t|s})\pi^\ell\big) \ , \quad 0 \leq s < t \leq 1 \ , \tag{1}$$

where $\tilde{x}_s$ denotes the one-hot encoding of x_s and $(\alpha_t)_{t \in [0,1]}$ is a monotone noise schedule with $\alpha_0 = 1$, $\alpha_1 \approx 0$, and $\alpha_{t|s} = \alpha_t / \alpha_s$. The time-reversal of this process is obtained by reversing the direction of time and is defined as the process $(X_{1-t})_{t \in [0,1]}$. Such a process allows by construction

to transport the source distribution p_1 to the target p_0. It is Markov with transitions given by

$$p_{s|t}(x_s|x_t) = \sum_{x_0} q_{s|0,t}(x_s|x_0, x_t)\, p_{0|t}(x_0|x_t)\ , \tag{2}$$

with $p_{0|t}(\cdot|x_t)$ the denoising posterior and $q_{s|0,t}$ the conditional distribution of X_s given (X_0, X_t), whose explicit expression is given in Appendix C. During training, these reverse transitions are learned using a factorized model parameterized by θ: $p^\theta_{s|t}(x_s|x_t) \coloneqq \prod_{\ell=1}^L p^{\theta,\ell}_{s|t}(x^\ell_s|x_t)$. To obtain a practical model for inference, we then consider the following joint distribution:

$$p^\theta_{0:n}(x_{t_{0:n}}) = p_1(x_{t_n}) \prod_{i=1}^n p^\theta_{s_i|t_i}(x_{s_i}|x_{t_i})\ , \tag{3}$$

for a choice of timesteps $0 = t_0 < \ldots < t_n = 1$, setting $s_i = t_{i-1}$ for notational convenience. A common parameterization used in the literature (Shi et al., 2024; Sahoo et al., 2024; Hoogeboom et al., 2021; Schiff et al., 2024) is $p^\theta_{s|t}(x_s|x_t) = q_{s|0,t}(x_s|\hat{x}^\theta_0(x_t, t), x_t)$, where $q_{s|0,t}$ denotes the extension of the bridge transition in (16) to simplex-valued inputs, and $\hat{x}^\theta_0$ denotes a neural network with values in $(\Delta_K)^L$. In what follows we assume, without any loss of generality, that it approximates the leave-one-out denoiser, whose ℓ-th component is $(t, x_t) \mapsto \mathbb{E}[X^\ell_0 \mid X^{-\ell}_t = x^{-\ell}_t]$; see Appendix C.

Rate matrices and discrete score. Another useful perspective on discrete diffusion is provided through rate matrices associated to the forward and backward processes (Campbell et al., 2022; Sun et al., 2023; Lou et al., 2024). The rate matrix associated to the forward transitions (1) is given for any $x, y \in \mathsf{X}$ and $t \in [0, 1]$ by

$$Q_t(y|x) = \sum_{\ell=1}^L \mathbb{1}_{y^{-\ell}=x^{-\ell}} Q^\ell_t(y^\ell|x^\ell)\ , \quad \text{where} \quad Q^\ell_t(y|x) = (-\dot{\alpha}_t/\alpha_t)\left\langle \tilde{y}, \pi^\ell \right\rangle \quad (x \neq y)\ . \tag{4}$$

Here $x, y \in \mathsf{V}$, and $Q^\ell_t(x|x) = -\sum_{y\neq x} Q^\ell_t(y|x)$ for all $\ell \in [\![1, L]\!]$. Let $(p_t)_{t\in[0,1]}$ denote the family of marginals of $(X_t)_{t\in[0,1]}$ with transition probabilities given by (1) and initial distribution p_0. By definition, $(p_t)_{t\in[0,1]}$ satisfies the Kolmogorov equation (Feller, 1940, Theorem 1): for $t \in (0, 1)$ and $x, y \in \mathsf{X}$,

$$\partial_t p_t(y) = \sum_{x\in\mathsf{X}} Q_t(y|x) p_t(x)\ . \tag{5}$$

Note that the forward transition $q_{t|s}(\cdot|x_s)$ specified in (1) are obtained by solving (5) started from δ_{x_s} at time s. On the other hand, the family of rate matrices $(\bar{Q}_t)_{t\in[0,1]}$ associated to the backward process satisfies for any $y \neq x$, such that $p_t(x) > 0$[1],

$$\bar{Q}_t(y|x) \coloneqq \frac{p_t(y)}{p_t(x)} Q_t(x|y) = \sum_{\ell=1}^L \mathbb{1}_{y^{-\ell}=x^{-\ell}} \frac{p^\ell_t(y^\ell|x^{-\ell})}{p^\ell_t(x^\ell|x^{-\ell})} Q^\ell_t(x^\ell|y^\ell)\ , \tag{6}$$

with $\bar{Q}_t(x|x) \coloneqq -\sum_{y\neq x} \bar{Q}_t(y|x)$. Using the definition of the reverse transitions in (2), (5) and (6), we obtain that the reverse transitions solve the following Kolmogorov equation: for any $x_t, y \in \mathsf{X}$, $p_t(x_t) > 0$,

$$\partial_s p_{s|t}(y|x_t) = -\sum_x \bar{Q}_s(y|x) p_{s|t}(x|x_t)\ , \qquad p_{t|t}(y|x_t) = \mathbb{1}_{x_t=y}\ . \tag{7}$$

Thus, initializing this CTMC from p_1 and running it backward generates the marginals $(p_t)_{t\in[0,1]}$.

3 Deterministic sampling via exponential residual flows

In this section, we construct a sampler whose trajectory is deterministic after a single random initialization and with one-time marginals agreeing with those of the reverse CTMC. The key idea is to augment the discrete state with the exponential variables used in Gillespie's first-reaction method and replace their resampling at jump times with a deterministic continuous evolution, resulting in the Exponential Residual Flow sampler described in Algorithm 1.

3.1 Exponential Residual Flow

Gillespie's algorithm. We recall Gillespie's algorithm, especially its first-reaction formulation Gillespie (2001), for simulating $(X_{1-t})_{t\geq 0}$ backward in time. We refer to Norris (1998) for the standard CTMC and Poisson-process results used here. A CTMC has piecewise-constant trajectories with random jump times. Given $X_t = x_t$, the total backward jump rate, or exit rate, $\bar{\lambda}_u(x_t) = -\bar{Q}_u(x_t|x_t) \geq 0$ for $u \leq t$ defines the intensity of an inhomogeneous Poisson process while x_t

[1]The definition of $\bar{Q}_t$ is arbitrary when $p_t(x) = 0$.

Algorithm 1 Gillespie / ERF algorithm.

Require: $0 \le s \le t \le 1$, reverse rates $(\bar{Q}_u)_{u\in[s,t]}$, forward exit rates $(\lambda_u)_{u\in[s,t]}$
Output: (X_s, E_s) at time s.

1: **Initialize:** $(X_t, E_t) \sim p_t \otimes \mathrm{Exp}^{\otimes(L\times K)}$, $X_s \leftarrow X_t$, $\tau^\star \leftarrow t$, $E_s \leftarrow E_t$
2: **while** $\tau^\star > s$ **do**
3: $\quad \tau^{\ell j} \leftarrow \sup\left\{u \in [s, \tau^\star] : \int_u^{\tau^\star} \bar{Q}_r^\ell(j|X_s)\,\mathrm{d}r \ge E_s^{\ell j}\right\}$ for $j \neq X_s^\ell$ ▷ in parallel
4: $\quad (\ell^\star, j^\star) \leftarrow \mathrm{argmax}_{(\ell,j):j\neq X_s^\ell}\ \tau^{\ell j}$
5: $\quad \tau^\star_{\mathrm{new}} \leftarrow \max\{s, \tau^{\ell^\star j^\star}\}$
6: $\quad E_s^{\ell j} \leftarrow E_s^{\ell j} - \int_{\tau^\star_{\mathrm{new}}}^{\tau^\star} \bar{Q}_r^\ell(j|X_s)\,\mathrm{d}r, \qquad j \neq X_s^\ell$ ▷ update active residuals
7: $\quad$ **if** $X_s^\ell = j$ **then** ▷ update passive residuals
8: $\quad\quad$ **either** $E_s^{\ell j} \sim \mathrm{Exp}(1)$ ▷ Gillespie
9: $\quad\quad$ **or** $E_s^{\ell j} \leftarrow E_s^{\ell j} + \int_{\tau^\star_{\mathrm{new}}}^{\tau^\star} \lambda_r^\ell(j)\,\mathrm{d}r$ ▷ ERF (ours)
10: $\quad$ **end if**
11: $\quad$ **if** $\tau^{\ell^\star j^\star} \ge s$ **then**
12: $\quad\quad X_s^{\ell^\star} \leftarrow j^\star$
13: $\quad$ **end if**
14: $\quad \tau^\star \leftarrow \tau^\star_{\mathrm{new}}$
15: **end while**
16: **return** X_s, E_s

is fixed. Concretely we obtain the jump time u by solving $\int_u^t \bar{\lambda}_r(x_t)\,\mathrm{d}r = E$ with $E \sim \mathrm{Exp}(1)$. Conditional on a jump at u, the next state is drawn with probabilities $\{\bar{Q}_u(y|x_t)/\bar{\lambda}_u(x_t)\}_{y\neq x_t}$. Repeating until the next jump falls below s gives the standard Gillespie simulation.

The categorical draw for the next state transition can be avoided by introducing independent Poisson processes for each of the $L(K-1)$ admissible transitions from x_t. This leads to the first-reaction formulation of Gillespie's algorithm. Suppose again that $X_t = x_t$. The Poisson process associated to a pair (ℓ, j), with $\ell \in [L]$ and $j \in [K] \setminus \{x_t^\ell\}$, has intensity $\bar{Q}_u^\ell(j|x_t)$ for $u \le t$. Its first arrival time $\tau^{\ell,j}$ appears in line 3 of Algorithm 1. Because time runs backward, the largest candidate time below the current time is the next jump: line 4 selects the pair $(\ell^\star, j^\star)$ and advances to its event time, or to s if no candidate exists. Throughout this interval, the discrete state remains equal to the current state x_t. The *active* residuals $E_t^{\ell,j}$ with indexes satisfying $j \neq x_t^\ell$, are translated in line 6, while *passive* residuals with $j = x_t^\ell$ are independently refreshed in line 8. If an event occurs we set $X_t^{\ell^\star}$ to $j^\star$ while the rest of the coordinates stay fixed. Using the memoryless property of the exponential distribution, the updated active residuals are still exponential random variable independent of the past. Iterating until no candidate event remains in $[s, \tau^\star]$ exactly simulates the backward process and is equivalent to the standard Gillespie simulation. We use this formulation below because it closely matches our deterministic construction; Appendix D.1 gives a formal construction of the full process $(X_t)_{t\in[0,1]}$.

Exponential Residual Flow. The exponential residual flow (ERF) replaces the stochastic sampling of Gillespie by a deterministic trajectory. To this end, we lift the discrete dynamics to the extended state $(X_t, E_t) \in \mathsf{X} \times [0,\infty)^{L\times K}$. It consists in a simple modification in Algorithm 1 changing the only step where randomness is introduced, *i.e.*, line 8 by the step in line 9. More precisely, we can deduce from the algorithm a process over $[0,1]$, $(X_t, E_t)_{t\in[0,1]}$. The algorithm defines the process at successive jump times $(\tau_k)_{k\in\ge 0}$. Then, for any k, on $(\tau_{k+1}, \tau_k]$, X_t is kept constant equal to X_{τ_k} and E_t solves the system of ODEs:

$$\frac{\mathrm{d}}{\mathrm{d}t}E_t^{\ell j} = \begin{cases} \bar{Q}_t^\ell(j|X_t), & j \neq X_t^\ell, \\ -\lambda_t^\ell(X_t^\ell), & j = X_t^\ell. \end{cases} \tag{8}$$

We call the process $(X_t, E_t)_{t\in[0,1]}$ the exponential residual flow (ERF) process. A simple illustration of this construction is provided in Figure 1. We show in our next result that the resulting marginal process $(X_t)_{t\ge 0}$ admits the same time marginals as the original CTMC. For completeness, a more formal construction of this process alongside Gillespie's first-reaction form is given in Appendix D.1.1.

Theorem 3.1 (Correctness of the exponential residual flow)**.** *Under standard assumptions stated in Appendix J the ERF process is almost surely well defined with finitely many jumps, and, for every* $t \in [0,1]$*,*

$$\mathcal{L}(X_t, E_t) = p_t \otimes \mathrm{Exp}(1)^{\otimes(L\times K)} \ . \tag{9}$$

To the best of our knowledge, Theorem 3.1 yields a novel deterministic sampling procedure for a CTMC that draws randomness only at initialization, similarly to the probability flow ODE in the continuous case (Song et al., 2021a).

We motivate the choice of the velocity (8) in Appendix J.1 by deriving a continuity equation on the extended state space from the Kolmogorov equation (5) satisfied by $(p_t)_{t\in[0,1]}$. Its vector field is precisely the one in (8). This argument, however, does not by itself prove Theorem 3.1. A complete proof instead characterizes by induction the density of the process at successive jump times and decomposes the path space according to the number of jumps. We give this proof in Appendix J.

3.2 Integrating the exponential residual flow

Implementing Algorithm 1 in practice raises computational issues that we now describe. For $y \neq x$, the reverse rates in (6) are commonly approximated as

$$\bar{Q}_t^\theta(y|x) \coloneqq \textstyle\sum_{\ell=1}^L \mathbb{1}_{y^{-\ell}=x^{-\ell}} \left\langle \tilde{y}^\ell, s_t^\theta(x)^\ell \right\rangle Q_t^\ell(x^\ell|y^\ell) \ , \tag{10}$$

where the score approximation $s_t^\theta(x)^\ell$ estimates the density ratios on the r.h.s. in (6) (Lou et al., 2024). We must then estimate the integrals of these approximate reverse rates in lines 3-6 and those of the associated forward exit rates in line 9. However, exact integration may be computationally challenging or infeasible.

To address this problem, most samplers introduce discretization strategies for these integration steps and consider a grid $0 = t_0 < \cdots < t_n = 1$. A first option is to adapt the τ-leaping algorithm by defining an approximate process successively on $[s_i, t_i]$: given X_{t_i}, $(X_u)_{u\in[s_i,t_i]}$ is the CTMC with constant reverse rate $\bar{Q}_u^\theta(y; t_i|x) = \bar{Q}_{t_i}^\theta(y|X_{t_i})$; more details are provided in Appendix D.4. Another option, given the form of the reverse rate and its approximation (10), is to freeze only the score approximation at time t_i, while retaining the forward rates $Q_u^\ell(y^\ell|x^\ell)$ in (4), whose time integrals are available in closed form, similarly to (Conforti et al., 2025). We develop a different alternative here. First, note that the score is related to the leave-one-out denoiser by

$$\langle \tilde{y}^\ell, s_t^\theta(x_t)^\ell \rangle = q_{t|0}(y^\ell|\hat{x}_0^\theta(x_t,t)^\ell) / q_{t|0}(x_t^\ell|\hat{x}_0^\theta(x_t,t)^\ell) \ . \tag{11}$$

This relation suggests to freeze only the leave-one-out denoiser at time t_i since if we plug (11) in (10), the integration with respect to time of the resulting approximate reverse rate can be made explicit for most models of interest. In addition, this choice also yields an exact sampler for a single token or a factorized distribution when the leave-one-out predictor is optimal; see Appendix D.5.2.

Practical implementation. Given $(\hat{X}_{t_i}, \hat{E}_{t_i})$, a single discretization step produces $(\hat{X}_{s_i}, \hat{E}_{s_i})$ by applying Algorithm 1 on $[s_i, t_i]$ with the following approximate reverse rate matrix:

$$\bar{Q}_r^\theta(y; t_i|x) \coloneqq \sum_{\ell=1}^L \mathbb{1}_{y^{-\ell}=x^{-\ell}} \frac{q_{r|0}^\ell(y^\ell|\hat{x}_0^\theta(\hat{X}_{t_i}, t_i)^\ell)}{q_{r|0}^\ell(x^\ell|\hat{x}_0^\theta(\hat{X}_{t_i}, t_i)^\ell)} Q_r^\ell(x^\ell|y^\ell) \ , \qquad r \in [s_i, t_i]. \tag{12}$$

For the forward exit rates, we keep the original rates $(\lambda_r^\ell)_{r\in[s_i,t_i]}$. In general, discretizing ERF requires the rates obtained by time-reversing the frozen reverse process on $[s_i, t_i]$, and using those of the original diffusion can lead to infinitely many jumps. For leave-one-out freezing, however, (12) is the exact time reversal, on $[s_i, t_i]$, of the original forward process started at the factorized distribution given by $\hat{x}_0^\theta(\hat{X}_{t_i}, t_i)$. Each interval is therefore an exact ERF step for this distribution and has finitely many jumps; see Appendix D.5.

Specialization to MDM and UDM. Having specified the partially frozen reverse rate matrix in (12), we now specify the calculation of jump times in line 3-4 of Algorithm 1 to MDM and UDM. For a current state X_r at time r for $r \in [s_i, t_i]$ and each active residual $E_r^{\ell j}$, the candidate time is obtained by solving $\int_u^r \bar{Q}_v^{\theta,\ell}(j; t_i|X_r)\, \mathrm{d}v = E_r^{\ell j}$ for u (line 3 of Algorithm 1). For MDM this is

straightforward since when $X_r^\ell = \mathsf{m}$,

$$\int_u^r \bar{Q}_v^{\theta,\ell}(j; t_i|z)\, \mathrm{d}v = \hat{x}_0^\theta(X_{t_i}, t_i)^{\ell j} \log\left(\frac{1-\alpha_r}{1-\alpha_u}\right) \tag{13}$$

and its candidate time follows by solving for u. Unmasked positions cannot jump so their candidate jump times are not computed. For UDM, however, the reverse rate integral is not explicitly invertible in u, and each candidate time must therefore be approximated by bisection. Since UDM can produce many jumps within a single interval, repeatedly solving these equations can be computationally costly. To control this cost, we split each interval into k subintervals and allow at most one jump per position in each, advancing all residuals to the subinterval boundary after that jump; see Appendix D.5. To reduce *both* the number of jumps and the cost of computing their times, we exploit the fact that different rate matrices can generate the same marginals $(p_t)_{t\in[0,1]}$ (Campbell et al., 2024; 2022; Gat et al., 2024) and use the net-flux reverse rates (Zhang et al., 2023), defined for $x \neq y$ as

$$\bar{Q}_t^{\mathrm{NF}}(y|x) := \frac{\left[p_t(x)\bar{Q}_t(y|x) - p_t(y)\bar{Q}_t(x|y)\right]_+}{p_t(x)} = -\sum_{\ell=1}^{L} \mathbb{1}_{y^{-\ell}=x^{-\ell}} \frac{\dot{\alpha}_t}{K\alpha_t}\left[\frac{p_t(y^\ell|x_t^{-\ell})}{p_t(x_t^\ell|x_t^{-\ell})} - 1\right]_+ \tag{14}$$

where $a_+ = \max\{a, 0\}$ and $\bar{Q}_t$ is the canonical UDM rate matrix. Surprisingly, the UDM net-flux forward and backward rates are tractable and their time integrals are also analytically invertible which eliminates the need for bisection search for the jump times and greatly increases computational speed, as we show in the experiments. The resulting updates are given in Appendix B, with the net-flux construction in Appendix E.

4 Application to Distillation

As in the continuous diffusion case (Salimans & Ho, 2022; Song et al., 2023), one application of deterministic sampling is to distill a teacher diffusion model into a student model that performs better with few sampling steps. The absence of a direct analogue of the probability flow ODE in the discrete setting has often been used as an argument against trajectory-based distillation (Sahoo et al., 2025; Deschenaux & Gulcehre, 2025; Zhang et al., 2023). Our exponential residual flow supplies deterministic transport between augmented states (X_t, E_t) and (X_s, E_s), allowing student predictions to be matched along a teacher trajectory.

Consistency distillation. We rely on consistency distillation (CD) following Song et al. (2023). From an augmented state, the student predicts the clean discrete state at time 0 reached by the teacher trajectory, it does not need to predict the endpoint exponential residuals. We train its predictions from two points on the same trajectory to agree. Writing $\Psi_{s|t}^\theta$ for the ERF transport of Algorithm 1, these points satisfy $(X_s, E_s) = \Psi_{s|t}^\theta(X_t, E_t)$ for $s < t$. In a previous approach, Sahoo et al. (2025) couple X_s and X_t through sampled Gaussian noise. Their construction is deterministic conditional on this same noise, but the transition remains stochastic when conditioned only on X_t, which contradicts what is required in CD. In contrast, our approach deterministically produces (X_s, E_s) from (X_t, E_t).

Since the ERF output at time 0 is uniquely determined by the augmented state at time t, a naive approach is to train a student to approximate this discrete-valued map directly. To make the learning problem more tractable, we instead relax the output at each position to an element of the probability simplex. More precisely, we parameterize the student as a factorized distribution over discrete states:

$$p_{0|t}^\phi(x_0|x_t, e_t) := \textstyle\prod_{\ell=1}^{L} p_{0|t}^{\phi,\ell}(x_0^\ell|x_t, e_t)\ .$$

For one current parameter ϕ and target parameter $\tilde{\phi}$, we consider the loss function:

$$\mathcal{L}_{\mathrm{CD}}(\phi; \bar{\phi}) = \mathbb{E}\left[\mathrm{KL}(p_{0|s}^{\bar{\phi}}(\cdot|X_s, E_s) \| p_{0|t}^{\phi}(\cdot|X_t, E_t))\right]\ , \tag{15}$$

where $(X_t, E_t) \sim p_t \otimes \mathrm{Exp}(1)^{\otimes L\times K}$, $(X_s, E_s) = \Psi_{s|t}^\theta(X_t, E_t)$, t is sampled from $\mathrm{Unif}(\delta, 1)$, and $s = t - \delta$, where δ is a hyperparameter. The target parameter $\bar{\phi}$ is either updated by EMA after each step or frozen for a training round and refreshed between rounds. As noted in Song et al. (2023), for the minimizer to be exact, we need to impose the boundary condition $p_{0|0}^\phi(\cdot|x_0, e_0) = \delta_{x_0}$ and give a non-zero probability to $s = 0$ so that at optimality, the student assigns all its mass to the clean endpoint of the ERF trajectory, and marginalizing over the augmented states at time t, these

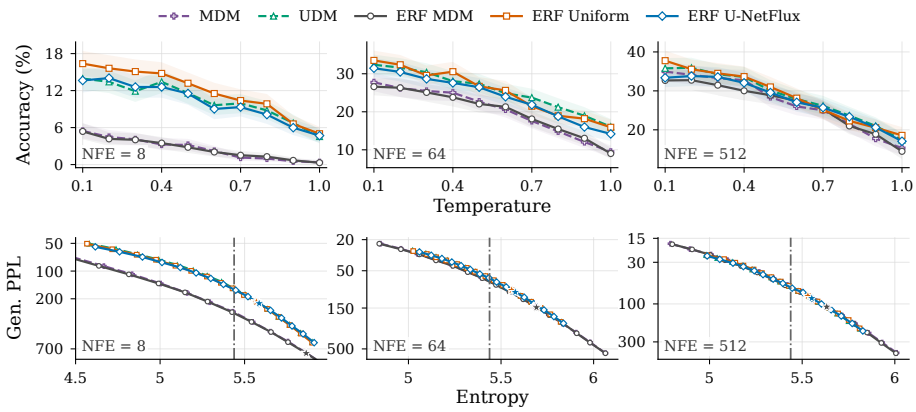

Figure 2: Sampling quality at matched NFE on TinyGSM (top) and OpenWebText (bottom). Dashed curves denote ancestral samplers; solid curves denote ERF. TinyGSM bands are 95% bootstrap intervals over questions. OWT curves trace the generative perplexity versus 1-gram entropy frontier by sweeping the temperature in $[0.8, 1.1]$; $\star$ refers to temperature 1, and the vertical dash-dotted line marks the OWT dataset entropy.

endpoints have law p_0. In practice, the boundary condition did not improve the performance so we don't use it. We also consider a biased ERF sampler that uses the frozen student in place of the pretrained teacher denoiser when computing the ERF rates. For UDM, we convert the leave-one-out student output to a denoising prediction before evaluating the loss using (29). The fixed-teacher and joint-refresh procedures are given in Algorithms 3 and 4.

Sampling. We consider the *refresh* and *bridge* transitions for the discrete state which are respectively $\sum_{x_0} q_{s|0}(x_s|x_0)\, p^{\phi}_{0|t}(x_0|x_t, e_t)$ and $\sum_{x_0} q_{s|0,t}(x_s|x_0, x_t)\, p^{\phi}_{0|t}(x_0|x_t, e_t)$, with fresh exponential variables drawn independently after each transition. Refresh is exact when the teacher transports the prescribed augmented marginals and the student learns its endpoint map. Bridge is not exact in general, but performs well empirically. Both procedures are given in Algorithm 5.

5 Experiments

Our experiments first study the empirical consequences of ERF's new joint law, showing that the resulting sampler achieves generation quality comparable to standard ancestral samplers at matched NFE. We then show that ERF consistency distillation (ERC) enables competitive generation with fewer NFEs than the teacher model.

General setting. We consider masked and uniform diffusion models trained on OpenWebText (OWT; Gokaslan & Cohen, 2019) and TinyGSM (Liu et al., 2023), using DDiT backbones with 170M and 168M parameters, respectively. OWT generation is unconditional with sequences of length 1,024; we measure GPT-2 Large generative perplexity (Radford et al., 2019) together with 1-gram entropy. TinyGSM models generate Python solutions on a 512-token canvas with fixed question tokens, and we evaluate answer accuracy on GSM8K (Cobbe et al., 2021), using all 1,319 test questions for the sampling comparison and the same first 1,024 questions for the main distillation comparison. For sampling, we also evaluate the released instruction-tuned DiffusionGemma model, with 26B total and approximately 4B active parameters (Google DeepMind, 2026), on MMLU-Pro (Wang et al., 2024).

5.1 Evaluation of the ERF sampler

Sampling quality. Within each diffusion family, we compare ERF and ancestral sampling using the same pretrained checkpoint at matched NFE budgets. We use a uniform time grid and compute the ERF Uniform jump times in line 3 of Algorithm 1 using 40 bisection iterations per solve. Further evaluations appear in Appendix H.1. The results are given in Figure 2. ERF achieves sampling quality comparable to the ancestral MDM and UDM samplers across the evaluated datasets and NFE budgets. On TinyGSM, the samplers exhibit similar accuracy and temperature dependence; on OWT, their entropy–perplexity curves across temperatures largely overlap within each diffusion family.

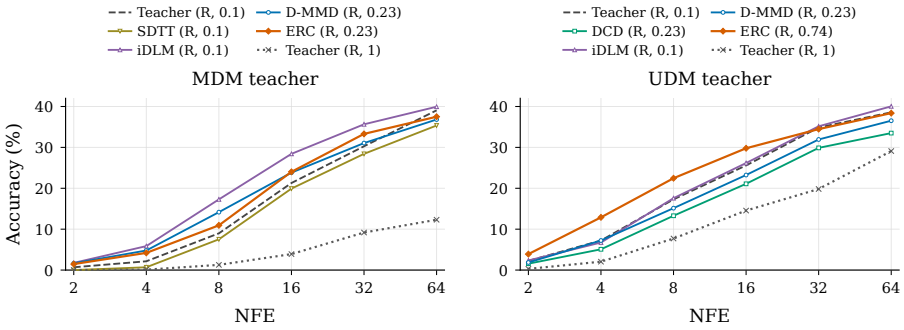

Figure 3: TinyGSM distillation at 2–64 NFEs. Each curve fixes its sampler and temperature to maximize mean accuracy across NFEs on the same 1,024 questions. ERC selects $\delta = 1/64$ for MDM and $1/8$ for UDM. R refers to refresh sampling.

The DiffusionGemma results support the same conclusion at a larger model scale (Table 1). At 48 NFEs and temperature 1, ERF Uniform and U-NetFlux are within 0.7 percentage points of the UDM ancestral sampler on MMLU-Pro questions. See Appendix I for experiment details.

Table 1: DiffusionGemma on 140 MMLU-Pro questions and three paired seeds. Brackets give 95% question-cluster bootstrap intervals.

Sampler	Accuracy (%)	s/question
Ancestral	57.4 [50.5, 64.3]	59.4
ERF Uniform	56.7 [49.8, 63.6]	61.1
ERF U-NetFlux	57.1 [50.2, 64.0]	60.4

ERF Uniform vs U-NetFlux. The U-NetFlux version is the more efficient ERF sampler for uniform diffusion because we can solve explicitly for the crossing time in line 3 without bisection; see Appendix E. Furthermore, it makes substantially fewer jumps and thus requires fewer solves. At 8 NFEs, trajectories average approximately one jump per generated position, compared with seven for Uniform, yielding speedups of 2.83× on TinyGSM and 2.00× on OWT (Table 2). This gain comes while maintaining similar sampling quality in Figure 2. We also evaluate ERF Uniform with a cap on jumps within the interval $[s_i, t_i]$; Figures 10 and 11 show its effect on sampling quality.

Table 2: Runtime and jump averages at 8 NFEs over three batches; times are seconds per sequence and jumps are per generated position.

Sampler	Time (s)	Jumps
TinyGSM		
Uniform	0.368	6.90
U-NetFlux	**0.130**	**1.02**
OWT		
Uniform	0.647	7.11
U-NetFlux	**0.323**	**1.04**

5.2 Distillation

Setting and baselines. We compare ERC with SDTT (Deschenaux & Gulcehre, 2025) for masked diffusion, DCD (Sahoo et al., 2025) for uniform diffusion, and D-MMD (Hoogeboom et al., 2026) and iDLM (Li et al., 2026) for both families. All results presented are from our own re-implementations and training runs, with all the students initialized from the corresponding pretrained teacher. Within each dataset and diffusion family, the methods share the *same* teacher checkpoint. For both MDM and UDM, the teacher models are trained using the standard ELBOs; Sahoo et al. (2024, Equation 11) and Schiff et al. (2024, Equation 19). In particular, for UDM our pre-trained model is a LOO denoiser. All the implementation details are given in Appendices F.5 and F.6.

Training. ERC conditions the student on Gumbel variables $-\log E_t$ through a learned projection; UDM ERC uses net-flux ERF rates. On both datasets we run all the methods with the same GPU-hour budget (100 for OWT and 50 for TinyGSM) and batch size 128. ERC uses four equal training rounds and the joint target-and-teacher refresh in Algorithm 4. We sweep $\delta \in \{1/8, 1/16, 1/32, 1/64\}$ and select one value per dataset and diffusion family using the evaluation criteria below. ERC uses forward KL on OWT and reverse KL on TinyGSM. For UDM, we apply the LOO-to-denoiser conversion formula to the student output before computing the loss (29). This gave improved results on all our experiments. Appendix H.5 gives the training and evaluation details. We disable the top-p truncation used by D-MMD (Hoogeboom et al., 2026) and fix the shared teacher's temperature to 1, so that all methods distill the same teacher distribution.

Sampling protocol. We evaluate all methods and the teacher with both *refresh* and *bridge* sampling at matched NFE, sweeping a common temperature grid to account for differences in logit

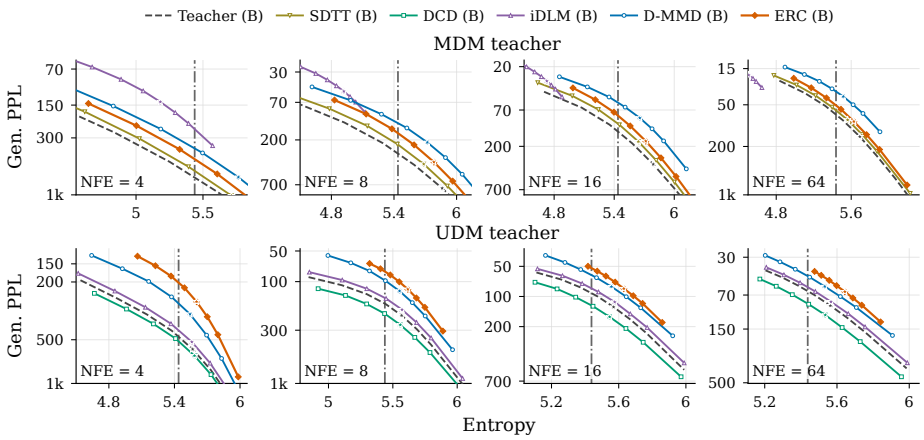

Figure 4: OWT distillation with bridge sampling at 4, 8, 16, and 64 NFEs, with students trained for 100 GPU-hours. Each point uses 2,048 samples, with temperature applied to denoiser probabilities over $[0.8, 1.2]$; $\star$ marks temperature 1, and the vertical dash-dotted line marks the OWT dataset entropy. B refers to bridge sampling.

sharpness across objectives. On TinyGSM, we select one sampler, temperature placement and temperature per method, together with one ERC gap δ, and hold these choices fixed along each curve. On OWT, the main figure compares generative perplexity–entropy curves (Pynadath et al., 2026). In the TinyGSM legend, B and R denote bridge and refresh sampling; the number gives the selected inference temperature. The OWT legend marks every method with B to indicate bridge sampling.

Results. On TinyGSM, UDM ERC doubles the accuracy of the strongest competing baseline at 2 and 4 NFEs and also records higher accuracy at 8 and 16; by 32 NFEs its accuracy is close to the teacher's. In this evaluation, it is the only method with accuracy above the teacher at some NFE budgets when fully sweeping temperature and samplers (even for the teacher). For MDM, ERC approximately matches the second-best baseline, see Figure 3. On OWT we see the same trend, UDM ERC has lower perplexity at matched entropy than the other baselines on the full frontier, while MDM ERC approximately matches the second-best baseline, see Figure 4. On Sudoku, ERC gives modest but consistent gains over the other methods in many evaluated settings. The complete sampler comparisons and Sudoku results appear in Appendices H.6 and H.9. Across these comparisons, the largest ERC gains occur with uniform diffusion, while masked-diffusion results remain competetitive with the considered baselines.

Ablations. We vary how the target and ERF transport (providing X_s starting from X_t in (15)) are updated, the training gap δ, the inference sampler, and whether ERC receives Gumbel inputs. For UDM, we also compare two placements of sampling temperature. The protocols and complete grids appear in Appendices H.7 to H.10.

6 Conclusion

In this work, we show that the prescribed marginals of a discrete diffusion can be transported by deterministic dynamics once the token state is augmented with continuous residuals. This construction generalizes Gillespie's algorithm by recycling the exponential variables drawn at every jump according to an appropriate ODE. This gives a pathwise target for consistency distillation. In practice, leave-one-out freezing retains the analytically known time dependence of the rates and is exact for factorized data. Across the evaluated tasks, ERF achieves sampling quality comparable to ancestral updates at matched model evaluations, and its consistency objective yields its clearest few-step gains for uniform diffusion.

AI USE STATEMENT

A generative AI assistant was used to support literature searches, discuss mathematical expressions and experimental results, and assist with manuscript drafting, editing, and LaTeX preparation. The authors take responsibility for all claims, citations, derivations, reported results, and the final content of the paper.

ETHICS STATEMENT

This work studies sampling and distillation methods for generative models and does not involve human-subject experiments. More efficient generation can also facilitate harmful uses of text and code models, while biases or unsafe behavior in a teacher model may persist in its distilled student. These risks should be considered when deploying models based on the proposed methods.

REPRODUCIBILITY STATEMENT

The method and its assumptions are described in Sections 3 and 4, with derivations and proofs in the appendix. Appendices G and H document the datasets, training and evaluation protocols, and additional results needed to assess the empirical claims.

REFERENCES

David F. Anderson. A modified Next Reaction Method for simulating chemical systems with time dependent propensities and delays. *The Journal of Chemical Physics*, 127(21):214107, 2007. doi: 10.1063/1.2799998.

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.

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, 2024. URL https://arxiv.org/abs/2402.04997.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

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

Justin Deschenaux and Caglar Gulcehre. Beyond autoregression: Fast LLMs via self-distillation through time. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uZ5K4HeNwd.

Luca Eyring, Vincent Pauline, Stefan Bauer, Alexey Dosovitskiy, and Zeynep Akata. Ddno: Discrete diffusion noise optimization. 2026.

Eugene A. Feinberg, Manasa Mandava, and Albert N. Shiryaev. On solutions of kolmogorov's equations for jump markov processes, 2013. URL https://arxiv.org/abs/1301.6998.

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

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f0d629a734b56a642701bba7bc8bb3ed-Abstract-Conference.html.

Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

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

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Google DeepMind. DiffusionGemma. https://deepmind.google/models/gemma/diffusiongemma/, June 2026. Accessed: 2026-06-10.

Samson Gourevitch, Yazid Janati, Dario Shariatian, Umut Simsekli, Eric Moulines, Eric P. Xing, and Alain Durmus. Uniform diffusion models revisited: Leave-one-out denoiser and absorbing state reformulation, 2026. URL https://arxiv.org/abs/2605.22765.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows: Learning categorical distributions with normalizing flows. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

Emiel Hoogeboom, David Ruhe, Jonathan Heek, Thomas Mensink, and Tim Salimans. Beyond single tokens: Distilling discrete diffusion models via discrete MMD, 2026. URL https://arxiv.org/abs/2603.20155.

David Li, Nikita Gushchin, Dmitry Abulkhanov, Eric Moulines, Ivan Oseledets, Maxim Panov, and Alexander Korotin. Idlm: Inverse-distilled diffusion language models, 2026. URL https://arxiv.org/abs/2602.19066.

Bingbin Liu, Sebastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. TinyGSM: achieving ¿80% on GSM8K with small language models, 2023. URL https://arxiv.org/abs/2312.09241.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Generative frontiers: Why evaluation matters for diffusion language models, 2026. URL https://arxiv.org/abs/2604.02718.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *arXiv preprint arXiv:2506.10892*, 2025.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*, 2024.

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

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.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23a.html.

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.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *arXiv preprint arXiv:2311.18828*, 2023.

Chi Zhang, Xixi Hu, Bo Liu, and Qiang Liu. Gumbel distillation for parallel text generation, 2026. URL https://arxiv.org/abs/2603.22216.

Pengze Zhang, Hubery Yin, Chen Li, and Xiaohua Xie. Formulating discrete probability flow through optimal transport, 2023. URL https://arxiv.org/abs/2311.03886.

APPENDIX CONTENTS

A Related Work **14**

B Rates and jump times used in the discretization **15**

C Bridge and simplex extension **16**

D CTMC integration schemes **16**

D.1 Gillespie's algorithm 16

D.1.1 Gillespie's first reaction form and ERF sampler 16

D.1.2 Gillespie's global form 17

D.2 Gillespie discretization with partially frozen rates 17

D.3 Equivalence of leave-one-out-frozen Gillespie and ancestral sampling 17

D.4 Euler and categorical tau-leaping 19

D.5 Explicit partially frozen updates for MDM and UDM 19

D.5.1 From integrated rates to jump updates 20

D.5.2 Exactness of leave-one-out freezing under factorization 21

D.5.3 A counterexample for score freezing 22

D.5.4 Why the explosion cannot occur with leave-one-out freezing 22

E Net-flux rate matrix **23**

E.1 Net flux construction and relation to prior work 23

E.2 Forward net-flux rates and passive residuals 24

F Additional material for distillation **24**

F.1 Consistency distillation training 24

F.2 Multi-step sampling 26

F.3 Bridge sampling and the ERF coupling 26

F.4 D-MMD 26

F.5 DCD 27

F.6 iDLM 28

G Experimental details **28**

G.1 Sudoku 28

H Additional experiments **30**

H.1 Sampling quality across NFE budgets 30

H.2 Euler and categorical tau-leaping 32

H.3 Uniform substeps 34

H.4 Runtime and jump-count protocol 37

H.5 Distillation evaluation 38

H.6 Sudoku distillation 39

H.7 Target and teacher updates 39

H.8 Consistency gap 43

H.9 Bridge and refresh sampling 46

H.10 Gumbel conditioning 54

I DiffusionGemma experiment **55**

J Proofs for the ERF sampler **55**

J.1 Intuition on the ODEs (8) 55

J.2 Setting 56

J.2.1 Reference rates 56

J.2.2 Residual dynamics 57

J.3 Theorem 57

J.4 Preliminaries 58

J.5 Change of variables 62

J.6 Induction on the number of jumps 66

J.7 Bounds on the partial sums and number of jumps 68

J.8 Proof of Theorem J.1 70

J.9 Last-jump distribution 70

A Related Work

CTMC sampling and discretization. Campbell et al. (2022) formulated discrete diffusion in continuous time, connecting their sampling to exact CTMC simulation (Gillespie, 1977) and τ-leaping (Gillespie, 2001). The modified next-reaction method (Anderson, 2007) preserves integrated-rate clocks across jumps but resamples the firing channel's exponential threshold. ERF instead becomes deterministic after a single random initialization, recycling residuals while preserving the exact one-time CTMC marginals, though generally with a different trajectory law. In practice, we freeze only the leave-one-out denoiser over each step while retaining the known time and state dependence of the rates, extending Euler/τ-leaping-style discretizations.

Deterministic sampling of discrete diffusion models. Zhang et al. (2023) introduce net-flux rates (14) to reduce jumps in a discrete analogue of the probability-flow ODE; see Appendix E, but the resulting process remains stochastic. Eyring et al. (2026) obtain deterministic trajectories by reusing Gumbel noise across discretization steps, potentially drifting from the target distribution. ERF instead gives deterministic transport after initialization while exactly preserving one-time marginals.

Distillation of discrete diffusion models. Distillation reduces sampling steps through distribution- or trajectory-based objectives. D-MMD (Hoogeboom et al., 2026) uses discrete moment matching, while IDLM (Li et al., 2026) uses a path space KL objective related to distribution matching (Yin et al., 2023). Both train auxiliary denoisers on student samples. Trajectory-based methods build on progressive and consistency distillation (Salimans & Ho, 2022; Song et al., 2023): SDTT (Deschenaux & Gulcehre, 2025) aggregates teacher probabilities across stochastic unmasking steps, while DCD (Sahoo et al., 2025) matches predictions along projected Gaussian trajectories with shared clean states and noise. These trajectories remain stochastic, leading to heuristic or biased objectives. Conditioning on exponential residuals instead makes the teacher trajectory deterministic given (X_t, E_t), enabling consistency distillation in the discrete setting.

B Rates and jump times used in the discretization

Fix $[s,t] = [s_i, t_i]$ and condition on $X_{t_i} = x$. The leave-one-out denoiser $\hat{x}_0^\theta(x, t_i)$ is cached throughout the step. Let z be the current state at time $r \in [s,t]$, and let $u \in [s,r]$ be a candidate jump time. An active residual $E_r^{\ell j}$ crosses zero when the integrated reverse rate from u to r equals $E_r^{\ell j}$; passive residuals increase by the integrated forward exit rate. More general freezing schemes and the derivations are given in Appendices D.5 and E.

For MDM, only $z^\ell = \mathsf{m}$ can jump, and then only to $j \neq \mathsf{m}$. The integrals needed for active and passive residuals are

$$\int_u^r \bar{Q}_v^{\theta,\ell}(j; t_i|z)\,\mathrm{d}v = \hat{x}_0^\theta(x,t_i)^{\ell j} \log\left(\frac{1-\alpha_r}{1-\alpha_u}\right) \qquad (z^\ell = \mathsf{m}),$$
$$\int_u^r \lambda_v^\ell(z^\ell)\,\mathrm{d}v = \mathbb{1}_{z^\ell \neq \mathsf{m}} \log\left(\frac{\alpha_u}{\alpha_r}\right).$$

If $\hat{x}_0^\theta(x,t_i)^{\ell j} > 0$ and $E_r^{\ell j} \leq \hat{x}_0^\theta(x,t_i)^{\ell j} \log((1-\alpha_r)/(1-\alpha_s))$, the candidate time for a masked position is

$$T^{\ell j} = \alpha^{-1}\left(1 - (1-\alpha_r)\exp\left\{-\frac{E_r^{\ell j}}{\hat{x}_0^\theta(x,t_i)^{\ell j}}\right\}\right),$$

and $T^{\ell j} = -\infty$ otherwise. In particular, no further active jump is possible at an unmasked position. Net flux does not change the MDM rates.

For UDM with net flux and $j \neq z^\ell$, the active and passive integrals are

$$\int_u^r \bar{Q}_v^{\mathrm{NF},\theta,\ell}(j; t_i|z)\,\mathrm{d}v = \frac{[\hat{x}_0^\theta(x,t_i)^{\ell j} - \hat{x}_0^\theta(x,t_i)^{\ell z^\ell}]_+}{K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - 1} \times \log\left(\frac{1 + (K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - 1)\alpha_u}{1 + (K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - 1)\alpha_r}\right),$$
$$\int_u^r b_v^{\mathrm{NF},\ell}(z)\,\mathrm{d}v = \frac{\sum_{j \in \mathsf{V}\setminus\{z^\ell\}} [\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - \hat{x}_0^\theta(x,t_i)^{\ell j}]_+}{K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - 1} \times \log\left(\frac{1 + (K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - 1)\alpha_u}{1 + (K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - 1)\alpha_r}\right).$$

Here $b_v^{\mathrm{NF},\ell}(z)$ is the forward net-flux exit rate at position ℓ. When $K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} = 1$, each logarithmic quotient divided by $K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - 1$ is understood as $\alpha_u - \alpha_r$. If $\hat{x}_0^\theta(x,t_i)^{\ell j} > \hat{x}_0^\theta(x,t_i)^{\ell z^\ell}$

and the active integral at $u = s$ is at least $E_r^{\ell j}$, its candidate time is

$$T^{\ell j} = \alpha^{-1}\left(\frac{\left(1 + (K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - 1)\alpha_r\right)\exp\left\{\frac{(K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell}-1)E_r^{\ell j}}{\hat{x}_0^\theta(x,t_i)^{\ell j}-\hat{x}_0^\theta(x,t_i)^{\ell z^\ell}}\right\} - 1}{K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} - 1}\right).$$

For $K\hat{x}_0^\theta(x,t_i)^{\ell z^\ell} = 1$, this becomes $T^{\ell j} = \alpha^{-1}(\alpha_r + E_r^{\ell j}/(\hat{x}_0^\theta(x,t_i)^{\ell j} - \hat{x}_0^\theta(x,t_i)^{\ell z^\ell}))$. When the stated conditions for a candidate jump fail, set $T^{\ell j} = -\infty$. The largest candidate time gives the next jump. After updating z, the residual integrals are recomputed with the same cached denoiser.

C Bridge and simplex extension

For $0 \le s < t \le 1$ with $\alpha_s > 0$, conditional independence across positions gives $q_{s|0,t}(x_s|x_0,x_t) = \prod_{\ell=1}^L q^\ell_{s|0,t}(x_s^\ell|x_0^\ell,x_t^\ell)$. For $x_0, x_t \in \mathsf{X}$ and a position ℓ such that $q^\ell_{t|0}(x_t^\ell|x_0^\ell) > 0$, Bayes' rule yields

$$q^\ell_{s|0,t}(\cdot|x_0^\ell,x_t^\ell) = \mathrm{Cat}\left(\cdot; \frac{\left[\alpha_{t|s}\tilde{x}_t^\ell + (1-\alpha_{t|s})\langle\tilde{x}_t^\ell,\pi^\ell\rangle\mathbf{1}_K\right]\odot\left[\alpha_s\tilde{x}_0^\ell + (1-\alpha_s)\pi^\ell\right]}{\alpha_t\langle\tilde{x}_t^\ell,\tilde{x}_0^\ell\rangle + (1-\alpha_t)\langle\tilde{x}_t^\ell,\pi^\ell\rangle}\right), \quad (16)$$

where $\tilde{x}_0^\ell = \tilde{x}_0^\ell$ and $\tilde{x}_t^\ell = \tilde{x}_t^\ell$. For $\nu \in \Delta_K$, we extend this expression by replacing x_0^ℓ with ν, equivalently setting $q^\ell_{s|0,t}(x_s^\ell|\nu,x_t^\ell) = q^\ell_{t|s}(x_t^\ell|x_s^\ell)q^\ell_{s|0}(x_s^\ell|\nu)/q^\ell_{t|0}(x_t^\ell|\nu)$, where $q^\ell_{r|0}(\cdot|\nu) = \mathrm{Cat}(\cdot;\alpha_r\nu + (1-\alpha_r)\pi^\ell)$. The numerator in (16) has nonnegative entries whose sum is the denominator, since $\alpha_{t|s}\alpha_s = \alpha_t$. Hence the extension is a probability vector whenever the denominator is positive. Completing zero-denominator rows by any probability vector and taking the positionwise product yields a stochastic reverse transition matrix on X; for masked diffusion, we use $\mathrm{Cat}(\cdot;\tilde{x}_t^\ell)$ at unmasked positions. Following Gourevitch et al. (2026, Proposition 1), define $\hat{x}^{\mathrm{loo}}_{0|t}(x_t)^\ell := \mathbb{E}[\tilde{X}_0^\ell \mid X_t^{-\ell} = x_t^{-\ell}]$. Conditioning first on $X_t^{-\ell}$ and then on X_t^ℓ gives, whenever $p_t(x_t) > 0$, $p^\ell_{s|t}(x_s^\ell|x_t) = q^\ell_{s|0,t}(x_s^\ell|\hat{x}^{\mathrm{loo}}_{0|t}(x_t)^\ell, x_t^\ell)$. Thus substituting the leave-one-out denoiser into the simplex extension recovers the exact positionwise reverse conditionals, and their product gives the factorized approximation used in (3).

D CTMC integration schemes

D.1 Gillespie's algorithm

D.1.1 Gillespie's first reaction form and ERF sampler

We recall Gillespie's algorithm in its first-reaction form (Gillespie, 2001), adapted to simulate $(X_u)_{u\in[s,t]}$ backward in time from $X_t = x$ using the reverse rate matrix $(\bar{Q}_u)_{u\in[s,t]}$. We define sequentially the jump times $(T_k^\star)_{k\le N}$ and states $(X_{T_k^\star})_{k\le N}$ up to a maximum number of jumps N in the following way. Set $T_0^\star = t$. At iteration $k \ge 0$, given the current jump time and state, for every $(\ell, j) \in [\![1,L]\!]\times[\![1,K]\!]$, independently draw $E_k^{\ell j} \sim \mathrm{Exp}(1)$ and define

$$T_{k+1}^{\ell j} = \sup\left\{u \in [s, T_k^\star] : \int_u^{T_k^\star} \bar{Q}_r^\ell(j|X_{T_k^\star})\,\mathrm{d}r \ge E_k^{\ell j}\right\}, \qquad (\ell_k^\star, j_k^\star) \in \operatorname*{argmax}_{(\ell,j)} T_{k+1}^{\ell j}. \quad (17)$$

We use here the convention $\sup\emptyset = -\infty$. If $T_{k+1}^{\ell_k^\star j_k^\star} = -\infty$, the procedure stops, so we set $N = k$ and $X_s = X_{T_k^\star}$. Otherwise, the $(k+1)$-st jump occurs at $T_{k+1}^\star = T_{k+1}^{\ell_k^\star j_k^\star}$, we set $X^{\ell_k^\star}_{T_{k+1}^\star} = j_k^\star$ and leave all other positions unchanged. The recursion continues until no candidate event occurs before s. Under appropriate conditions, (see the discussion in Feinberg et al. (2013)), N is almost surely finite and the algorithm produces a trajectory $(X_u)_{u\in[s,t]}$ whose law is that of the reverse CTMC conditioned on $X_t = x$.

The ERF sampler follows the same first-reaction construction on the augmented state. Draw $E_t^{\ell j} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$ once, including for the passive pairs $j = x^\ell$, and set $T_0^\star = t$. At iteration k,

use the current residuals instead of drawing fresh exponentials: for each $j \neq X^{\ell}_{T^\star_k}$, define $T^{\ell j}_{k+1}$ by (17) with $E^{\ell j}_k$ replaced by $E^{\ell j}_{T^\star_k}$, and select the largest candidate time. At the selected jump time $u = T^\star_{k+1}$, or at $u = s$ if no candidate exists, update every residual using the state before the jump:

$$E^{\ell j}_u = E^{\ell j}_{T^\star_k} + \begin{cases} -\displaystyle\int_u^{T^\star_k} \bar{Q}^{\ell}_v(j|X_{T^\star_k})\,\mathrm{d}v, & j \neq X^{\ell}_{T^\star_k}, \\ \displaystyle\int_u^{T^\star_k} \lambda^{\ell}_v(X^{\ell}_{T^\star_k})\,\mathrm{d}v, & j = X^{\ell}_{T^\star_k}. \end{cases} \tag{18}$$

If a jump occurs, change the winning position to $j^\star_k$ without resampling any residual; the triggering residual is zero and becomes passive. Otherwise, set $X_s = X_{T^\star_k}$ and stop. Thus the initial draw determines the entire ERF trajectory.

D.1.2 Gillespie's global form

The global form of Gillespie's algorithm (Gillespie, 1977) simulates the same backward trajectory but replaces the transition-specific exponential race by a single clock for the total exit rate. As above, set $T^\star_0 = t$ and $X_{T^\star_0} = x$, and define recursively the jump times $(T^\star_k)_{k \leq N}$ and the corresponding states. At iteration $k \geq 0$, given $(T^\star_k, X_{T^\star_k})$, define the total reverse exit rate $\bar{\lambda}^{\theta}_r(X_{T^\star_k}) := \sum_{\ell=1}^{L} \sum_{j: j \neq X^{\ell}_{T^\star_k}} \bar{Q}^{\theta,\ell}_r(j|X_{T^\star_k})$, draw a single variable $E_k \sim \mathrm{Exp}(1)$, and set

$$T^\star_{k+1} = \sup\left\{ u \in [s, T^\star_k] : \int_u^{T^\star_k} \bar{\lambda}^{\theta}_r(X_{T^\star_k})\,\mathrm{d}r \geq E_k \right\}. \tag{19}$$

We again use the convention $\sup \emptyset = -\infty$. If $T^\star_{k+1} = -\infty$, the procedure stops, so we set $N = k$ and $X_s = X_{T^\star_k}$. Otherwise, conditionally on the event time $T^\star_{k+1}$, we draw a transition $(\ell^\star_k, j^\star_k)$ with probability proportional to its instantaneous reverse rate, namely $\bar{Q}^{\theta,\ell^\star_k}_{T^\star_{k+1}}(j^\star_k|X_{T^\star_k}) / \bar{\lambda}^{\theta}_{T^\star_{k+1}}(X_{T^\star_k})$. We then set $X^{\ell^\star_k}_{T^\star_{k+1}} = j^\star_k$, leave all other positions unchanged, and repeat from $T^\star_{k+1}$.

The first-reaction form of D.1 draws one exponential variable for every possible transition and obtains the event time and transition label jointly by selecting the largest candidate time. The global form draws one exponential variable for the total exit rate and samples the transition label afterward from the normalized instantaneous rates. The resulting backward CTMC trajectories have the same law.

D.2 Gillespie discretization with partially frozen rates

Exact simulation with time-inhomogeneous Gillespie dynamics requires integrating the time-dependent transition rates and reevaluating their learned component after every one-position jump. To reduce the number of neural function evaluations, consider the grid $0 = t_0 < \cdots < t_n = 1$, with $s_i := t_{i-1}$, and the partially frozen reverse rate matrix $\bar{Q}^{\theta}_r(\cdot; t_i|\cdot)$ introduced in Section 3.2. On each interval $[s_i, t_i]$, we use a single forward pass of the model and may freeze the same component as for ERF: the full rate matrix, the score, the denoiser, or the leave-one-out denoiser. Gillespie discretization offers slightly more flexibility because it resamples the exponential clocks after every jump and therefore does not require the forward rates associated with the approximate frozen backward rates.

Applying the first-reaction construction to this approximate rate matrix gives Algorithm 2. Note that this discretization is almost the same as the discretization of the ERF sampler.

D.3 Equivalence of leave-one-out-frozen Gillespie and ancestral sampling

Fix an interval $[s, t] = [s_i, t_i]$, condition on $X_t = x$, assume that the denominators below are positive, and denote the cached leave-one-out predictions by $\widehat{p}^{\ell}_i := \hat{x}^{\theta}_0(x, t)^{\ell}$. Each prediction,

Algorithm 2 Gillespie discretization with partially frozen reverse rates

Require: Grid $0 = t_0 < \cdots < t_n = 1$ and a partial-freezing parameterization
Draw $X_1 \sim p_1$
for $i = n, \dots, 1$ **do** ▷ Discretize time
  Set $r \leftarrow t_i$, $s \leftarrow s_i$, and $z \leftarrow X_{t_i}$
  Evaluate the learned component at (t_i, z) and define $\bar{Q}^{\theta}_{\cdot}(\cdot; t_i|\cdot)$ ▷ Approximate the rates
  while $r > s$ **do**
    Draw $E^{\ell j} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$ for every (ℓ, j) such that $j \neq z^{\ell}$
    $T^{\ell j} \leftarrow \sup\left\{u \in [s, r] : \int_u^r \bar{Q}^{\theta,\ell}_v(j; t_i|z)\, \mathrm{d}v \geq E^{\ell j}\right\}$ ▷ Use the approximate rates
    $(\widehat{\ell}, \widehat{j}) \leftarrow \mathrm{argmax}_{(\ell,j)}\, T^{\ell j}$ and $u \leftarrow T^{\widehat{\ell}\widehat{j}}$
    if $u < s$ **then**
      $r \leftarrow s$
    else
      $z^{\widehat{\ell}} \leftarrow \widehat{j}$ and $r \leftarrow u$
    end if
  end while
  $X_{s_i} \leftarrow z$
end for
return X_0

together with the observation x^{ℓ}, induces the denoising posterior

$$p^{\theta,\ell}_{0|t}(j|x) := \frac{\langle \tilde{j}, \widehat{p}^{\ell}_i \rangle q^{\ell}_{t|0}(x^{\ell}|j)}{q^{\ell}_{t|0}(x^{\ell}|\widehat{p}^{\ell}_i)}, \qquad j \in \mathsf{V}.$$

At population optimum this is $\mathbb{P}(X^{\ell}_0 = j \mid X_t = x)$, by first conditioning on $X^{-\ell}_t = x^{-\ell}$ and then applying Bayes' rule to X^{ℓ}_t. Consequently, $\prod_{\ell=1}^{L} p^{\theta,\ell}_{0|t}(\cdot|x)$ is precisely the usual factorized approximation of the denoising posterior.

Consider the ancestral update obtained by marginalizing the forward bridge against this factorized posterior. Since both the bridge and the approximate posterior factorize across positions, Bayes' rule gives

$$\begin{aligned} p^{\theta}_{s|t}(w|x) &:= \sum_{x_0 \in \mathsf{X}} q_{s|0,t}(w|x_0, x) \prod_{\ell=1}^{L} p^{\theta,\ell}_{0|t}(x^{\ell}_0|x) \\ &= \prod_{\ell=1}^{L} \frac{q^{\ell}_{s|0}(w^{\ell}|\widehat{p}^{\ell}_i) q^{\ell}_{t|s}(x^{\ell}|w^{\ell})}{q^{\ell}_{t|0}(x^{\ell}|\widehat{p}^{\ell}_i)} = q_{s|0,t}(w|\widehat{p}_i, x). \end{aligned} \tag{20}$$

The last expression is the ancestral kernel obtained by applying (16) positionwise.

We now consider Gillespie sampling with the leave-one-out prediction $\widehat{p}^{\ell}_i$ frozen throughout $[s, t]$. For a fixed position ℓ, $q^{\ell}_{r|0}(z^{\ell}|\widehat{p}^{\ell}_i)$ is the time-r marginal of the one-token forward process initialized at time zero from $\mathrm{Cat}(\widehat{p}^{\ell}_i)$. If the current configuration is z and $y \neq z^{\ell}$, the leave-one-out-frozen reverse rate at this position is

$$\bar{Q}^{\theta,\ell,\mathrm{loo}}_r(y; t_i|z) = Q^{\ell}_r(z^{\ell}|y) \frac{q^{\ell}_{r|0}(y|\widehat{p}^{\ell}_i)}{q^{\ell}_{r|0}(z^{\ell}|\widehat{p}^{\ell}_i)}.$$

This is exactly the usual time-reversal formula for that one-token forward process. Therefore, when the reverse process starts from x^{ℓ} at time t, its transition probability to w^{ℓ} at time s is

$$\frac{q^{\ell}_{s|0}(w^{\ell}|\widehat{p}^{\ell}_i) q^{\ell}_{t|s}(x^{\ell}|w^{\ell})}{q^{\ell}_{t|0}(x^{\ell}|\widehat{p}^{\ell}_i)} = q^{\ell}_{s|0,t}(w^{\ell}|\widehat{p}^{\ell}_i, x^{\ell}).$$

Conditionally on the cached predictions, the frozen generator is a sum of these one-position generators, so its transition kernel is the product of the positionwise transitions. Exact Gillespie simulation,

including all within-interval jumps, consequently produces

$$\prod_{\ell=1}^{L} q^{\ell}_{s|0,t}(w^{\ell}|\widehat{p}^{\ell}_{i}, x^{\ell}) = p^{\theta}_{s|t}(w|x).$$

Thus leave-one-out-frozen Gillespie recovers exactly the ancestral update in (20).

The same conclusion does not hold in general when freezing the score or the ordinary denoising posterior.

D.4 Euler and categorical tau-leaping

Token-wise Euler. Token-wise Euler freezes the state and time dependence of the rate matrix, effectively drawing the positions conditionally independently. For each ℓ, its categorical probabilities are

$$\mathbb{P}(X^{\ell}_{s} = y \mid X_t = x) = \begin{cases} h\bar{Q}^{\theta,\ell}_{t}(y|x) & y \neq x^{\ell}\,, \\ 1 + h\bar{Q}^{\theta,\ell}_{t}(x^{\ell}|x) & y = x^{\ell}\,. \end{cases} \tag{21}$$

The step is valid whenever $h\sum_{y\neq x^{\ell}} \bar{Q}^{\theta,\ell}_{t}(y|x) \leq 1$ for every position ℓ.

Categorical τ-leaping. Categorical τ-leaping uses the same frozen rates but solves Kolmogorov exactly for these frozen rate until the first jump. Draw independently

$$N^{\ell y}_{i} \sim \mathrm{Poisson}\Big(h\bar{Q}^{\theta,\ell}_{t}(y|x)\Big)\,, \qquad \ell \in [\![1, L]\!]\,, \quad y \neq x^{\ell}\,. \tag{22}$$

If $\sum_{y\neq x^{\ell}} N^{\ell y}_{i} = 1$, position ℓ is set to the unique y for which $N^{\ell y}_{i} = 1$; otherwise it remains equal to x^{ℓ}. Equivalently, with $\bar{\lambda}^{\theta,\ell}_{t}(x) := \sum_{y\neq x^{\ell}} \bar{Q}^{\theta,\ell}_{t}(y|x)$, the update probabilities are $\mathbb{P}(X^{\ell}_{s} = y \mid X_t = x) = h\bar{Q}^{\theta,\ell}_{t}(y|x)\exp\{-h\bar{\lambda}^{\theta,\ell}_{t}(x)\}$ for $y \neq x^{\ell}$, and the remaining mass is assigned to x^{ℓ}.

D.5 Explicit partially frozen updates for MDM and UDM

Here we describe the rate matrices associated with the different methods introduced in Section 3.2. Fix the discretization interval $[s, t] = [s_i, t_i]$ and condition on $X_{t_i} = x$. The discrete state x is fixed throughout the interval: the learned component is evaluated once at (t_i, X_{t_i}) and cached. By contrast, z denotes the current discrete state of the trajectory at a time $r \in [s, t]$. Initially $r = t_i$ and $z = x$, but z is updated after every jump while the cached output evaluated at (t_i, X_{t_i}) is kept unchanged. In the formulas below, $y \neq z^{\ell}$ is a candidate token at position ℓ, and we suppose that the current discrete state remains equal to z as time decreases from r to u.

Besides the score representation in (10), the density ratio in (6) admits the identities

$$\frac{p^{\ell}_{t}(w^{\ell}|x^{-\ell})}{p^{\ell}_{t}(x^{\ell}|x^{-\ell})} = \sum_{x_0\in\mathsf{V}} \frac{q^{\ell}_{t|0}(w^{\ell}|x_0)}{q^{\ell}_{t|0}(x^{\ell}|x_0)} p^{\ell}_{0|t}(x_0|x)\,, \tag{23}$$

$$\frac{p^{\ell}_{t}(w^{\ell}|x^{-\ell})}{p^{\ell}_{t}(x^{\ell}|x^{-\ell})} = \frac{q^{\ell}_{t|0}(w^{\ell}|\hat{x}^{\mathrm{loo}}_{0|t}(x)^{\ell})}{q^{\ell}_{t|0}(x^{\ell}|\hat{x}^{\mathrm{loo}}_{0|t}(x)^{\ell})}\,. \tag{24}$$

MDM. The only possible reverse transition at position ℓ is from $z^{\ell} = \mathsf{m}$ to a clean token $y \neq \mathsf{m}$. Under score freezing, the rate and its integral are

$$\bar{Q}^{\theta,\ell}_{v}(y; t_i|z) = \langle \tilde{y}, s^{\theta}_{t_i}(X_{t_i})^{\ell}\rangle \left(-\frac{\dot{\alpha}_v}{\alpha_v}\right), \quad \int_{u}^{r} \bar{Q}^{\theta,\ell}_{v}(y; t_i|z)\,\mathrm{d}v = \langle \tilde{y}, s^{\theta}_{t_i}(X_{t_i})^{\ell}\rangle \log\left(\frac{\alpha_u}{\alpha_r}\right).$$

Under denoising-posterior freezing, they are

$$\bar{Q}^{\theta,\ell}_{v}(y; t_i|z) = p^{\theta,\ell}_{0|t_i}(y|X_{t_i}) \left(-\frac{\dot{\alpha}_v}{1-\alpha_v}\right), \quad \int_{u}^{r} \bar{Q}^{\theta,\ell}_{v}(y; t_i|z)\,\mathrm{d}v = p^{\theta,\ell}_{0|t_i}(y|X_{t_i}) \log\left(\frac{1-\alpha_r}{1-\alpha_u}\right).$$

For leave-one-out-denoiser freezing, the rate and its integral have the same form as under denoising-posterior freezing, with $p^{\theta,\ell}_{0|t_i}(y|X_{t_i})$ replaced by $\hat{x}^{\theta}_0(X_{t_i},t_i)^{\ell y}$; the resulting practical updates are collected in Appendix B. All rates at an already unmasked position vanish.

UDM. Under score freezing, every $y \neq z^\ell$ has

$$\bar{Q}^{\theta,\ell}_v(y;t_i|z) = \frac{\langle \tilde{y}, s^{\theta}_{t_i}(X_{t_i})^\ell\rangle}{K}\left(-\frac{\dot{\alpha}_v}{\alpha_v}\right), \quad \int_u^r \bar{Q}^{\theta,\ell}_v(y;t_i|z)\,\mathrm{d}v = \frac{\langle \tilde{y}, s^{\theta}_{t_i}(X_{t_i})^\ell\rangle}{K}\log\left(\frac{\alpha_u}{\alpha_r}\right).$$

For denoising-posterior freezing, the rate is

$$\bar{Q}^{\theta,\ell}_v(y;t_i|z) = \frac{1}{K}\left(-\frac{\dot{\alpha}_v}{\alpha_v}\right)\Big(1 - p^{\theta,\ell}_{0|t_i}(y|X_{t_i}) - p^{\theta,\ell}_{0|t_i}(z^\ell|X_{t_i}) \\ + p^{\theta,\ell}_{0|t_i}(y|X_{t_i})\frac{1+(K-1)\alpha_v}{1-\alpha_v} + p^{\theta,\ell}_{0|t_i}(z^\ell|X_{t_i})\frac{1-\alpha_v}{1+(K-1)\alpha_v}\Big),$$

whose integral is

$$\int_u^r \bar{Q}^{\theta,\ell}_v(y;t_i|z)\,\mathrm{d}v = \frac{1}{K}\log\left(\frac{\alpha_u}{\alpha_r}\right) + p^{\theta,\ell}_{0|t_i}(y|X_{t_i})\log\left(\frac{1-\alpha_r}{1-\alpha_u}\right) \\ + \frac{p^{\theta,\ell}_{0|t_i}(z^\ell|X_{t_i})}{K-1}\log\left(\frac{1+(K-1)\alpha_r}{1+(K-1)\alpha_u}\right).$$

Finally, under leave-one-out-denoiser freezing,

$$\bar{Q}^{\theta,\ell}_v(y;t_i|z) = \frac{1}{K}\left(-\frac{\dot{\alpha}_v}{\alpha_v}\right)\frac{\alpha_v\langle \tilde{y}, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle + (1-\alpha_v)/K}{\alpha_v\langle \tilde{z}^\ell, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle + (1-\alpha_v)/K},$$

$$\int_u^r \bar{Q}^{\theta,\ell}_v(y;t_i|z)\,\mathrm{d}v = \frac{1}{K}\log\left(\frac{\alpha_u}{\alpha_r}\right) \\ + \frac{\langle \tilde{y}, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle - \langle \tilde{z}^\ell, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle}{K\langle \tilde{z}^\ell, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle - 1}\log\left(\frac{1+\big(K\langle \tilde{z}^\ell, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle - 1\big)\alpha_u}{1+\big(K\langle \tilde{z}^\ell, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle - 1\big)\alpha_r}\right).$$

When $K\langle \tilde{z}^\ell, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle - 1 = 0$, the second term is understood by continuity as

$$\big(\langle \tilde{y}, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle - \langle \tilde{z}^\ell, \hat{x}^{\theta}_0(X_{t_i},t_i)^\ell\rangle\big)\,(\alpha_u - \alpha_r).$$

After every UDM transition, z^ℓ is replaced by the selected token and the displayed expressions are reevaluated for the new current token using the cached posterior or leave-one-out prediction before solving for the following event.

D.5.1 From integrated rates to jump updates

Suppose that the evolving configuration is z at time r. In the first-reaction form of Gillespie's algorithm, we draw independent $E^{\ell y}_r \sim \mathrm{Exp}(1)$ for every possible transition (ℓ, y) with $y \neq z^\ell$. Its candidate time is $T^{\ell y} = -\infty$ if $\int_s^r \bar{Q}^{\theta,\ell}_v(y;t_i|z)\,\mathrm{d}v < E^{\ell y}_r$; otherwise, $T^{\ell y}$ is the largest $u \in [s,r]$ satisfying

$$\int_u^r \bar{Q}^{\theta,\ell}_v(y;t_i|z)\,\mathrm{d}v = E^{\ell y}_r\,. \tag{25}$$

The first transition in the backward evolution is $(\widehat{\ell},\widehat{y}) \in \mathrm{argmax}_{\ell,y} T^{\ell y}$. If all candidate times equal $-\infty$, no transition occurs before s. Otherwise, we set $u = T^{\widehat{\ell}\widehat{y}}$, $z^{\widehat{\ell}} \leftarrow \widehat{y}$, and $r \leftarrow u$, then start a new exponential race while keeping the learned component fixed.

Closed-form update for MDM. In MDM, only transitions from $z^\ell = \mathsf{m}$ to $y \neq \mathsf{m}$ are possible. Moreover, all active rates share the same time-dependent factor. It is therefore unnecessary to compute and compare every candidate time: one first selects the transition with the smallest normalized residual and then evaluates α^{-1} only once for this transition.

Throughout this paragraph, a quotient with zero denominator is understood as $+\infty$.

Under score freezing, select

$$(\widehat{\ell}, \widehat{y}) \in \operatorname*{argmin}_{\substack{\ell: z^\ell = \mathsf{m} \\ y \neq \mathsf{m}}} \frac{E_r^{\ell y}}{\langle \tilde{y}, s_{t_i}^\theta(X_{t_i})^\ell \rangle}.$$

If the minimum exceeds $\log(\alpha_s/\alpha_r)$, no transition occurs before s. Otherwise, the selected transition occurs at the single time

$$T^\star = \alpha^{-1}\left(\alpha_r \exp\left\{ \frac{E_r^{\widehat{\ell}\widehat{y}}}{\langle \tilde{\widehat{y}}, s_{t_i}^\theta(X_{t_i})^{\widehat{\ell}} \rangle} \right\}\right).$$

For denoising-posterior freezing, the selection and inversion have the form in Appendix B, with its cached denoiser probability replaced by $p_{0|t_i}^{\theta,\ell}(y|X_{t_i})$. In MDM, the leave-one-out and standard denoisers coincide at masked positions.

Root finding for UDM. Score-frozen UDM candidate times are also explicit. The candidate time for (ℓ, y) lies in $[s, r]$ when

$$\frac{\langle \tilde{y}, s_{t_i}^\theta(X_{t_i})^\ell \rangle}{K} \log\left(\frac{\alpha_s}{\alpha_r}\right) \geq E_r^{\ell y},$$

and, when its coefficient is positive,

$$T^{\ell y} = \alpha^{-1}\left(\alpha_r \exp\left\{ \frac{K E_r^{\ell y}}{\langle \tilde{y}, s_{t_i}^\theta(X_{t_i})^\ell \rangle} \right\}\right).$$

For denoising-posterior and leave-one-out-denoiser freezing, the displayed integrals combine several logarithmic terms, so (25) has no closed-form solution in general. After checking that $\int_s^r \bar{Q}_v^{\theta,\ell}(y; t_i|z)\,\mathrm{d}v \geq E_r^{\ell y}$, we obtain $T^{\ell y}$ by bisection, since $u \mapsto \int_u^r \bar{Q}_v^{\theta,\ell}(y; t_i|z)\,\mathrm{d}v$ is continuous and monotone. Once all candidate times have been obtained, we apply the transition with the largest one, update z, and recompute the integrals with the new current token and the same cached output evaluated at (t_i, X_{t_i}).

D.5.2 Exactness of leave-one-out freezing under factorization

Suppose that the data distribution factorizes across positions, and write $p_0^\ell := \mathbb{E}[\tilde{X}_0^\ell] \in \Delta_K$ for its ℓ-th marginal. Since the forward process also acts independently across positions, its marginal remains factorized at every time r, with ℓ-th marginal $q_{r|0}^\ell(\cdot|p_0^\ell)$. Independence then gives

$$\hat{x}_{0|r}^{\mathrm{loo}}(z)^\ell = \mathbb{E}[\tilde{X}_0^\ell \mid X_r^{-\ell} = z^{-\ell}] = p_0^\ell$$

for every r and z. Thus, at optimum, the cached leave-one-out prediction is independent of both the endpoint (t, x) and the evolving discrete state. If z and w differ only at position ℓ, the leave-one-out-frozen rate consequently satisfies

$$\bar{Q}_r^{\theta,\mathrm{loo}}(w; t_i|z) = Q_r^\ell(z^\ell|w^\ell)\frac{q_{r|0}^\ell(w^\ell|p_0^\ell)}{q_{r|0}^\ell(z^\ell|p_0^\ell)} = \bar{Q}_r(w|z). \tag{26}$$

Hence the partially frozen generator is the exact reverse generator throughout every discretization interval, and its exact integration introduces no discretization error. The single-token case is the special case $L = 1$: the conditioning context is empty, so the leave-one-out prediction is simply the data distribution p_0.

Score and denoising-posterior freezing do not have this property in general. Even under factorization, the exact score at position ℓ depends on the current time through the ratio of the marginals at time r and on the current token z^ℓ. Similarly, the ordinary denoising posterior reduces to $\mathbb{P}(X_0^\ell = y \mid X_r^\ell = z^\ell)$, which still depends on both r and z^ℓ. Caching either quantity at (t, x) therefore generally changes the reverse rates as soon as time advances or the position jumps. For example, in a single-token binary UDM with uniform reference distribution and $p_0(1) = \rho \neq 1/2$, one has $p_r(1) = \alpha_r\rho + (1-\alpha_r)/2$. Both the score ratio and the denoising posterior vary with α_r, so freezing them at t cannot recover the exact reverse generator over $[s, t]$. In MDM, the ordinary and

leave-one-out denoisers coincide at a masked position, so denoising-posterior freezing is also exact in this particular factorized setting but the distinction concerns general corruption kernels such as UDM.

D.5.3 A COUNTEREXAMPLE FOR SCORE FREEZING

We now show that freezing the score for the ERF procedure while keeping the original forward rates can lead to an infinite number of jumps in finite time. Consider UDM with one position, two tokens, and the schedule $\alpha_t = \exp(-2t)$. The original forward rates are constant and satisfy

$$Q_t(2|1) = Q_t(1|2) = 1.$$

Freeze the score at time $t_i = 1/10$ and state $X_{t_i} = 1$, with cached value $s_{t_i}(1) = (1, 2)$. The score-frozen rates in UDM are then

$$\bar{Q}_t^\theta(2; t_i|1) = 2, \qquad \bar{Q}_t^\theta(1; t_i|2) = 1.$$

The ERF nevertheless retains the original forward exit rate, equal to one, as the passive speed in either token. These active and passive speeds are incompatible: if the frozen reverse rates were the time reversal of the original forward rates under a positive law q_t, the two directions would require simultaneously $q_t(2)/q_t(1) = 2$ and $q_t(1)/q_t(2) = 1$.

Let $(E_{t_i}^{11}, E_{t_i}^{12}) = (A, B)$ with A and B independent $\mathrm{Exp}(1)$ variables. Starting from $X_{t_i} = 1$ and writing $S := A + B/2$, the first jump occurs after backward elapsed time $B/2$ and leaves the residual pair $(S, 0)$ in token 2. After another time S, the process returns to 1 with residuals $(0, S)$. More generally, starting from 1 with residuals $(0, R)$, a two-jump cycle takes time R and returns to 1 with residuals $(0, R/2)$. The successive cycles therefore have durations $S, S/2, S/4, \ldots$, so infinitely many jumps accumulate at

$$T_{\mathrm{acc}} = \frac{B}{2} + S + \sum_{n=0}^{\infty} \frac{S}{2^n} = 3A + 2B < \infty. \tag{27}$$

For every $h > 0$, the event $3A + 2B < h$ has probability $1 - 3\exp(-h/3) + 2\exp(-h/2) > 0$. Hence the explosion occurs with positive probability on any interval of positive length, whereas Gillespie sampling with the same bounded frozen reverse rates is nonexplosive.

The issue is not with the score approximation, the cached score can be population-optimal. Indeed, choose

$$p_0(1) = \frac{1}{2} - \frac{\exp(1/5)}{6}, \qquad p_0(2) = \frac{1}{2} + \frac{\exp(1/5)}{6}.$$

Then $p_{t_i}(1) = 1/3$ and $p_{t_i}(2) = 2/3$, so the exact score at 1 is $(1, 2)$. Conditional on $X_{t_i} = 1$, the score-frozen ERF explodes before time zero on the event $3A + 2B < t_i$, which has positive probability.

D.5.4 WHY THE EXPLOSION CANNOT OCCUR WITH LEAVE-ONE-OUT FREEZING

We now show that the same phenomenon cannot happen with leave-one-out freezing. Fix the cached leave-one-out predictions $\widehat{p}_i^\ell := \hat{x}_0^\theta(X_{t_i}, t_i)^\ell$, for $\ell \in [\![1, L]\!]$, and define the product law

$$\widetilde{p}_{r|i}(z) := \prod_{\ell=1}^{L} q_{r|0}^\ell(z^\ell|\widehat{p}_i^\ell).$$

Each factor is a marginal of the original forward process, hence $\widetilde{p}_{r|i}$ is transported by the original forward generator. If z and w differ only at position ℓ, then (12) becomes

$$\bar{Q}_r^{\theta,\mathrm{loo}}(w; t_i|z) = \frac{\widetilde{p}_{r|i}(w)}{\widetilde{p}_{r|i}(z)} Q_r(z|w).$$

Thus the leave-one-out-frozen reverse generator is the exact time reversal of the original forward generator under $\widetilde{p}_{r|i}$. Its associated forward exit rates are exactly $\lambda_r^\ell(z^\ell)$, which are the passive speeds retained by ERF. The incompatibility responsible for the score-freezing counterexample is therefore absent.

The product law $\widetilde{p}_{t|i}$ is only an auxiliary reference law: the actual marginal of X_t entering the interval, conditional on the cached prediction, is in general not the same. Nevertheless, this is sufficient to rule out explosion. Let $\mathcal{B}_i$ be the set of augmented initial states from which the leave-one-out-frozen residual flow explodes on a regular interval $[s, t]$. By Theorem 3.1,

$$\left(\widetilde{p}_{t|i} \otimes \mathrm{Exp}(1)^{\otimes LK}\right)(\mathcal{B}_i) = 0.$$

When $\widetilde{p}_{t|i}$ is strictly positive, as on a regular UDM interval, this measure is equivalent to counting measure on X times Lebesgue measure on the residual orthant. Consequently, the exploding residual vectors form a Lebesgue-null set separately for every starting configuration. The actual token marginal may therefore differ from $\widetilde{p}_{t|i}$ without affecting nonexplosion: any token entry law combined with an absolutely continuous conditional residual law assigns zero probability to $\mathcal{B}_i$. In particular, the open positive-volume set of exploding residuals exhibited above for score freezing cannot exist under leave-one-out freezing.

E Net-flux rate matrix

E.1 Net flux construction and relation to prior work

The relevant decomposition is that of the probability flux induced by the reverse rates, rather than of the rates alone. Assume that $p_t(x) > 0$ and, for $x \neq w$, define

$$F_t(x, w) \coloneqq p_t(x)\bar{Q}_t(w|x), \quad S_t(x, w) \coloneqq \frac{F_t(x, w) + F_t(w, x)}{2}, \quad J_t(x, w) \coloneqq F_t(x, w) - F_t(w, x).$$

Here S_t is the symmetric traffic along the unordered pair $\{x, w\}$ and J_t is its antisymmetric current. Using $\bar{Q}_t(x|x) = -\sum_{w \neq x} \bar{Q}_t(w|x)$, the reverse Kolmogorov equation becomes

$$\partial_t p_t(x) = -\sum_w p_t(w)\bar{Q}_t(x|w) = \sum_{w \neq x} \big(F_t(x, w) - F_t(w, x)\big) = \sum_{w \neq x} J_t(x, w).$$

Thus the marginal evolution depends only on J_t: the symmetric term S_t contributes equally in both directions and cancels in every difference $F_t(x, w) - F_t(w, x)$. Keeping the same current while forbidding simultaneous flux in both directions uniquely gives $F_t^{\mathrm{NF}}(x, w) = [J_t(x, w)]_+$. Dividing this flux by the mass at its source yields

$$\bar{Q}_t^{\mathrm{NF}}(w|x) \coloneqq \frac{[J_t(x, w)]_+}{p_t(x)} = \frac{\big[p_t(x)\bar{Q}_t(w|x) - p_t(w)\bar{Q}_t(x|w)\big]_+}{p_t(x)},$$

with diagonal entries chosen so that each row sums to zero. Since $a_+ - (-a)_+ = a$, the net-flux rates preserve J_t and hence the Kolmogorov equation, while $\bar{Q}_t^{\mathrm{NF}}(w|x)\bar{Q}_t^{\mathrm{NF}}(x|w) = 0$. Moreover, any pair of nonnegative opposing fluxes with difference $J_t(x, w)$ has total flux at least $|J_t(x, w)|$, with equality for the net-flux construction. It is therefore the minimum-activity generator with the same pairwise current.

The time-reversal identity gives $p_t(w)\bar{Q}_t(x|w) = p_t(x)Q_t(w|x)$, and consequently $\bar{Q}_t^{\mathrm{NF}}(w|x) = [\bar{Q}_t(w|x) - Q_t(w|x)]_+$.

A similar construction appears in Campbell et al. (2024), however the two constructions are not equivalent and ours leads to an even less stochastic process. Let us write $\bar{Q}_t^{X_0}(\cdot|\cdot)$ for the reverse generator conditioned on the clean endpoint. Campbell et al. (2024) applies net flux conditionally and then averages under the posterior of X_0, whereas ours first forms the unconditional generator and then applies net flux. Thus, for $x \neq w$, the two rates are respectively

$$\mathbb{E}\Big[\bar{Q}_t^{\mathrm{NF}, X_0}(w|x) \mid X_t = x\Big] \qquad \text{and} \qquad \bar{Q}_t^{\mathrm{NF}}(w|x),$$

and they are not equal in general because the positive-part operation in the net-flux projection does not commute with averaging. More precisely, its convexity gives

$$\bar{Q}_t^{\mathrm{NF}}(w|x) \leq \mathbb{E}\Big[\bar{Q}_t^{\mathrm{NF}, X_0}(w|x) \mid X_t = x\Big].$$

The inequality is strict when different clean endpoints induce opposing currents between x and w. Hence our unconditional net-flux generator has no larger off-diagonal rate for any transition, and therefore no larger total jump intensity, while inducing the same marginal evolution. Interestingly, applying net flux once more after averaging the conditional construction recovers our generator.

E.2 Forward net-flux rates and passive residuals

ERF requires the forward rates associated with the chosen reverse generator to evolve its passive residuals. For strictly positive marginals and exact canonical rates, time reversal gives, for $x \neq w$,

$$Q_t^{\mathrm{NF}}(w|x) := \frac{p_t(w)}{p_t(x)} \bar{Q}_t^{\mathrm{NF}}(x|w) = \big[Q_t(w|x) - \bar{Q}_t(w|x)\big]_+ . \tag{28}$$

Indeed, substituting (14) gives $[p_t(w)\bar{Q}_t(x|w) - p_t(x)\bar{Q}_t(w|x)]_+ / p_t(x)$, and the canonical time-reversal identity reduces its first term to $Q_t(w|x)$. Diagonal entries are the negatives of the corresponding row sums. Thus both net-flux generators are obtained by subtracting the canonical forward and reverse rates in opposite orders and taking positive parts; no additional model or density evaluation is needed.

For position ℓ, the passive speed must therefore be replaced by $b_t^{\mathrm{NF},\ell}(x) := \sum_{w:\, w^{-\ell}=x^{-\ell},\, w^\ell \neq x^\ell} Q_t^{\mathrm{NF}}(w|x)$. This choice satisfies the required boundary balance $p_t(x) b_t^{\mathrm{NF},\ell}(x) = \sum_{w:\, w^{-\ell}=x^{-\ell},\, w^\ell \neq x^\ell} p_t(w) \bar{Q}_t^{\mathrm{NF}}(x|w)$. Together with preservation of the Kolmogorov equation, this gives the augmented marginal law under the regularity assumptions of the ERF construction. Moreover, the net-flux reverse rates are pointwise no larger than the canonical rates, so their expected jump count is no larger under the common exact marginals. Their leave-one-out-frozen integrals and jump times are given in Appendix B.

F Additional material for distillation

F.1 Consistency distillation training

The objective and its motivation are given in Section 4; the round-based training procedure is summarized below. Write D for the selected forward or reverse KL divergence. In Algorithm 3, only the frozen target is refreshed between rounds and the pretrained teacher generates every trajectory. The pretrained teacher is the original diffusion model and is evaluated from the current tokens and time.

Algorithm 3 Exponential-residual consistency distillation with roundwise target refresh

Require: Pre-trained teacher parameters θ, student parameters ϕ, gap δ, numbers of rounds N and updates M, learning rate η, averaging decay ρ, and KL choice D
Ensure: Distilled student parameters ϕ

$\phi^{\mathrm{EMA}} \leftarrow \phi$
for $n = 1, \dots, N$ **do**
  $\bar{\phi} \leftarrow \phi^{\mathrm{EMA}}$ ▷ Freeze the target for this round
  for $m = 1, \dots, M$ **do**
    Sample $t \sim \mathrm{Unif}(\delta, 1)$ and set $s \leftarrow t - \delta$
    Sample $X_0 \sim p_0$ and $X_t \sim q_{t|0}(\cdot|X_0)$
    Sample $E_t^{\ell j} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$ independently of X_t
    $(X_s, E_s) \leftarrow \Psi_{s|t}^{\theta}(X_t, E_t)$
    $\widehat{\mathcal{L}} \leftarrow D\Big(p_{0|t}^{\phi}(\cdot|X_t, E_t), \mathrm{sg}\Big[p_{0|s}^{\bar{\phi}}(\cdot|X_s, E_s)\Big]\Big)$
    $\phi \leftarrow \phi - \eta \nabla_\phi \widehat{\mathcal{L}}$
    $\phi^{\mathrm{EMA}} \leftarrow \rho \phi^{\mathrm{EMA}} + (1-\rho)\phi$
  end for
end for
return ϕ

The joint-refresh variant also replaces the model that supplies the ERF rollout predictions between rounds. The first round uses the pretrained diffusion model. In every later round, the rollout uses the same frozen EMA copy $\bar{\phi}$ as the consistency target. To advance from t to s, we split the interval into numerical ERF substeps. At the beginning of a substep, say at time u, the frozen ERC network is evaluated with token input X_u, time input u, and residual input $-\log E_u$. For MDM, its output is the mean denoising distribution used in the masked ERF rates. For UDM, its output is the native leave-one-out distribution used directly in the UDM ERF rates. This output is frozen while one ERF substep advances the residuals and changes the tokens. The resulting (X, E) becomes the network input at the next substep. Thus, the model is reevaluated at every numerical substep, but not after each token jump occurring within that substep. The complete procedure is summarized in Algorithm 4.

Algorithm 4 Exponential-residual consistency distillation with roundwise target-and-teacher refresh

Require: Pre-trained teacher parameters θ, student parameters ϕ, gap δ, numbers of rounds N and updates M, learning rate η, averaging decay ρ, and KL choice D
Ensure: Distilled student parameters ϕ

$\phi^{\mathrm{EMA}} \leftarrow \phi$
for $n = 1, \dots, N$ **do**
  $\bar{\phi} \leftarrow \phi^{\mathrm{EMA}}$ ▷ Freeze the target for this round
  for $m = 1, \dots, M$ **do**
    Sample $t \sim \mathrm{Unif}(\delta, 1)$ and set $s \leftarrow t - \delta$
    Sample $X_0 \sim p_0$ and $X_t \sim q_{t|0}(\cdot|X_0)$
    Sample $E_t^{\ell j} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$ independently of X_t
    Choose the teacher grid $t = u_0 > u_1 > \cdots > u_J = s$
    for $q = 0, \dots, J-1$ **do**
      if $n = 1$ **then**
        Evaluate the pretrained teacher with network inputs (X_{u_q}, u_q)
      else
        Evaluate the frozen ERC model $\bar{\phi}$ with network inputs $(X_{u_q}, u_q, -\log E_{u_q})$
      end if
      Advance (X_{u_q}, E_{u_q}) to $(X_{u_{q+1}}, E_{u_{q+1}})$ by one ERF substep with this output frozen
    end for
    $\widehat{\mathcal{L}} \leftarrow D\Big(p^{\phi}_{0|t}(\cdot|X_t, E_t), \mathrm{sg}\Big[p^{\bar{\phi}}_{0|s}(\cdot|X_s, E_s)\Big]\Big)$
    $\phi \leftarrow \phi - \eta \nabla_{\phi} \widehat{\mathcal{L}}$
    $\phi^{\mathrm{EMA}} \leftarrow \rho \phi^{\mathrm{EMA}} + (1-\rho)\phi$
  end for
end for
return ϕ

For UDM, the ERC network natively predicts a leave-one-out distribution. The two uses of this prediction must be distinguished. During a rollout, the pretrained teacher supplies its usual leave-one-out prediction, while a refreshed trajectory teacher supplies the residual-conditioned leave-one-out prediction of the frozen ERC model; in both cases, this leave-one-out distribution is used directly to construct the UDM ERF rates. In the consistency loss, both the current student prediction at (X_t, E_t) and the frozen target prediction at (X_s, E_s) are instead converted to mean denoising distributions before applying D. The conversion is pointwise in any additional residual conditioning, so both sides retain the same dependence on E_u. It is given by

$$\hat{x}_0^{\omega}(x,u)^{\ell} = \mathrm{softmax}\left(\log \hat{x}_0^{\omega,\mathrm{loo}}(x,u)^{\ell} + \log\left(1 + \frac{K\alpha_u}{1-\alpha_u}\right)\tilde{x}^{\ell}\right). \tag{29}$$

F.2 MULTI-STEP SAMPLING

Algorithm 5 Multi-step refresh or bridge sampling with a distilled student

Require: Student parameters ϕ, decreasing grid $1 = t_k > t_{k-1} > \cdots > t_1 > t_0 = 0$ with $s_i := t_{i-1}$, and choice of refresh or bridge sampling
Output: A clean state $\widehat{X}_0$

Sample $X_{t_k} \sim p_1$ and $E^{\ell j}_{t_k} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$
for $i = k, k-1, \ldots, 1$ **do**
  Sample $\widehat{X}_0 \sim p^{\phi}_{0|t_i}(\cdot|X_{t_i}, E_{t_i})$
  if $i > 1$ **then**
    either sample $X_{s_i} \sim q_{s_i|0}(\cdot|\widehat{X}_0)$ ▷ Refresh
    or sample $X_{s_i} \sim q_{s_i|0,t_i}(\cdot|\widehat{X}_0, X_{t_i})$ ▷ Bridge
    Sample a fresh $E^{\ell j}_{s_i} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$
  end if
end for
return $\widehat{X}_0$

F.3 BRIDGE SAMPLING AND THE ERF COUPLING

The bridge $q_{s|0,t}(\cdot|y,x)$ is defined by the original forward diffusion; see (16). Therefore, to obtain samples from the marginal p_s, it suffices to sample $X_0 \sim p_{0|t}(\cdot|x)$ where $p_{0|t}(\cdot|x)$ is the original posterior, i.e., $\frac{p_0(y)q_{t|0}(x|y)}{p_t(x)}$. However, the ERF process generally changes the coupling of X_t and X_0. The coupling learned by the student is

$$p^{\phi}_{0|t}(y|x_t) := \mathbb{E}_{E_t}\left[p^{\phi}_{0|t}(y|x_t, E_t)\right]. \tag{30}$$

Consequently, the ideal endpoint distribution $p^{\phi}_{0|t}(\cdot|x_t, E_t)$ need not give a sample from the forward posterior $p_{0|t}(\cdot|x)$ and the bridge generally does not produce p_s. Refresh sampling is always valid since it requires only $\widehat{X}_0 \sim p_0$: then $\sum_y p_0(y)q_{s|0}(z|y) = p_s(z)$, and fresh independent exponential residuals restore the augmented marginal.

For masked diffusion, however, each position jumps at most once, from m to a token. A passive residual changed by ERF can therefore never affect a later jump. The ERF discrete path is therefore equivalent to Gillespie since the passive residuals are never re-used. MDM ERF thus has the original reverse-chain law, so its ideal endpoint student satisfies $p^{\phi}_{0|t}(y|x_t) = p_{0|t}(y|x_t)$ and bridge sampling is exact in this case.

F.4 D-MMD

We implement D-MMD following Hoogeboom et al. (2026). Let ϕ, θ, and ψ denote the parameters of the student, frozen teacher, and auxiliary model. For masked diffusion, training uses the path

$$X_0 \sim p_0, \quad X_t \sim q_{t|0}(\cdot|X_0), \quad \widehat{X}^{\ell}_0 \sim \mathrm{Cat}\left(\hat{x}^{\phi}_0(X_t, t)^{\ell}\right), \quad X_s \sim q_{s|0,t}(\cdot|\widehat{X}_0, X_t), \qquad s < t.$$

All sampled states, including $\widehat{X}_0$ and X_s, are held fixed during differentiation, as in Hoogeboom et al. (2026, Algorithm 1). During the generator update, only the explicit student probabilities are differentiated; teacher and auxiliary predictions are fixed. During the auxiliary update, the student and teacher predictions are fixed and only the auxiliary parameters are differentiated. Thus the displayed expressions specify surrogate losses with these stop-gradient conventions, rather than derivatives through the parameter-dependent sampling law. Writing $w_s = (1 - \alpha_s)^{-1}$, for masked

diffusion we alternate the objectives

$$\begin{aligned}\mathcal{L}_{\text{gen}}(\phi) &= -\mathbb{E}\left[w_s \sum_{\ell: X_s^\ell = \mathsf{m}} \left\langle \hat{x}_0^\phi(X_t,t)^\ell, \log \hat{x}_0^\theta(X_s,s)^\ell - \log \hat{x}_0^\psi(X_s,s)^\ell \right\rangle\right], \\ \mathcal{L}_{\text{aux}}(\psi) &= -\mathbb{E}\left[w_s \sum_{\ell: X_s^\ell = \mathsf{m}} \left\langle \hat{x}_0^\phi(X_t,t)^\ell + \hat{x}_0^\theta(X_s,s)^\ell, \log \hat{x}_0^\psi(X_s,s)^\ell \right\rangle\right].\end{aligned} \tag{31}$$

For uniform diffusion, the student, teacher, and auxiliary model all use leave-one-out predictions. We sample $\widehat{X}_0^\ell \sim \mathrm{Cat}(\hat{x}_0^{\phi,\mathrm{loo}}(X_t,t)^\ell)$ and $X_s \sim q_{s|0,t}(\cdot|\widehat{X}_0, X_t)$, and use the correction of Hoogeboom et al. (2026, App. D): for $r < s$,

$$\begin{aligned}\mathcal{L}_{\text{gen}}^{\text{loo}}(\phi) = \mathbb{E}\Bigg[\frac{1}{s-r}\sum_{\ell=1}^{L} \Big\langle & q_{r|0,s}^\ell(\cdot|\hat{x}_0^{\phi,\mathrm{loo}}(X_t,t)^\ell, X_s^\ell), \\ & \log q_{r|0,s}^\ell(\cdot|\hat{x}_0^{\psi,\mathrm{loo}}(X_s,s)^\ell, X_s^\ell) - \log q_{r|0,s}^\ell(\cdot|\hat{x}_0^{\theta,\mathrm{loo}}(X_s,s)^\ell, X_s^\ell)\Big\rangle\Bigg], \\ \mathcal{L}_{\text{aux}}^{\text{loo}}(\psi) = -\mathbb{E}\Bigg[\frac{1}{s-r}\sum_{\ell=1}^{L} \Big\langle & q_{r|0,s}^\ell(\cdot|\widehat{X}_0^\ell, X_s^\ell) + q_{r|0,s}^\ell(\cdot|\hat{x}_0^{\theta,\mathrm{loo}}(X_s,s)^\ell, X_s^\ell), \\ & \log q_{r|0,s}^\ell(\cdot|\hat{x}_0^{\psi,\mathrm{loo}}(X_s,s)^\ell, X_s^\ell)\Big\rangle\Bigg].\end{aligned}$$

F.5 DCD

For a fair comparison, every uniform-diffusion baseline starts from the same checkpoints trained with the UDM objective. We therefore implement DCD from Sahoo et al. (2025) directly in the UDM state space by marginalizing its Gaussian construction. Sahoo et al. draw $G^{\ell j} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$ and define

$$X_u^\ell := \operatorname*{argmax}_{j \in [\![1,K]\!]} \left\{\widetilde{\alpha}_u \mathbb{1}_{j=X_0^\ell} + \sqrt{1-\widetilde{\alpha}_u^2}\, G^{\ell j}\right\}, \qquad u \in \{s,t\}.$$

The same $G^{\ell j}$ is used at both times, and $\widetilde{\alpha}_u$ is chosen so that $X_u^\ell \mid X_0^\ell \sim \mathrm{Cat}(\alpha_u \tilde{X}_0^\ell + (1-\alpha_u)\mathbf{1}/K)$. Since the ordering of the Gaussian coordinates $j \neq X_0^\ell$ does not depend on u, these variables can be marginalized exactly, giving

$$X_s^\ell \mid (X_0^\ell, X_t^\ell) \sim \mathrm{Cat}\left(\frac{\alpha_s - \alpha_t}{1-\alpha_t}\tilde{X}_0^\ell + \frac{1-\alpha_s}{1-\alpha_t}\tilde{X}_t^\ell\right).$$

This is a maximal coupling of the two one-position UDM marginals conditional on X_0^ℓ. We sample it directly, without Gaussian variables, by drawing independently across positions $U^\ell \sim \mathrm{Unif}(0,1)$ and $Y^\ell \sim \mathrm{Cat}(\mathbf{1}/K)$ once and setting

$$X_u^\ell = \begin{cases} Y^\ell, & U^\ell < 1-\alpha_u, \\ X_0^\ell, & U^\ell \geq 1-\alpha_u. \end{cases} \qquad u \in \{s,t\}.$$

The coupling induces the transition

$$p_{s|t}^{\max}(x_s|x_t) = \sum_{x_0 \in \mathsf{X}} p_{0|t}(x_0|x_t) \prod_{\ell=1}^{L} \mathrm{Cat}\left(x_s^\ell; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\tilde{x}_0^\ell + \frac{1-\alpha_s}{1-\alpha_t}\tilde{x}_t^\ell\right).$$

At the beginning of each round, we freeze the current student parameters ϕ^- and double the time separation,

$$\phi^- \leftarrow \phi, \qquad s = (t - \Delta_k)_+, \qquad \Delta_{k+1} = 2\Delta_k.$$

We convert both leave-one-out outputs to mean predictions using the UDM identity above and minimize the reverse KL

$$\mathcal{L}_{\mathrm{DCD}}(\phi;\phi^-) = \mathbb{E}\left[\frac{1}{L}\sum_{\ell=1}^{L}\mathrm{KL}(\hat{x}_0^{\phi}(X_t,t)^{\ell}\|\hat{x}_0^{\phi^-}(X_s,s)^{\ell})\right]. \tag{32}$$

For fixed ϕ^- and t, if the target probabilities are strictly positive and their conditional logarithms are integrable, the pointwise minimizer of (32) is therefore, for $j \in [\![1,K]\!]$,

$$\left\langle \tilde{j}, \hat{x}_0^{\phi^\star}(x_t,t)^{\ell}\right\rangle = \frac{\exp\left\{\mathbb{E}_{x_s\sim p^{\max}_{s|t}(\cdot|x_t)}\left[\log\left\langle \tilde{j}, \hat{x}_0^{\phi^-}(x_s,s)^{\ell}\right\rangle\right]\right\}}{\sum_{k=1}^{K}\exp\left\{\mathbb{E}_{x_s\sim p^{\max}_{s|t}(\cdot|x_t)}\left[\log\left\langle \tilde{k}, \hat{x}_0^{\phi^-}(x_s,s)^{\ell}\right\rangle\right]\right\}}.$$

If a target probability can vanish, this expression instead requires a nonempty common support with a positive finite normalizing constant. In particular, the exact mean-prediction boundary at $s = 0$ is a point mass at X_0^{ℓ}. If distinct clean values remain possible conditionally on X_t, the expected reverse KL against that boundary is infinite for every student distribution. An implementation including $s = 0$ must therefore specify positive target smoothing or a positive lower diffusion time; the softmax conversion above also requires an endpoint convention when $\alpha_s = 1$. At inference, each prediction is re-noised to the next time on the sampling grid, $X_s^{\ell} \sim q^{\ell}_{s|0}(\cdot|\hat{x}_0^{\phi}(X_t,t)^{\ell})$.

F.6 iDLM

We implement iDLM (Li et al., 2026) with a frozen teacher, a student, and a trainable auxiliary denoiser, initialized from the same pretrained checkpoint used by the other methods in the corresponding diffusion family. Training alternates student and auxiliary updates. The student minimizes the difference between teacher and auxiliary denoising losses on its generated distribution, while the auxiliary denoiser is trained on that distribution. For masked diffusion, sampled categorical states are held fixed during differentiation and gradients act on the student probabilities appearing explicitly in the loss.

For uniform diffusion, we retain the Gaussian relaxation used by iDLM to obtain differentiable noisy inputs, with relaxation temperature 10^{-3}. The teacher remains the shared pretrained UDM checkpoint; no separately pretrained Gaussian diffusion model is introduced. Thus the Gaussian relaxation is retained for iDLM training, whereas the Gaussian variables in DCD's coupled forward samples are marginalized as described in Appendix F.5. At evaluation, iDLM generates discrete tokens and is subject to the same sampler and temperature controls as the other students.

G Experimental Details

G.1 Sudoku

We train one MDM and one UDM teacher per difficulty. Each example concatenates an 81-digit quiz, a separator, its 81-digit solution, and an end-of-sequence token; 0 denotes a blank quiz position. The quiz and separator remain fixed, while the model predicts the solution positions. The vocabulary includes ten digits and five special tokens for padding, separation, masking, end of sequence, and unknown inputs. Tables 3 to 5 summarize the configuration. Exact-grid accuracy requires all 81 predicted positions to match the reference solution; constraint violations are also computed.

Table 3: Sudoku datasets, generated with seed 42.

Difficulty	Clues	Training examples	Validation examples	Workers
Easy	40	48,000	2,000	4
Medium	35	48,000	2,000	4
Hard	30	48,000	2,000	8

Table 4: Sudoku teacher architecture and diffusion configuration.

Parameter	MDM UDM
Backbone	DiT-style transformer
Transformer blocks	8
Hidden dimension	512
Attention heads	8
Conditioning dimension	128
Padded/truncated sequence length	164
Vocabulary size	15
Dropout	0.1
Trainable parameters	28,562,191

Table 5: Shared Sudoku teacher training recipe, applied to all six runs.

Parameter	Value
Optimizer steps	20,000
Batch size	256
Numerical precision	bfloat16
Optimizer	AdamW
Peak learning rate	3×10^{-4}
Learning-rate schedule	2,500 warmup steps, then constant
AdamW (β_1, β_2)	$(0.9, 0.999)$
Optimizer epsilon	10^{-8}
Weight decay	0
Gradient clipping	Maximum norm 1
Diffusion loss	ELBO
Antithetic time sampling	Enabled
Sampling epsilon	10^{-3}
Training seed	1
EMA decay	0.9999

H ADDITIONAL EXPERIMENTS

H.1 SAMPLING QUALITY ACROSS NFE BUDGETS

Figures 5 and 6 extend the main comparison to all eight evaluated NFE budgets, from 8 to 1,024. The samplers share a pretrained checkpoint within each diffusion family. TinyGSM evaluations use all 1,319 GSM8K test questions and temperatures $\{0.1, 0.2, \ldots, 1\}$. Accuracy is the fraction of correct answers, with pointwise 95% percentile-bootstrap intervals over questions. OpenWebText evaluations use temperatures $\{0.80, 0.82, \ldots, 1.10\}$ and report GPT-2 Large generative perplexity together with token entropy. The ancestral evaluations contain 960 generated sequences per point and the ERF evaluations contain 480.

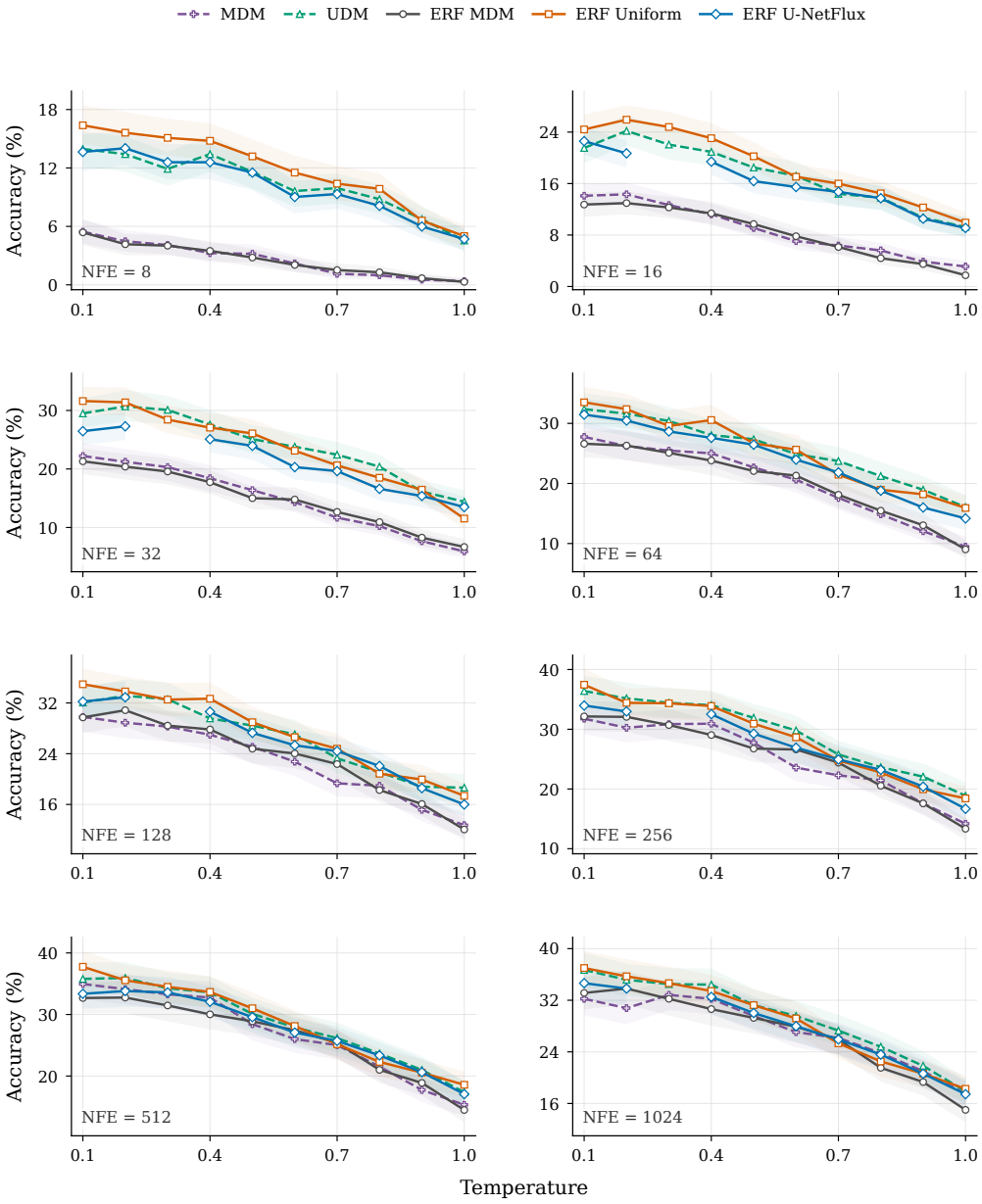

Figure 5: TinyGSM sampling across all evaluated NFE budgets. Each panel gives GSM8K accuracy across sampling temperatures, with pointwise 95% bootstrap intervals. Dashed lines denote ancestral sampling and solid lines denote ERF.

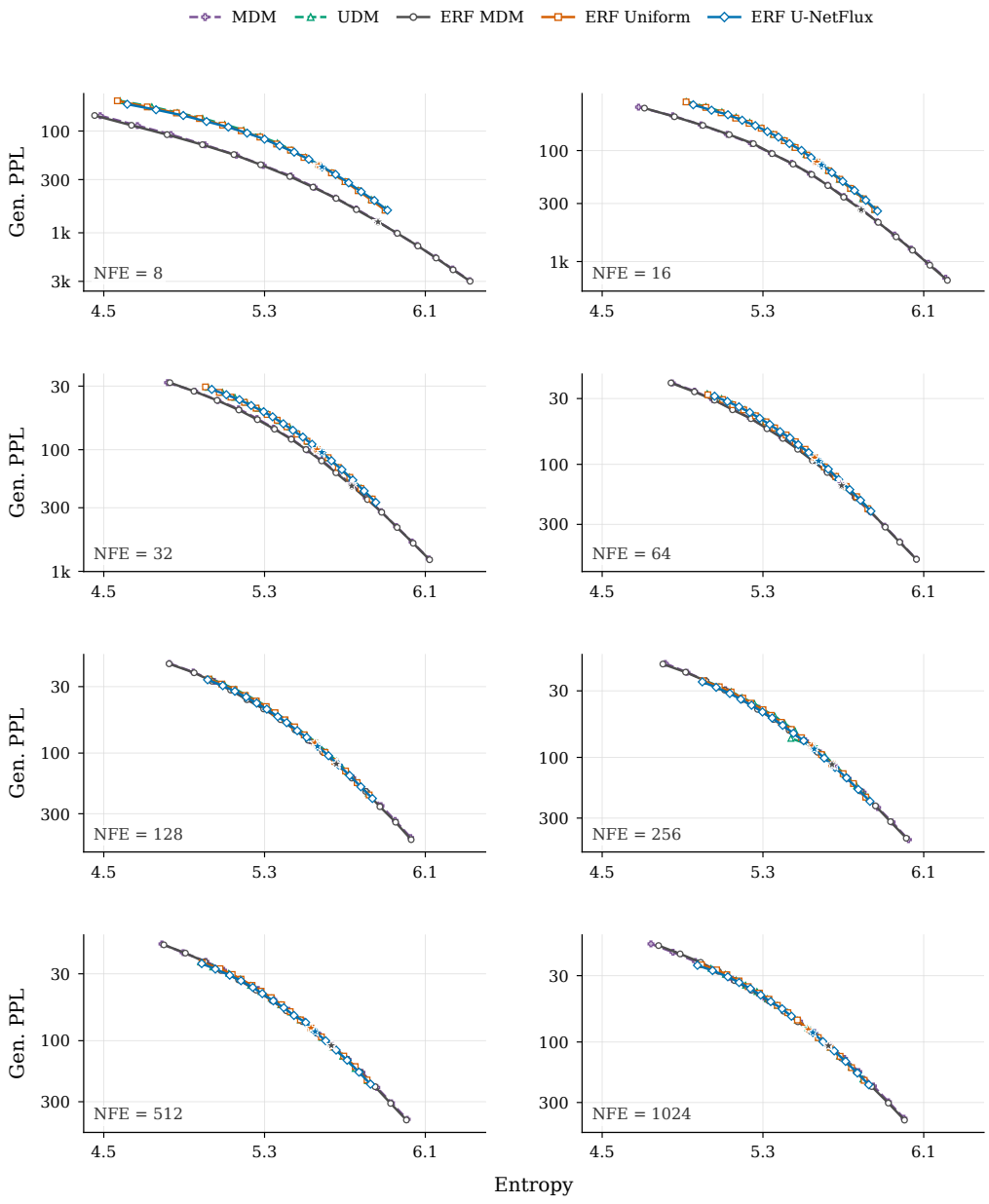

Figure 6: OpenWebText sampling across all evaluated NFE budgets. Markers follow the temperature grid from 0.80 to 1.10, with stars at 1. Perplexity uses an inverted logarithmic axis, scaled independently in each panel.

H.2 EULER AND CATEGORICAL TAU-LEAPING

Figures 7 to 9 compare ERF with the Euler and categorical τ-leaping schemes described in Appendix D.4, separately for masked, uniform, and U-NetFlux rates at 8, 64, and 512 NFEs. Each rate family uses one pretrained checkpoint, and the masked and uniform comparisons also include the ancestral controls from the main sampling experiment. TinyGSM evaluations use all 1,319 GSM8K test questions at temperatures $\{0.1, 0.2, \ldots, 1\}$. OpenWebText evaluations use temperatures $\{0.80, 0.82, \ldots, 1.10\}$, with 960 sequences per point for the reused ancestral controls and 480 for ERF, Euler, and categorical τ-leaping; all runs use seed 1. The ancestral UDM reference applies temperature to the raw model logits before the prediction conversion, whereas the other uniform samplers apply it to the converted prediction.

The Euler implementation clips the categorical weights and normalizes them when the frozen-rate step violates the probability constraint, while categorical τ-leaping rejects multiple proposed jumps at a position. Both baselines replace the final scheduled update with a denoiser prediction.

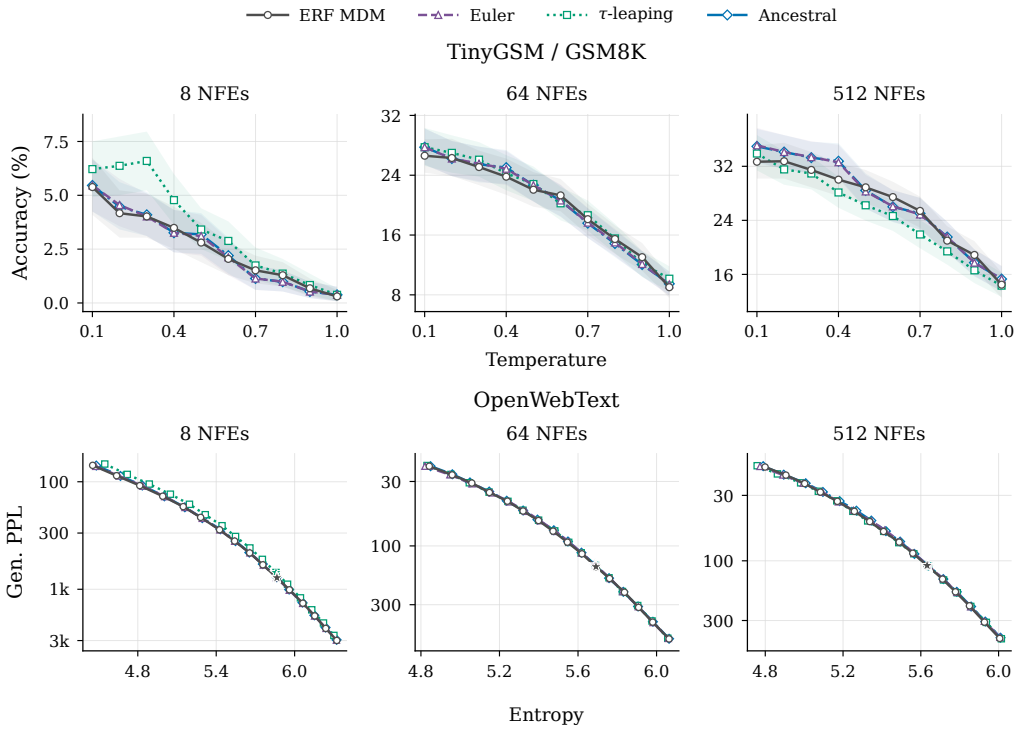

Figure 7: ERF, ancestral sampling, Euler, and categorical τ-leaping for masked diffusion. The top row shows GSM8K accuracy across sampling temperatures, with pointwise 95% bootstrap intervals; the bottom row shows OpenWebText entropy and GPT-2 Large generative perplexity. Stars mark temperature 1. The perplexity axes are logarithmic and inverted, and panel limits follow the observed ranges.

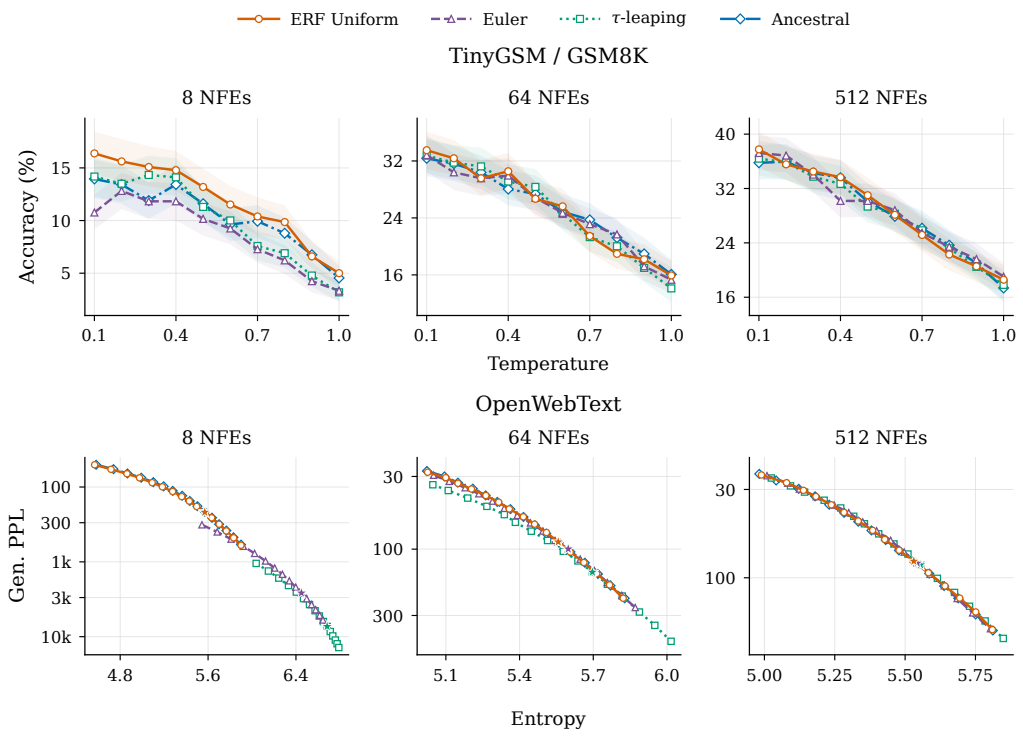

Figure 8: ERF Uniform, ancestral UDM, Euler, and categorical τ-leaping for uniform diffusion, using the same plotting conventions as Figure 7. The ancestral reference uses temperature scaling before prediction conversion and 960 OpenWebText sequences per point; the other methods use scaling after conversion and 480 sequences. All four samplers share the same pretrained checkpoint.

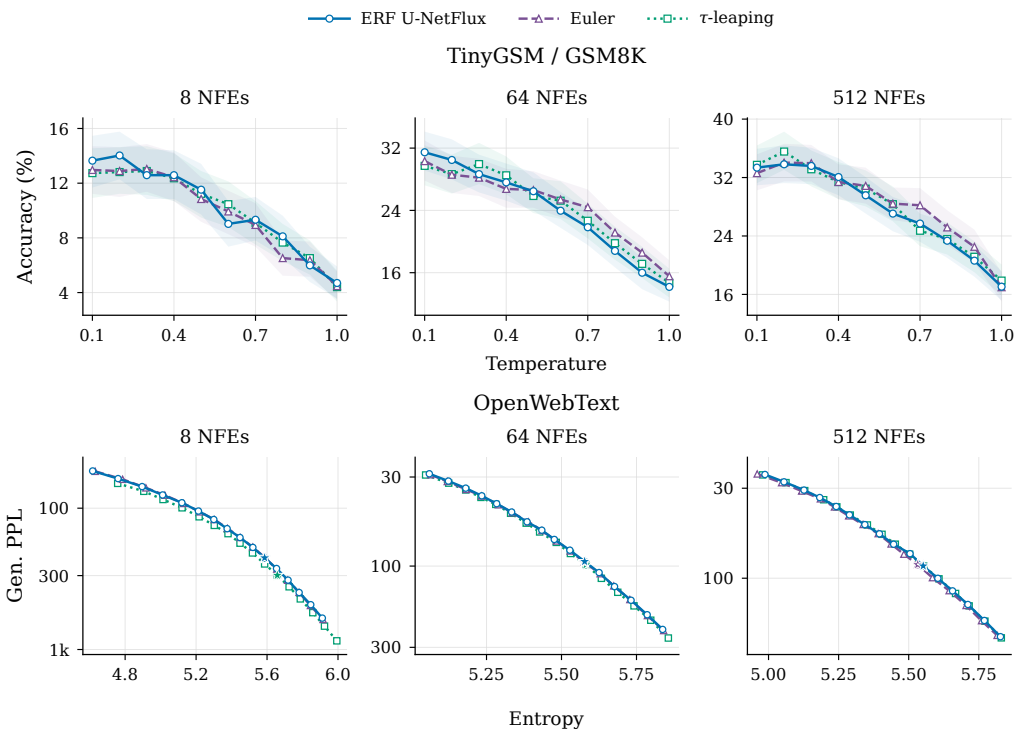

Figure 9: ERF U-NetFlux, Euler, and categorical τ-leaping with U-NetFlux rates, using the same evaluation and plotting conventions as Figure 7. Each panel compares the three integration schemes with the same pretrained uniform-diffusion checkpoint.

H.3 Uniform substeps

The bounded variants divide each frozen-model interval into $1, 2, 4, 8, 16$, or 32 substeps and permit at most one jump per position within a substep. The unbounded variant resolves every crossing. Substeps do not introduce additional model evaluations. Figures 10 and 11 display all completed settings across the eight NFE budgets. The TinyGSM grid uses temperatures 0.1, 0.5, and 1; the OpenWebText grid uses $\{0.80, 0.84, \ldots, 1.08\}$.

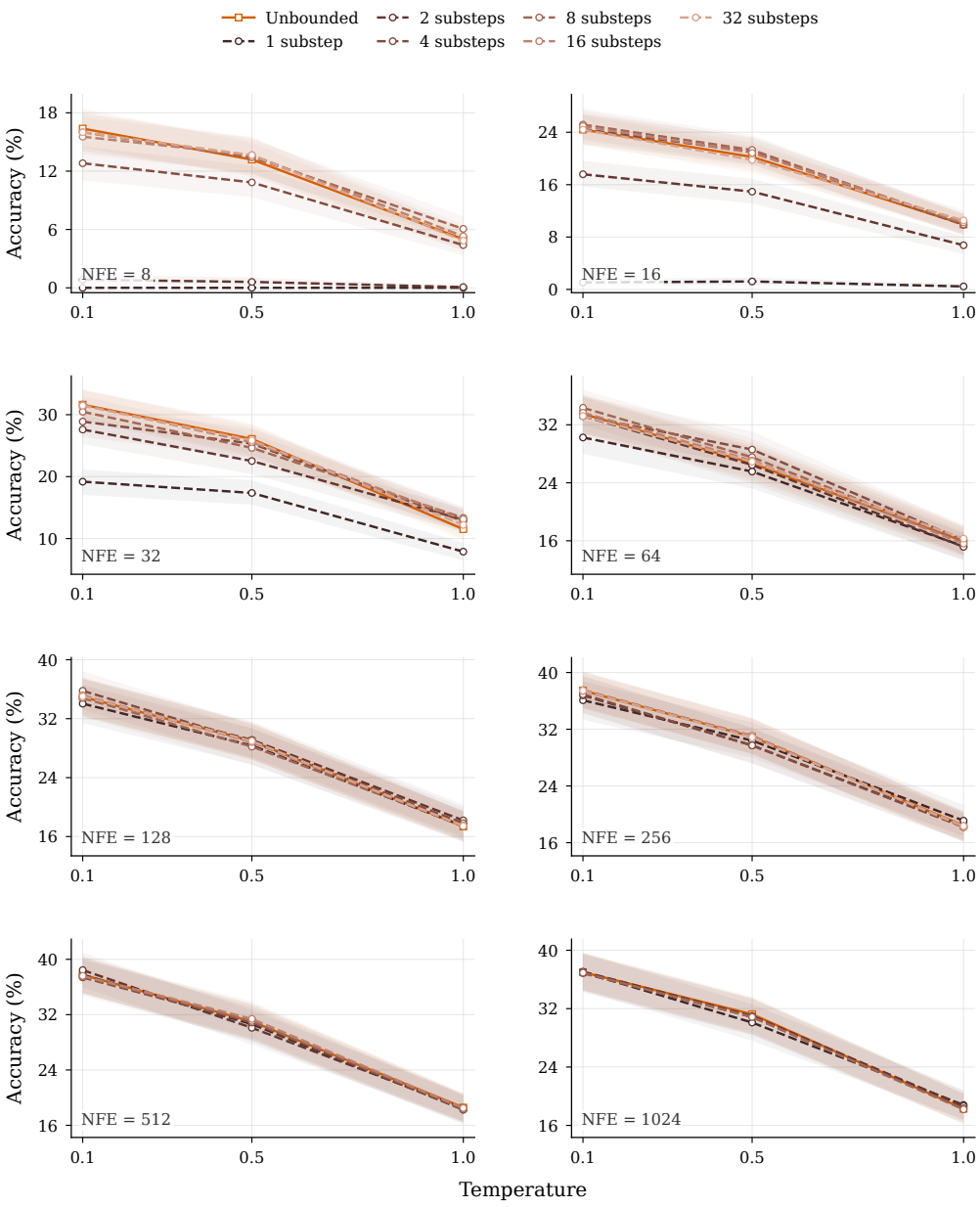

Figure 10: The full TinyGSM substep comparison. Curves show GSM8K accuracy at fixed NFE as the sampling temperature varies; bands are pointwise 95% bootstrap intervals. Each legend entry specifies the number of substeps per frozen-model interval. Missing high-substep settings at larger NFE budgets are left absent.

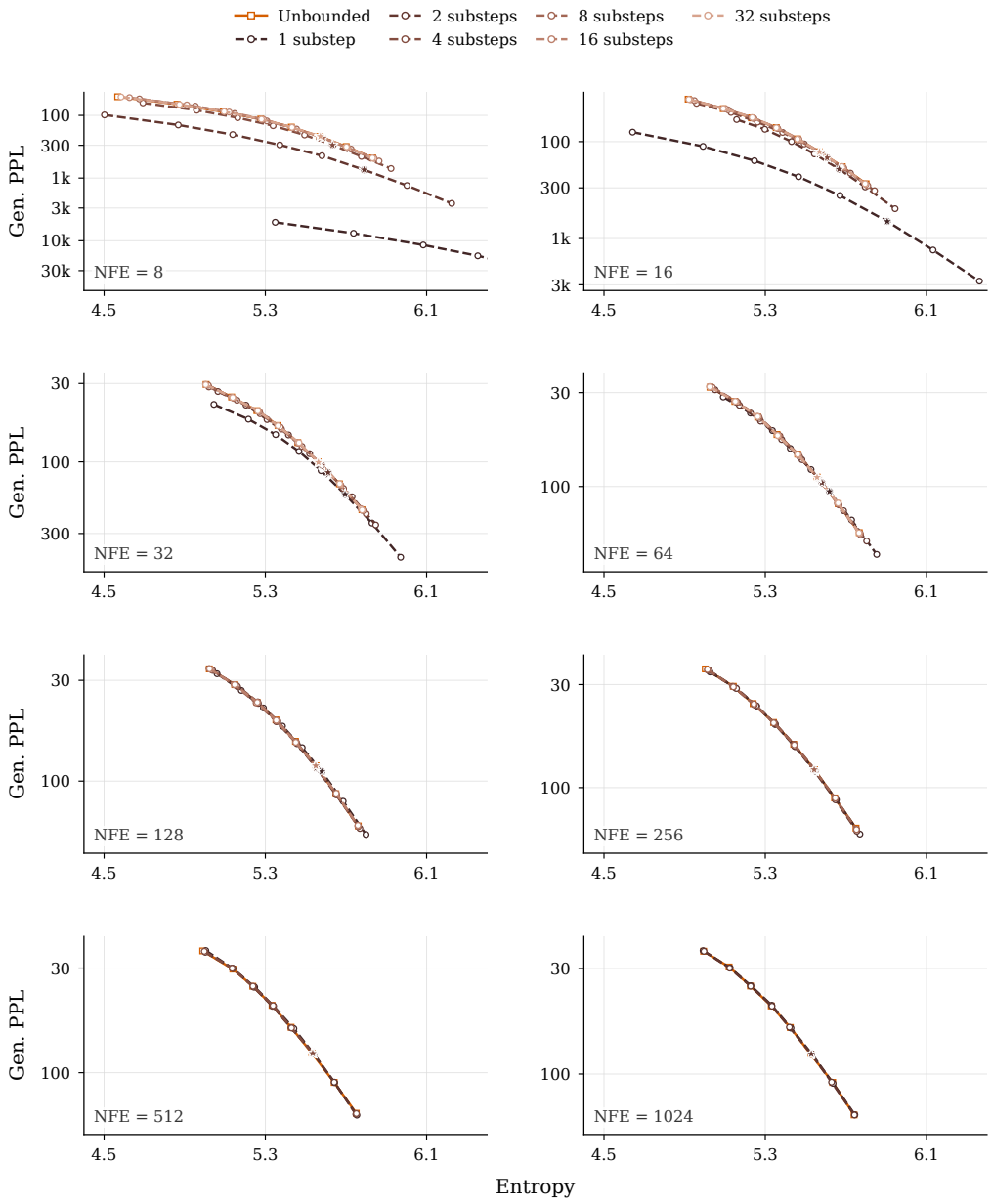

Figure 11: The full OpenWebText substep comparison. Curves show the entropy–perplexity trade-off over eight temperatures, with stars at 1. The inverted logarithmic perplexity axis retains the observed range in each panel. Only completed evaluations are drawn, with lines interrupted at missing temperatures.

H.4 Runtime and jump-count protocol

The profiles use one AMD MI210, batches of four sequences, and generator seeds 1, 2, and 3. For TinyGSM, the three batches contain the first twelve GSM8K test prompts, four successive prompts per seed. Prompt positions remain fixed in a 512-token canvas. OpenWebText generation is unconditional with length 1,024. Sampling temperatures are 0.1 for TinyGSM and 1 for OpenWebText. Both methods use the same checkpoint, bfloat16 model autocast, and float64 residual clocks.

Uniform explicitly uses 40 bisection iterations per crossing-time solve. Its terminal time is 10^{-6}, while U-NetFlux uses terminal time zero. Each method receives one warmup before measurement. Timings include sampling and exclude checkpoint loading and metric computation. Device synchronization brackets each timing interval.

Jump counts come from a separate replay with the same generator seed and prompt batch. We check that its final tokens agree exactly with the uninstrumented run. Counts measure actual token changes and are divided by the number of generated positions, excluding fixed prompt positions. Figure 12 exposes the three individual measurements; Table 6 gives their means. These runs measure sampling cost on the stated batches.

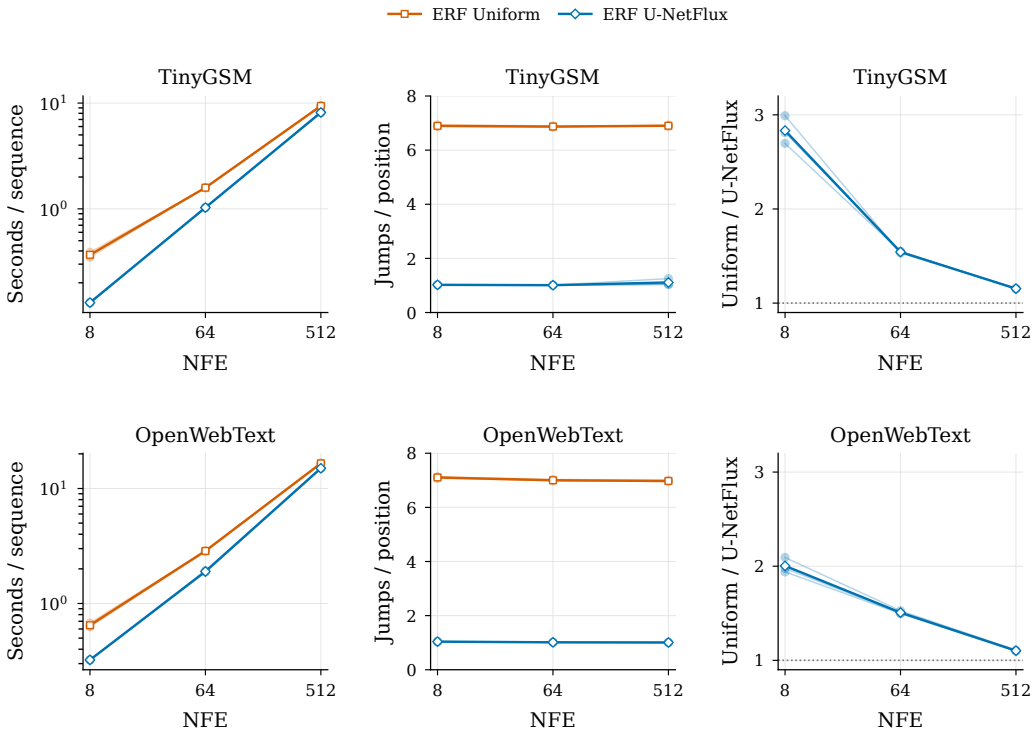

Figure 12: Individual runtime and jump-count measurements. Faint curves and points show the three seeds; the main curves show their means. In the right column, the main curve is the ratio of mean Uniform and U-NetFlux runtimes, while faint curves give the three paired ratios. The dotted line marks equal runtime.

Table 6: Mean sampling time in seconds per sequence, speedup, and jumps per generated position across the three profile batches. Speedup is the ratio of mean Uniform and U-NetFlux runtimes.

Dataset	NFE	Uniform (s)	U-NetFlux (s)	Speedup	Uniform jumps	U-NetFlux jumps
TinyGSM	8	0.368	0.130	2.83	6.90	1.02
TinyGSM	64	1.586	1.028	1.54	6.87	1.01
TinyGSM	512	9.407	8.154	1.15	6.90	1.11
OWT	8	0.647	0.323	2.00	7.11	1.04
OWT	64	2.862	1.899	1.51	7.00	1.01
OWT	512	16.544	15.005	1.10	6.98	1.01

H.5 Distillation evaluation

The main distillation figures use the completed runs with budgets of 50 GPU-hours per student on TinyGSM and 100 GPU-hours on OpenWebText, all with global batch size 128. The budget is the number of allocated GPUs multiplied by elapsed training-job time, including initialization, checkpointing and training validation, but excluding queue time and independent evaluation jobs. The number of student updates therefore varies by method. All methods share the original temperature-1 teacher checkpoint within each dataset and diffusion family, with top-p truncation disabled. Uniform iDLM retains its Gaussian relaxation with the shared teacher, as detailed in Appendix F.6.

The main figures use the final student EMA checkpoints from training seed 1. All student and matching original-teacher evaluations use sampling seed 101, so no across-seed uncertainty band is shown. ERC refreshes its frozen EMA target and rollout teacher at 25%, 50% and 75% of the training budget. We evaluate gaps $\delta \in \{1/8, 1/16, 1/32, 1/64\}$ and select one gap per dataset and diffusion family, held fixed across NFEs. The selected gaps are $1/64$ for masked TinyGSM, $1/8$ for uniform TinyGSM, $1/64$ for masked OpenWebText and $1/16$ for uniform OpenWebText. UDM predicts mean LOO probabilities and converts them to mean predictions for ERC. The ERC objective is reverse KL on TinyGSM and forward KL on OpenWebText.

The main DCD curves use the final native consistency-distillation runs. The completed evaluation grid contains both bridge and refresh sampling for every student and teacher, with denoiser temperature placement for MDM and both denoiser and LOO placement for UDM. At temperature 1, the two UDM temperature placements reuse the same evaluation within each sampler; these entries are aliases, not independent samples.

TinyGSM students and teachers in the main figure are evaluated on the same first 1,024 GSM8K test questions, with a 512-token canvas. The common sampling temperatures are 0.1, 0.23, 0.36, 0.49, 0.61, 0.74, 0.87 and 1, evaluated at $\{2, 4, 8, 16, 32, 64\}$ NFEs. We maximize accuracy averaged with equal weight over these six NFE budgets to select one sampler, temperature placement and temperature per method; ERC additionally selects one gap. The selected configuration is fixed for the entire curve, and exact temperature ties favor the lower temperature. These are descriptive grid selections without a separate tuning set. The complete sampler-specific comparisons appear in Appendix H.9.

The main OpenWebText figure uses bridge sampling and denoiser temperature placement for every method and teacher, at temperatures 0.8, 0.85, 0.9, 0.95, 1, 1.05, 1.1 and 1.2. Every point uses 2,048 sequences of length 1,024. To select the ERC gap within each family, we restrict each NFE comparison to the entropy interval shared by all four gaps, linearly interpolate log-perplexity as a function of entropy, and integrate over that interval, dividing by its width. We choose the gap with the lowest average of these normalized integrals over $\{2, 4, 8, 16, 32, 64\}$ NFEs. This compares perplexity at matched entropy without extrapolation and fixes the same gap throughout the figure. The displayed curves contain measured points only. Appendices H.8 and H.9 give the complete measured gap and sampler sweeps; their OpenWebText panels fit the visible points subject to entropy bounds 4–6 and a generative-perplexity ceiling of 1,000, clipping values outside these limits.

H.6 SUDOKU DISTILLATION

Tables 7 and 8 report exact-grid accuracy on the 2,000 validation puzzles at each difficulty and NFE budget in $\{1, 2, 4, 8, 16, 32, 64\}$. We use EMA weights throughout and average sampling seeds 101 and 303 before selecting a configuration. For each method and NFE budget, the tables retain the highest two-seed mean over the ancestral bridge and categorical refresh samplers and over temperatures $\{0.1, 0.25, 0.5, 0.7, 0.8, 0.9, 1, 1.1\}$. The ERC checkpoint is selected at 8 NFEs with refresh sampling and then held fixed across NFE budgets. These selections are descriptive maxima over the evaluation grid rather than estimates from a separate validation sweep.

Table 7: Sudoku MDM exact-grid accuracy in percent. Subscripts report the selected sampler and temperature, with R denoting categorical refresh and B denoting the ancestral bridge. Bold denotes the highest accuracy within each difficulty and NFE budget.

Difficulty	Method	1 NFE	2 NFE	4 NFE	8 NFE	16 NFE	32 NFE	64 NFE
Easy	Teacher	$79.28_{\mathrm{R},\,.1}$	$84.98_{\mathrm{R},\,.1}$	$91.08_{\mathrm{R},\,.1}$	$95.08_{\mathrm{R},\,.25}$	$97.08_{\mathrm{R},\,.1}$	$98.20_{\mathrm{R},\,.1}$	$\mathbf{99.20}_{\mathrm{R},\,.1}$
	ERC (ours)	$\mathbf{81.08}_{\mathrm{B},\,.1}$	$\mathbf{86.15}_{\mathrm{R},\,.1}$	$\mathbf{92.88}_{\mathrm{R},\,.1}$	$\mathbf{95.85}_{\mathrm{R},\,.1}$	$\mathbf{97.45}_{\mathrm{R},\,.25}$	$\mathbf{98.45}_{\mathrm{R},\,.1}$	$98.90_{\mathrm{R},\,.5}$
	D-MMD	$78.43_{\mathrm{B},\,.5}$	$84.18_{\mathrm{R},\,.1}$	$91.98_{\mathrm{R},\,.25}$	$95.03_{\mathrm{R},\,.1}$	$97.13_{\mathrm{R},\,1}$	$98.40_{\mathrm{R},\,.1}$	$99.13_{\mathrm{R},\,.5}$
	iDLM	$79.70_{\mathrm{B},\,.1}$	$85.03_{\mathrm{R},\,.1}$	$91.25_{\mathrm{R},\,.1}$	$95.08_{\mathrm{R},\,.7}$	$97.10_{\mathrm{R},\,.25}$	$98.23_{\mathrm{R},\,.1}$	$99.10_{\mathrm{R},\,.9}$
	SDTT	$68.73_{\mathrm{R},\,.1}$	$74.98_{\mathrm{R},\,.1}$	$84.10_{\mathrm{R},\,.1}$	$89.95_{\mathrm{R},\,.1}$	$93.78_{\mathrm{R},\,.25}$	$96.05_{\mathrm{R},\,.25}$	$97.43_{\mathrm{R},\,.9}$
Medium	Teacher	$39.40_{\mathrm{R},\,.1}$	$50.18_{\mathrm{R},\,.1}$	$66.53_{\mathrm{R},\,.25}$	$76.88_{\mathrm{R},\,.1}$	$84.33_{\mathrm{R},\,.1}$	$89.93_{\mathrm{R},\,.1}$	$92.65_{\mathrm{R},\,.1}$
	ERC (ours)	$41.43_{\mathrm{B},\,.1}$	$\mathbf{51.85}_{\mathrm{R},\,.25}$	$\mathbf{68.28}_{\mathrm{R},\,.25}$	$\mathbf{78.60}_{\mathrm{R},\,.7}$	$\mathbf{84.88}_{\mathrm{R},\,.1}$	$88.98_{\mathrm{R},\,.5}$	$92.53_{\mathrm{R},\,.25}$
	D-MMD	$39.05_{\mathrm{B},\,.1}$	$49.48_{\mathrm{B},\,.7}$	$66.20_{\mathrm{R},\,.25}$	$77.33_{\mathrm{R},\,1.1}$	$83.58_{\mathrm{R},\,1}$	$89.05_{\mathrm{R},\,.5}$	$92.43_{\mathrm{R},\,.1}$
	iDLM	$\mathbf{41.58}_{\mathrm{B},\,.1}$	$51.35_{\mathrm{R},\,.1}$	$66.50_{\mathrm{R},\,.7}$	$76.65_{\mathrm{R},\,.25}$	$83.75_{\mathrm{R},\,.25}$	$\mathbf{90.40}_{\mathrm{R},\,.1}$	$\mathbf{93.05}_{\mathrm{R},\,.1}$
	SDTT	$12.45_{\mathrm{B},\,.1}$	$21.75_{\mathrm{R},\,.1}$	$40.53_{\mathrm{R},\,.1}$	$55.15_{\mathrm{R},\,.25}$	$66.85_{\mathrm{R},\,.1}$	$76.05_{\mathrm{R},\,.1}$	$82.50_{\mathrm{R},\,.1}$
Hard	Teacher	$2.68_{\mathrm{B},\,.1}$	$7.23_{\mathrm{B},\,.1}$	$20.75_{\mathrm{R},\,.1}$	$36.10_{\mathrm{R},\,.1}$	$\mathbf{47.88}_{\mathrm{R},\,.1}$	$57.30_{\mathrm{R},\,.1}$	$66.35_{\mathrm{R},\,.25}$
	ERC (ours)	$3.70_{\mathrm{B},\,.25}$	$8.10_{\mathrm{R},\,.25}$	$\mathbf{22.23}_{\mathrm{R},\,.1}$	$\mathbf{36.58}_{\mathrm{R},\,.1}$	$47.75_{\mathrm{R},\,.1}$	$\mathbf{57.38}_{\mathrm{R},\,.8}$	$\mathbf{66.43}_{\mathrm{R},\,.8}$
	D-MMD	$3.58_{\mathrm{B},\,.25}$	$8.05_{\mathrm{B},\,.25}$	$20.95_{\mathrm{R},\,.25}$	$34.70_{\mathrm{R},\,.25}$	$46.48_{\mathrm{R},\,.25}$	$56.38_{\mathrm{R},\,.8}$	$65.05_{\mathrm{R},\,1.1}$
	iDLM	$\mathbf{4.13}_{\mathrm{B},\,.1}$	$\mathbf{8.83}_{\mathrm{B},\,.1}$	$21.85_{\mathrm{R},\,.1}$	$35.72_{\mathrm{R},\,.5}$	$46.75_{\mathrm{R},\,.25}$	$56.45_{\mathrm{R},\,.1}$	$66.10_{\mathrm{R},\,.7}$
	SDTT	$0.55_{\mathrm{B},\,.1}$	$2.65_{\mathrm{R},\,.1}$	$11.45_{\mathrm{R},\,.25}$	$22.60_{\mathrm{R},\,.1}$	$31.78_{\mathrm{R},\,.1}$	$41.25_{\mathrm{R},\,.1}$	$51.43_{\mathrm{R},\,.1}$

Table 8: Sudoku UDM exact-grid accuracy in percent. Mean- and LOO-parameterized UDM use distinct teachers and method coverage. Selection, subscripts and boldface follow Table 7.

Difficulty	Method	1 NFE	2 NFE	4 NFE	8 NFE	16 NFE	32 NFE	64 NFE
Easy	Teacher	$74.50_{\mathrm{R},\,.1}$	$92.85_{\mathrm{R},\,.1}$	$96.83_{\mathrm{R},\,.8}$	$98.05_{\mathrm{R},\,.8}$	$98.70_{\mathrm{R},\,.8}$	$99.35_{\mathrm{R},\,.1}$	$\mathbf{99.63}_{\mathrm{R},\,.5}$
	ERC (ours)	$\mathbf{74.65}_{\mathrm{B},\,.1}$	$\mathbf{93.28}_{\mathrm{R},\,.1}$	$\mathbf{96.95}_{\mathrm{R},\,1.1}$	$\mathbf{98.25}_{\mathrm{R},\,1.1}$	$\mathbf{98.93}_{\mathrm{R},\,.1}$	$\mathbf{99.55}_{\mathrm{R},\,.1}$	$99.58_{\mathrm{R},\,.7}$
	D-MMD	$73.13_{\mathrm{B},\,.1}$	$92.58_{\mathrm{R},\,.25}$	$96.55_{\mathrm{R},\,.1}$	$98.18_{\mathrm{R},\,.8}$	$98.73_{\mathrm{R},\,1.1}$	$99.30_{\mathrm{R},\,.9}$	$99.48_{\mathrm{R},\,.1}$
	DCD	$74.50_{\mathrm{B},\,.1}$	$92.30_{\mathrm{R},\,.5}$	$96.55_{\mathrm{R},\,.1}$	$97.80_{\mathrm{R},\,.8}$	$98.50_{\mathrm{R},\,.5}$	$99.13_{\mathrm{R},\,.25}$	$99.28_{\mathrm{R},\,.25}$
Medium	Teacher	$40.03_{\mathrm{R},\,.1}$	$74.38_{\mathrm{R},\,.25}$	$85.95_{\mathrm{R},\,.7}$	$90.78_{\mathrm{R},\,1.1}$	$\mathbf{94.00}_{\mathrm{R},\,1}$	$95.55_{\mathrm{R},\,.1}$	$\mathbf{97.00}_{\mathrm{R},\,.25}$
	ERC (ours)	$40.95_{\mathrm{B},\,.1}$	$\mathbf{75.78}_{\mathrm{R},\,.1}$	$\mathbf{86.88}_{\mathrm{R},\,.25}$	$\mathbf{91.23}_{\mathrm{R},\,1}$	$93.93_{\mathrm{R},\,.7}$	$\mathbf{95.85}_{\mathrm{R},\,1}$	$96.98_{\mathrm{R},\,1.1}$
	D-MMD	$37.73_{\mathrm{B},\,.1}$	$74.88_{\mathrm{R},\,.25}$	$86.00_{\mathrm{R},\,1.1}$	$90.40_{\mathrm{R},\,.8}$	$93.48_{\mathrm{R},\,.7}$	$95.55_{\mathrm{R},\,.7}$	$96.60_{\mathrm{R},\,.1}$
	DCD	$\mathbf{41.25}_{\mathrm{B},\,.1}$	$74.58_{\mathrm{R},\,.8}$	$85.63_{\mathrm{R},\,.5}$	$90.53_{\mathrm{R},\,1.1}$	$93.28_{\mathrm{R},\,.1}$	$95.20_{\mathrm{R},\,.9}$	$96.65_{\mathrm{R},\,.25}$
Hard	Teacher	$7.80_{\mathrm{B},\,.1}$	$39.78_{\mathrm{B},\,.1}$	$57.55_{\mathrm{R},\,.25}$	$\mathbf{68.35}_{\mathrm{R},\,.8}$	$\mathbf{74.78}_{\mathrm{R},\,.7}$	$80.25_{\mathrm{R},\,.9}$	$85.73_{\mathrm{R},\,1}$
	ERC (ours)	$7.50_{\mathrm{B},\,.1}$	$\mathbf{40.80}_{\mathrm{R},\,.1}$	$\mathbf{58.88}_{\mathrm{R},\,.1}$	$67.38_{\mathrm{R},\,1}$	$74.63_{\mathrm{R},\,.8}$	$\mathbf{80.40}_{\mathrm{R},\,.8}$	$\mathbf{86.20}_{\mathrm{R},\,1.1}$
	D-MMD	$6.95_{\mathrm{B},\,.1}$	$39.08_{\mathrm{B},\,.25}$	$57.58_{\mathrm{R},\,.9}$	$66.80_{\mathrm{R},\,.5}$	$74.05_{\mathrm{R},\,1.1}$	$79.55_{\mathrm{R},\,.1}$	$85.00_{\mathrm{R},\,.8}$
	DCD	$\mathbf{7.93}_{\mathrm{B},\,.25}$	$39.25_{\mathrm{B},\,.25}$	$57.78_{\mathrm{R},\,.25}$	$67.98_{\mathrm{R},\,.5}$	$74.70_{\mathrm{R},\,1}$	$79.55_{\mathrm{R},\,.1}$	$84.68_{\mathrm{R},\,1.1}$

H.7 TARGET AND TEACHER UPDATES

The target is the stopped-gradient student on the right of the consistency loss; the transport teacher is the denoiser used to generate the paired ERF states. At a refresh, the target is replaced by the student EMA; target-and-teacher refresh also uses that copy for transport. Over 10,000 updates, we compare a frozen target, per-step EMA, target refresh every 2,500 updates, and target-and-teacher refresh. Each uses either a fixed gap $\delta = 1/64$ or a gap increasing over four rounds from $1/64$ to $1/8$. At fixed gap, we also test 5,000-update target refresh and EMA restricted to the target's conditioning projection. TinyGSM uses reverse KL and both inference samplers; OWT uses forward KL and bridge sampling. The figures use sampling seed 101 and the measured temperature grid at 4–64 NFEs.

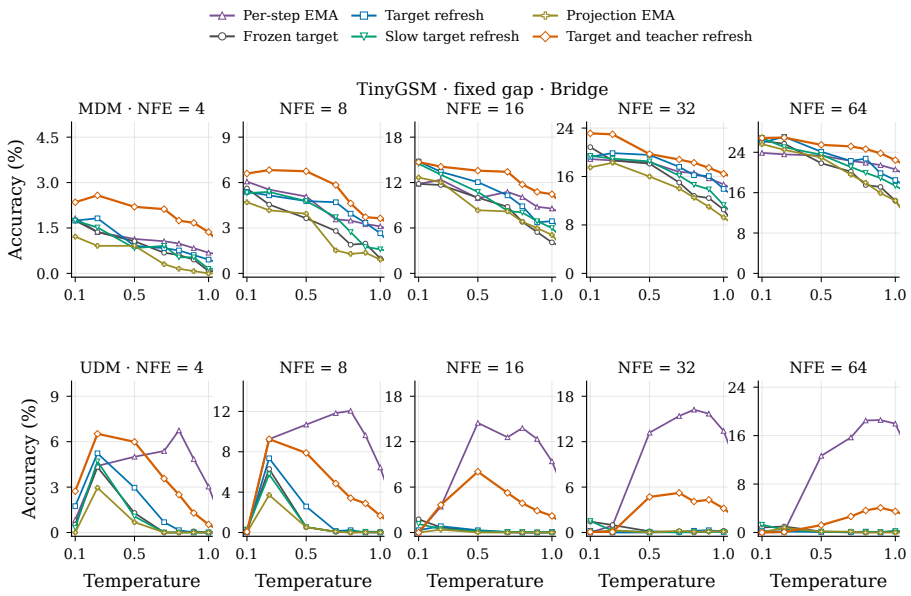

Figure 13: TinyGSM target-update policies with a fixed training gap, under bridge sampling.

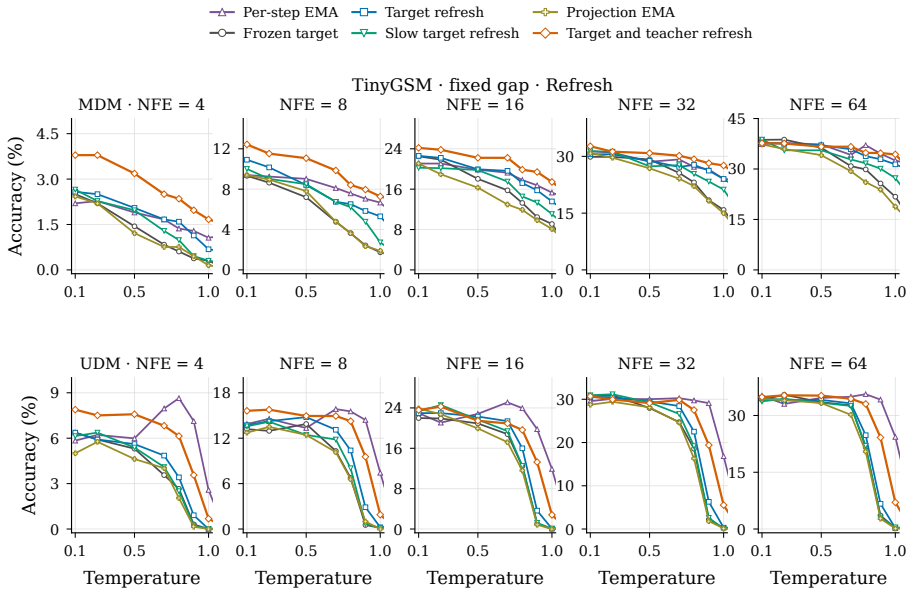

Figure 14: The same fixed-gap policies under refresh sampling.

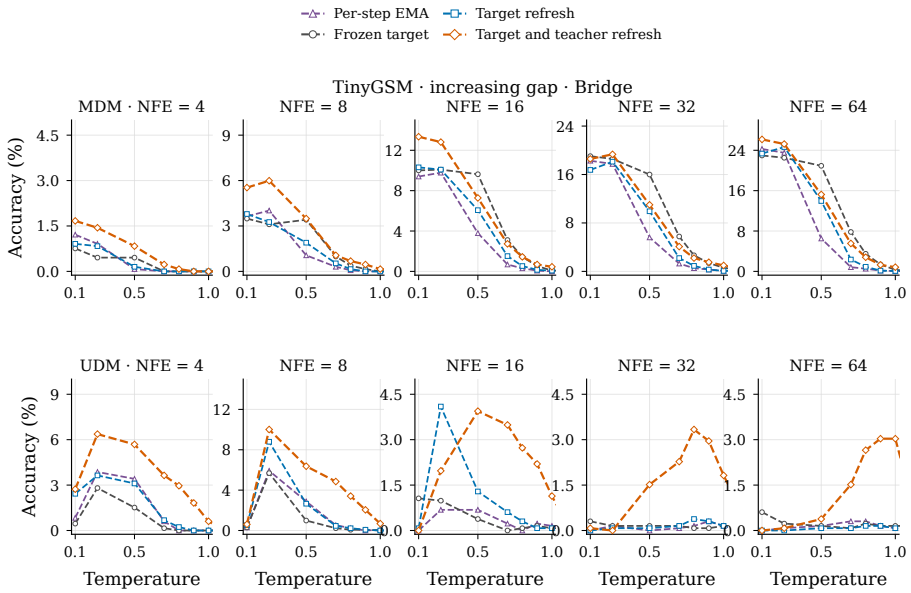

Figure 15: TinyGSM target-update policies with an increasing training gap, under bridge sampling.

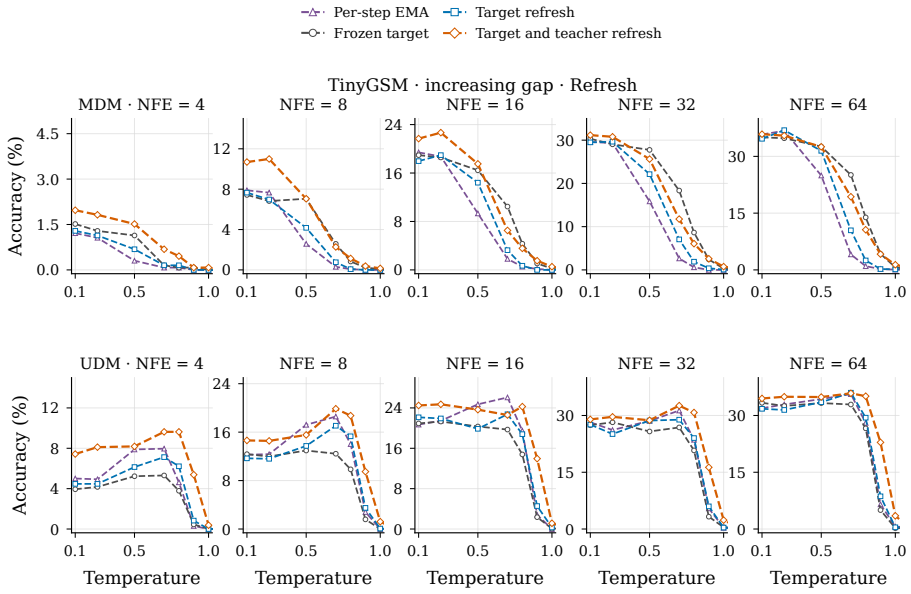

Figure 16: The same increasing-gap policies under refresh sampling.

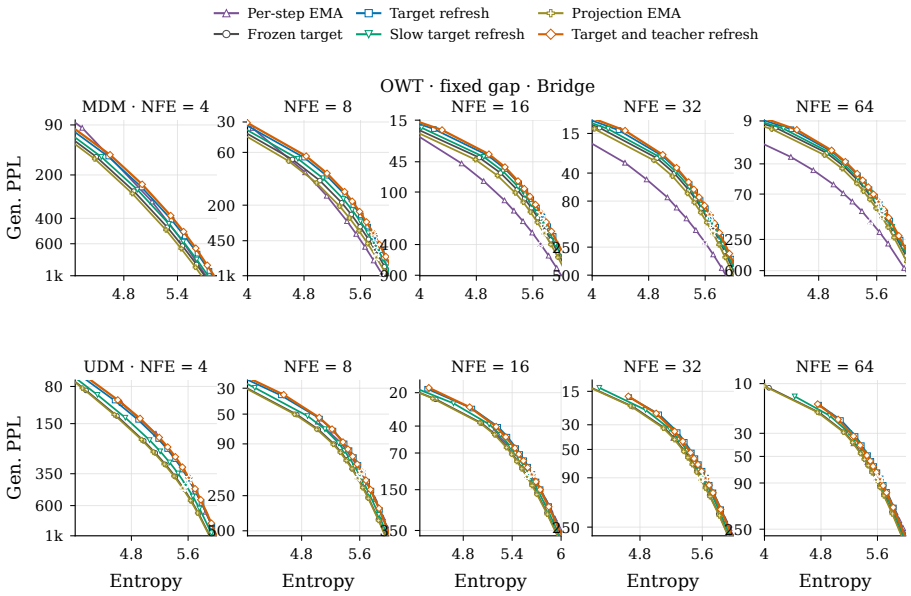

Figure 17: OWT target-update policies with a fixed training gap, under bridge sampling. Stars mark temperature 1.

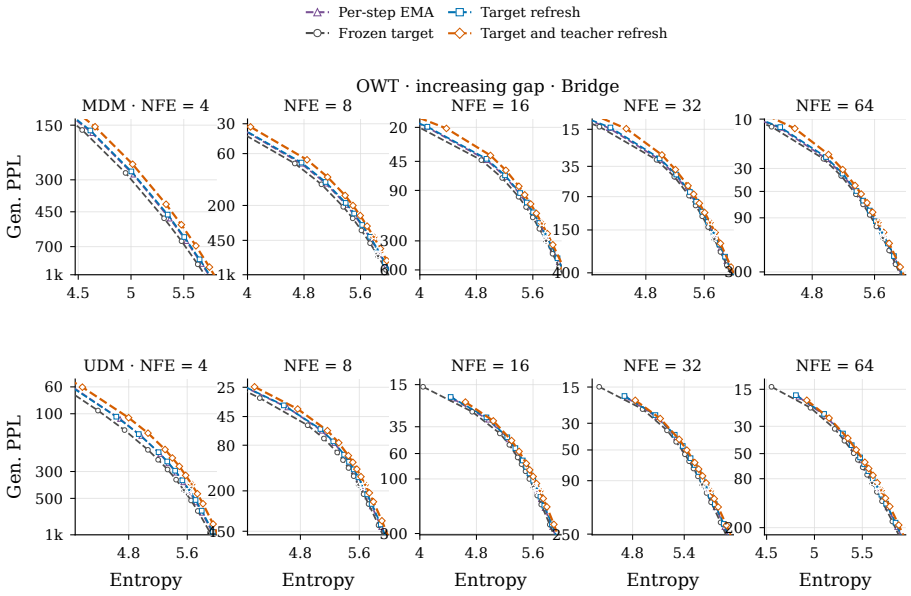

Figure 18: The same OWT policies with an increasing gap. Axes fit visible points within entropy 4–6 and perplexity at most 1,000.

H.8 CONSISTENCY GAP

We compare $\delta \in \{1/8, 1/16, 1/32, 1/64\}$ with the target refreshed every round, either keeping the pretrained transport denoiser fixed or refreshing it from the same student EMA. For TinyGSM, each curve in Figure 19 uses one sampler and temperature selected by mean accuracy across the six NFEs. For OWT, Figure 20 averages log-perplexity over the entropy interval shared by all eight curves at each NFE, then across NFEs; lower scores correspond to lower matched-entropy perplexity. The measured OWT temperature curves appear in Figures 21 to 24.

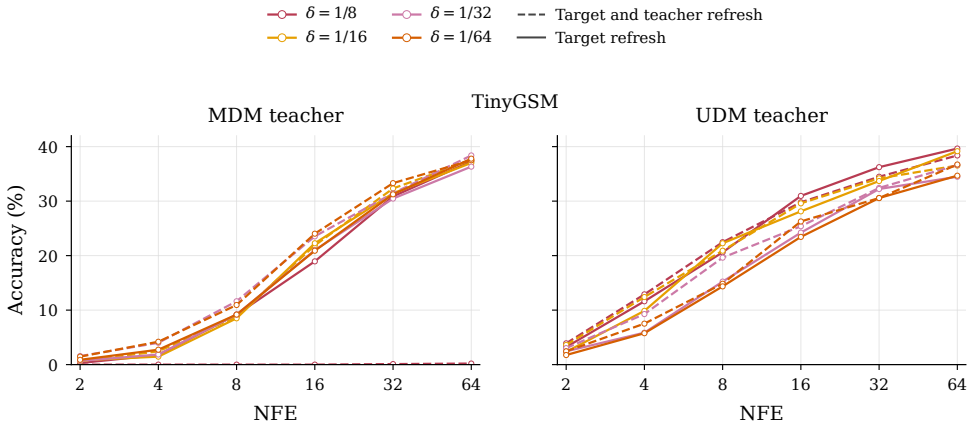

Figure 19: TinyGSM accuracy across training gaps and refresh policies. Solid curves refresh only the target; dashed curves also refresh the rollout teacher.

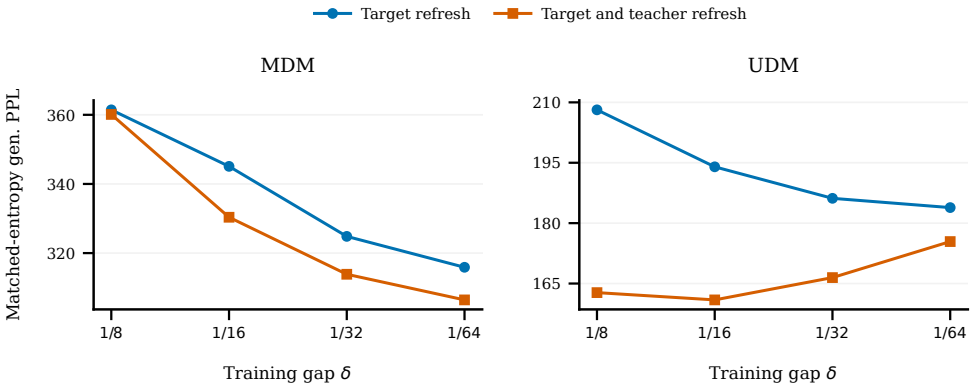

Figure 20: OWT matched-entropy generative perplexity, averaged across six NFEs, for each training gap and refresh policy.

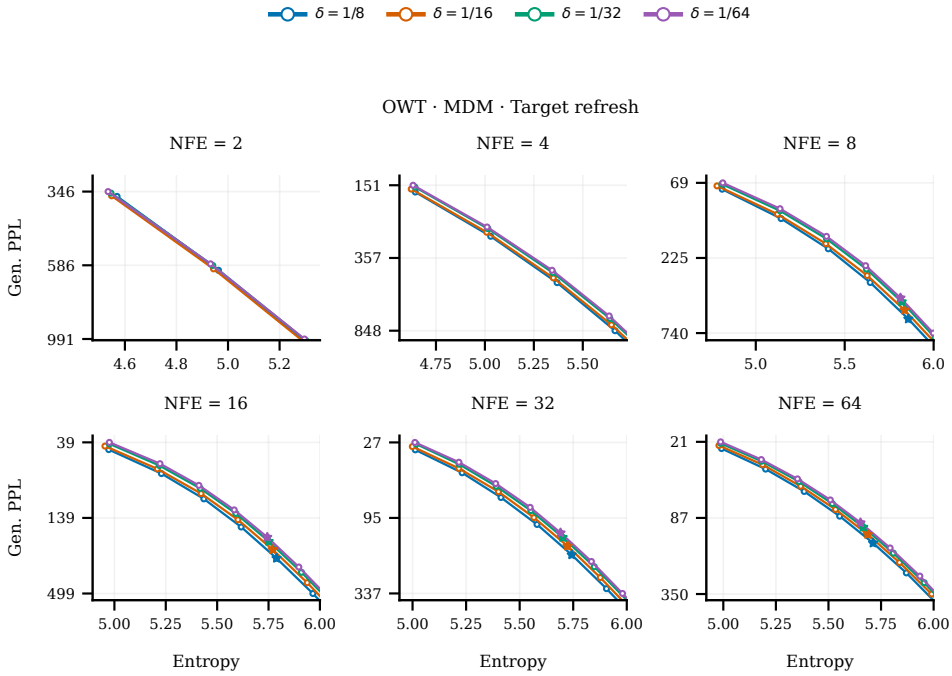

Figure 21: OWT MDM with target refresh: measured bridge-sampling temperature curves at each NFE. Stars mark temperature 1.

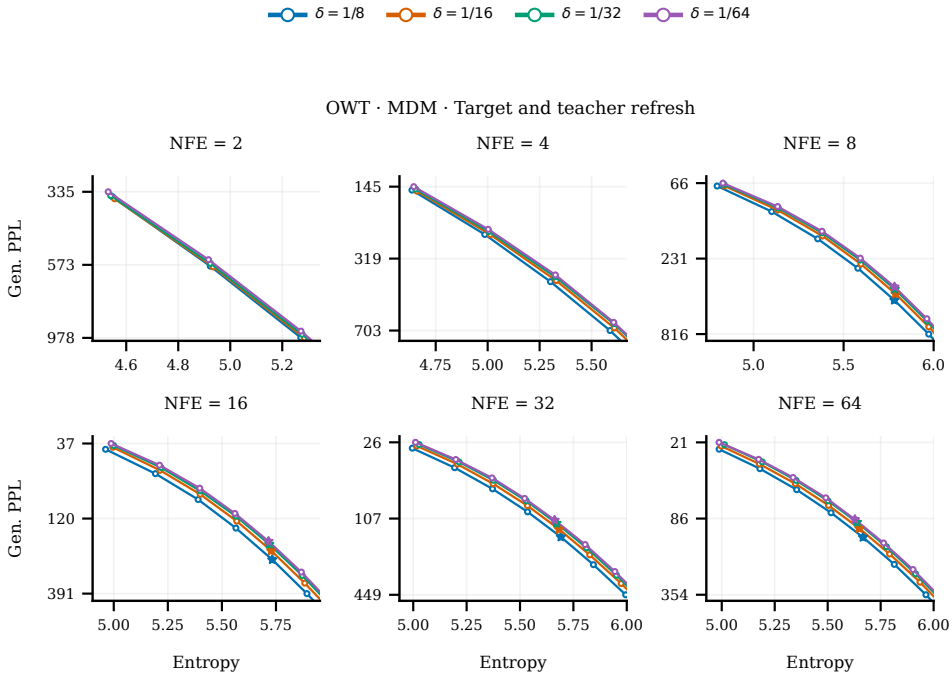

Figure 22: OWT MDM with joint target-and-teacher refresh, under the same evaluation grid as Figure 21.

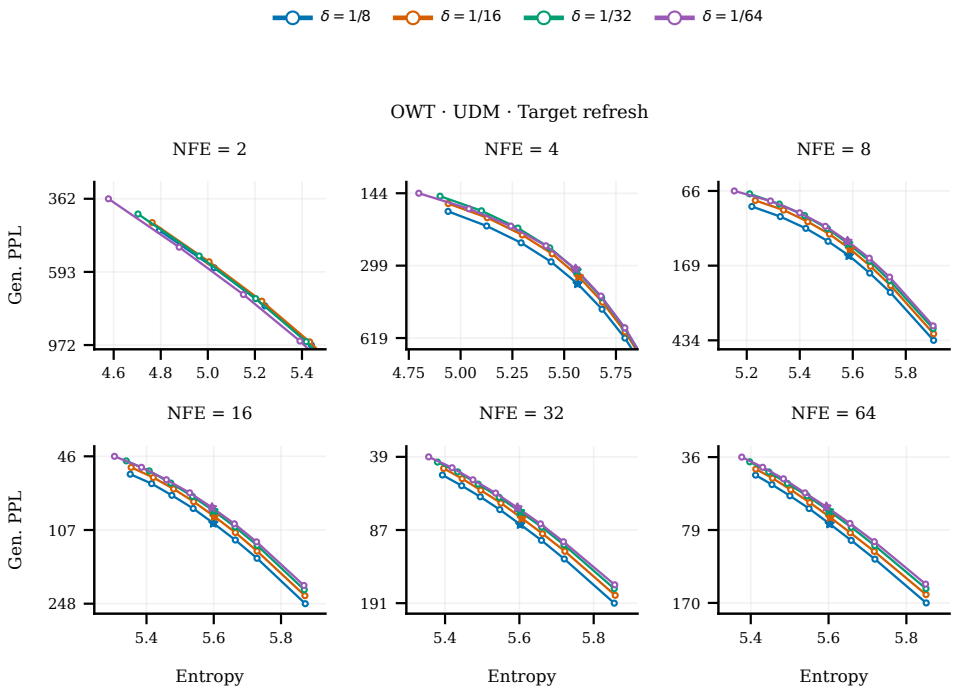

Figure 23: OWT UDM with target refresh. Axes show visible points within entropy 4–6 and perplexity at most 1,000.

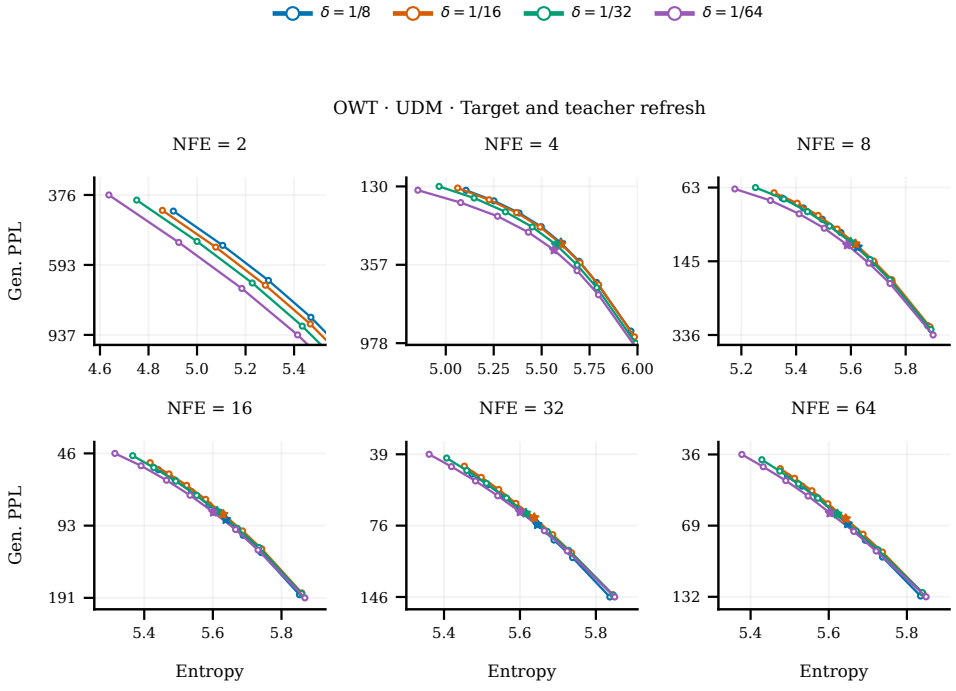

Figure 24: OWT UDM with joint target-and-teacher refresh, under the same evaluation grid as Figure 23.

H.9 BRIDGE AND REFRESH SAMPLING

We compare bridge sampling, which conditions the next state on the current state and the student's clean prediction, with refresh sampling, which re-noises that prediction and draws new residuals. For UDM, temperature is applied either to LOO probabilities before conversion or to denoising probabilities after conversion; MDM uses denoiser temperature. Each comparison includes the shared teacher, iDLM, D-MMD, SDTT for MDM or DCD for UDM, and ERC. ERC uses the selected gap in Appendix H.5, fixed for each dataset and diffusion family.

For TinyGSM, Figures 25 to 27 show accuracy against NFE with bridge and refresh in adjacent panels. Within each panel, one of the eight measured temperatures is selected separately for each method by its mean accuracy over the six NFE budgets; that temperature is printed in the legend and held fixed across NFEs. Figures 28 to 30 instead show all measured temperatures at every NFE for both samplers, with pointwise 95% question-bootstrap bands. All points use the same $1{,}024$ GSM8K questions.

For OpenWebText, each panel of Figures 31 to 36 fixes an NFE budget and sweeps the eight measured temperatures. Axes fit the visible points without extending below entropy 4, above entropy 6, or above generative perplexity $1{,}000$; points outside these bounds are not shown, and a star identifies temperature 1 when visible. The perplexity axis is logarithmic and inverted. All points use $2{,}048$ generated sequences. Denoiser and LOO placement at temperature 1 are aliases of one evaluation within each UDM sampler, not independent replicates.

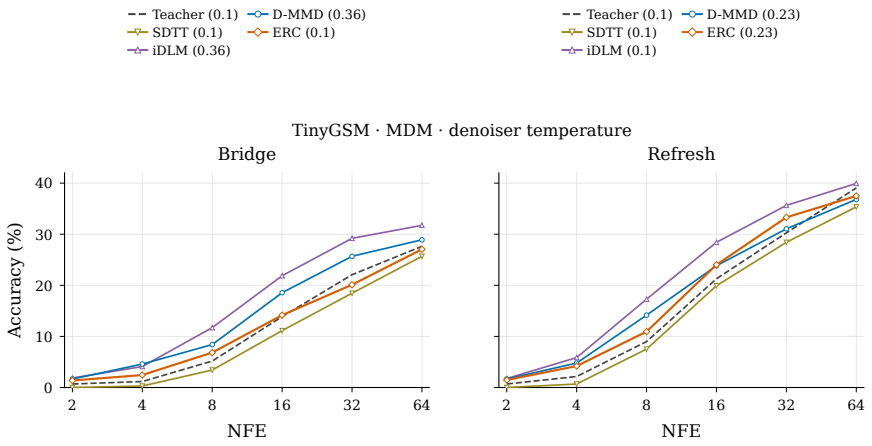

Figure 25: TinyGSM MDM accuracy versus NFE under bridge and refresh sampling. Each method retains the temperature selected by mean accuracy across six NFE budgets; the teacher, SDTT, iDLM, D-MMD, and ERC are included.

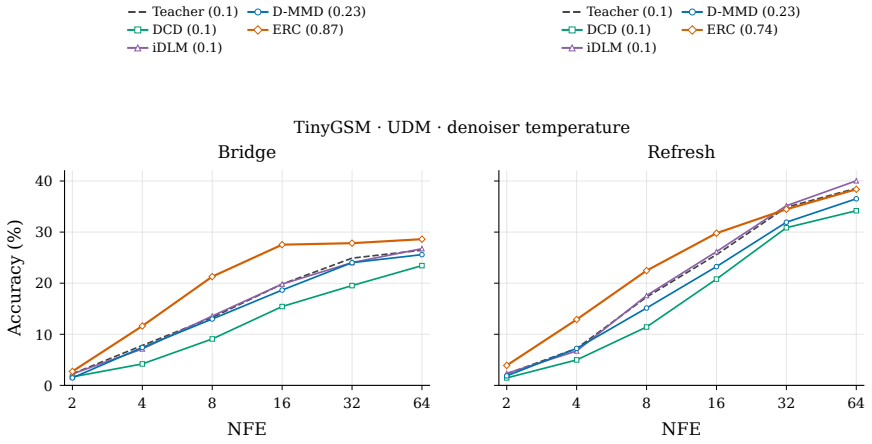

Figure 26: TinyGSM UDM accuracy versus NFE with temperature applied to denoiser predictions, comparing bridge and refresh sampling and all five methods.

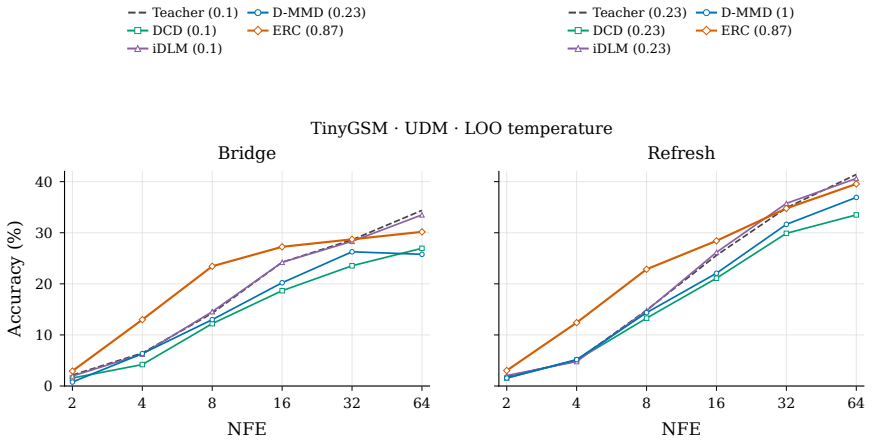

Figure 27: TinyGSM UDM accuracy versus NFE with temperature applied to LOO predictions. Bridge and refresh panels contain the teacher, iDLM, D-MMD, DCD, and ERC.

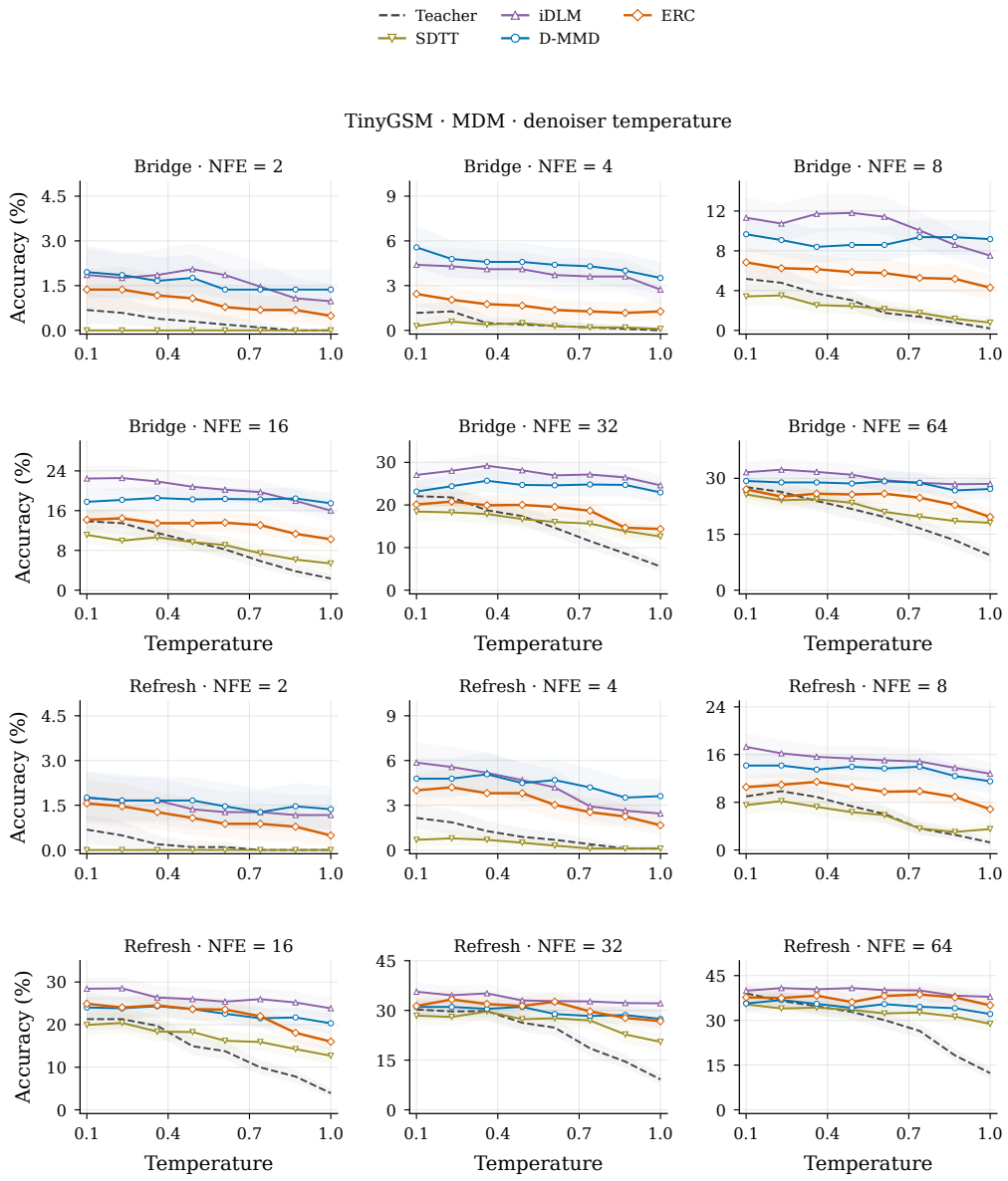

Figure 28: TinyGSM MDM accuracy across eight denoiser temperatures and six NFE budgets for bridge and refresh sampling. Shaded bands are pointwise 95% bootstrap intervals over questions.

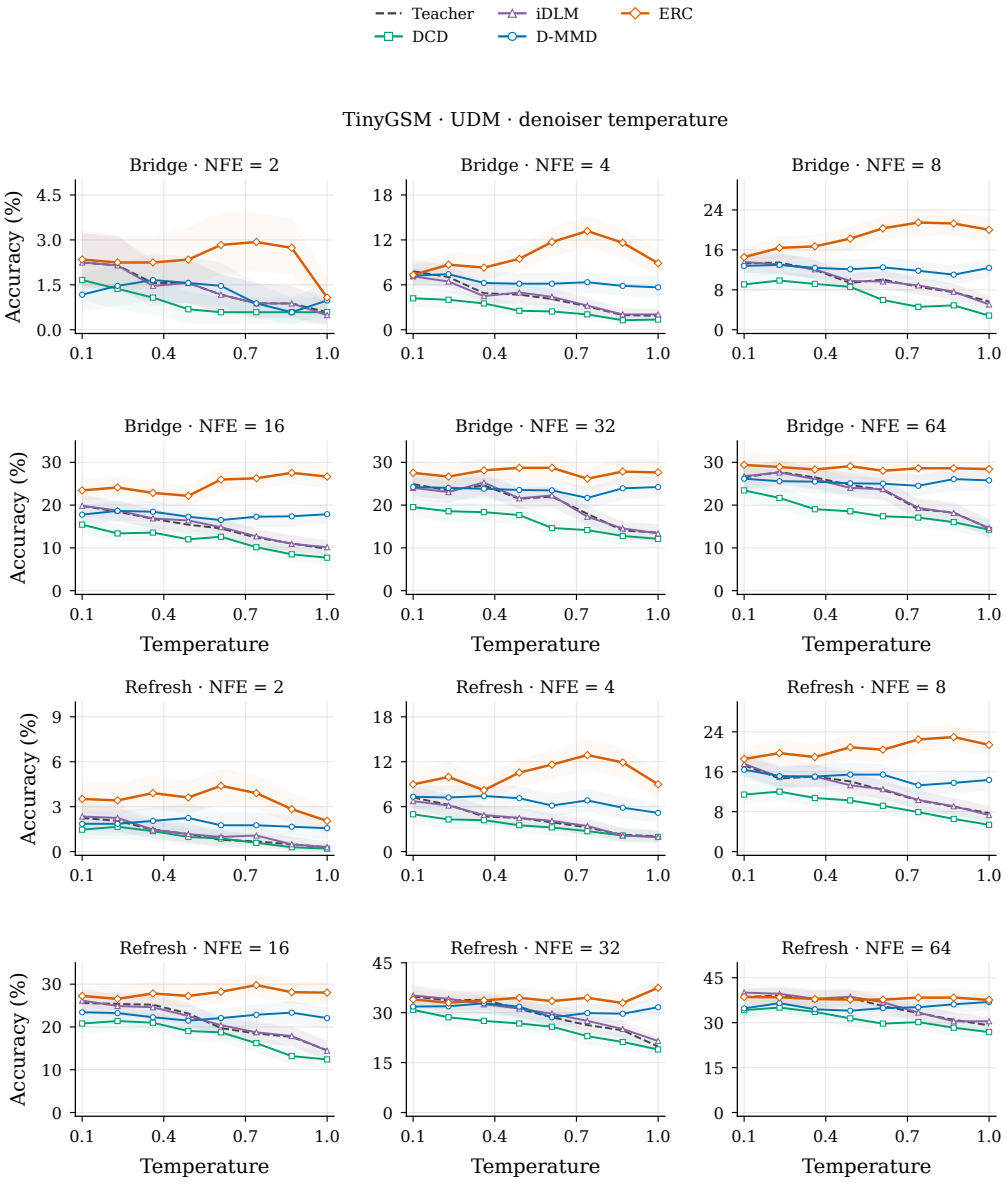

Figure 29: TinyGSM UDM denoiser-temperature sweeps for both samplers at all six NFE budgets. Bands give pointwise 95% question-bootstrap intervals.

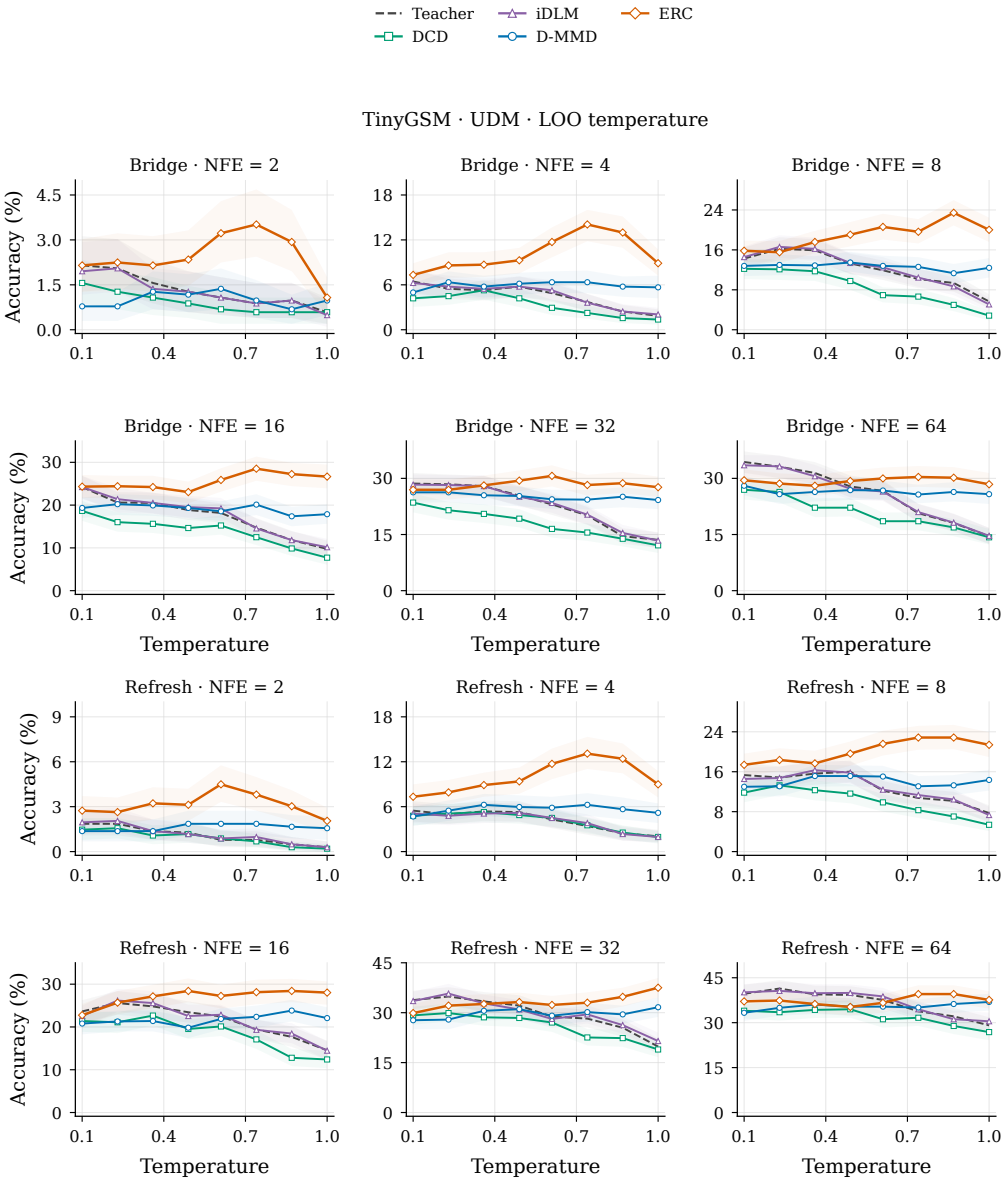

Figure 30: TinyGSM UDM LOO-temperature sweeps for bridge and refresh sampling. All eight temperatures and six NFE budgets are retained, with pointwise 95% question-bootstrap intervals.

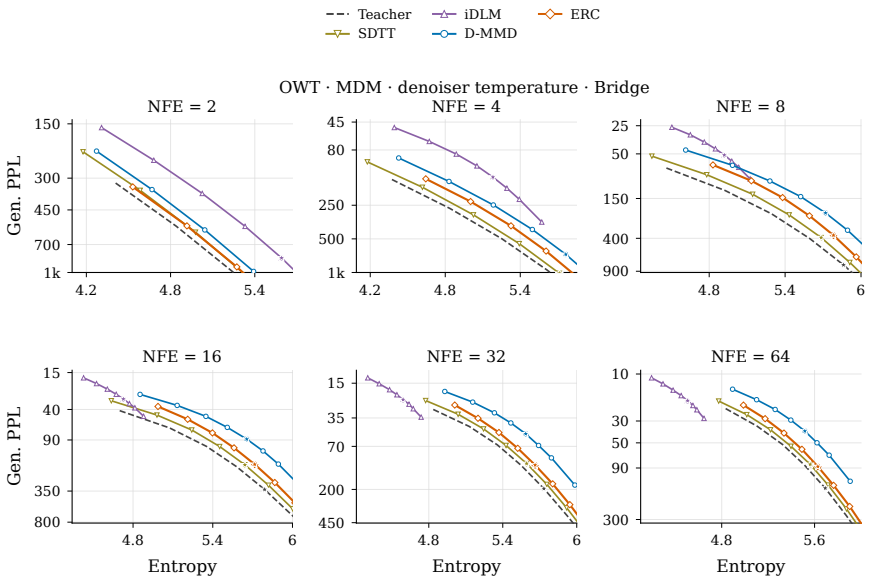

Figure 31: OWT MDM entropy–perplexity curves under bridge sampling.

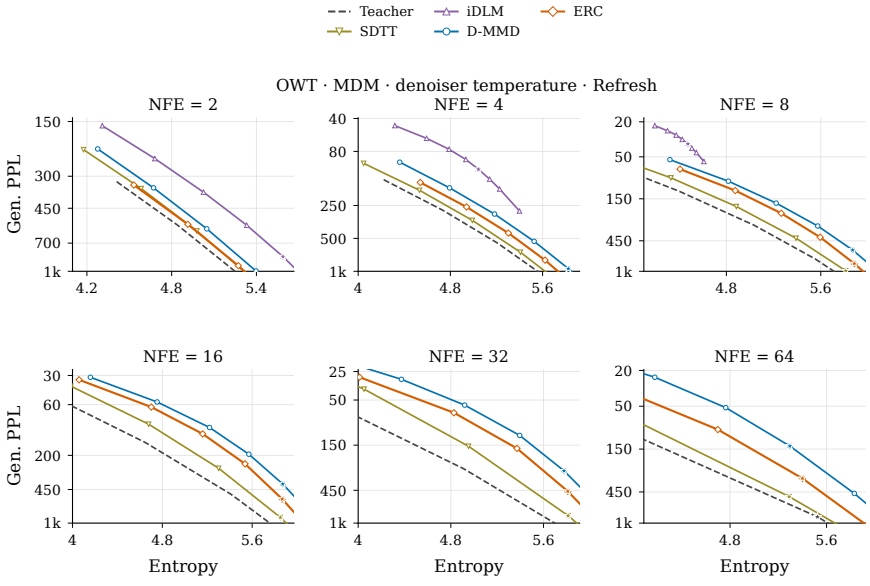

Figure 32: The same OWT MDM comparison under refresh sampling.

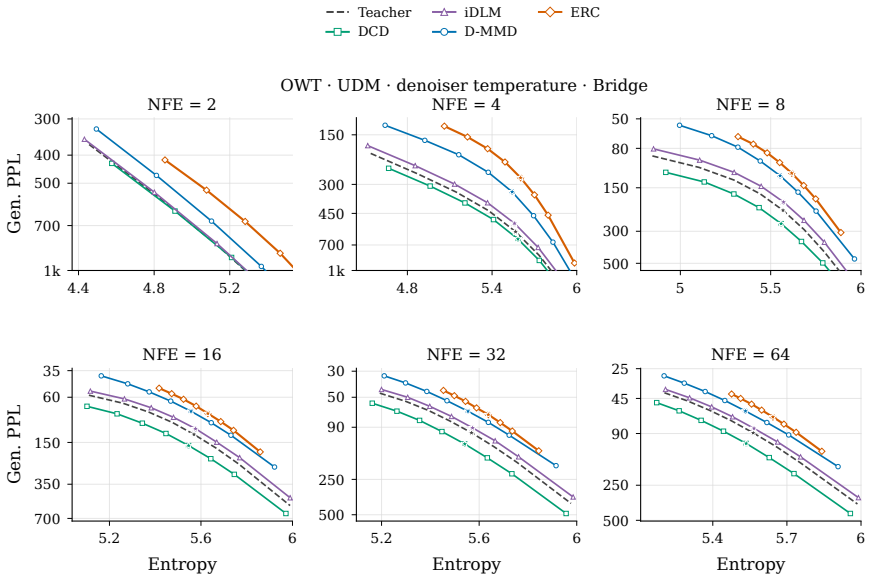

Figure 33: OWT UDM entropy–perplexity curves with denoiser temperature under bridge sampling.

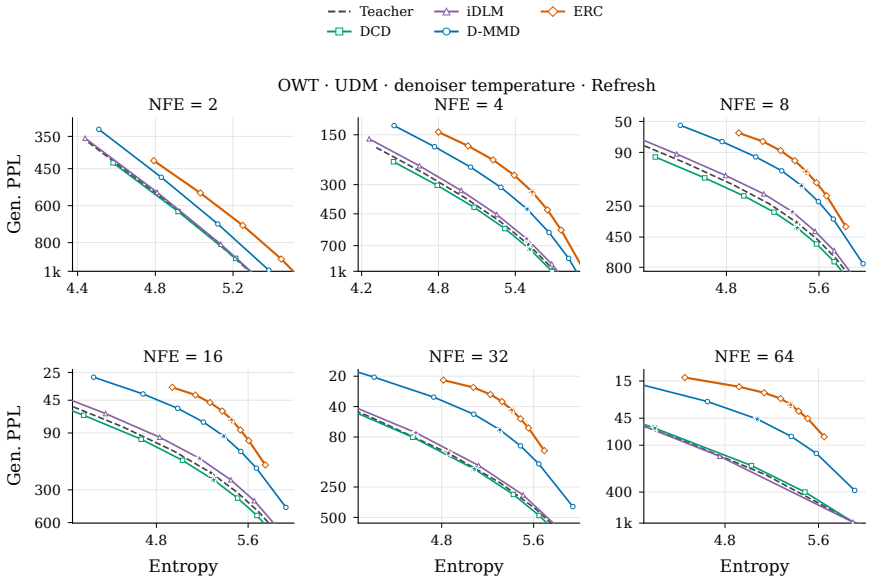

Figure 34: The same OWT UDM denoiser-temperature comparison under refresh sampling.

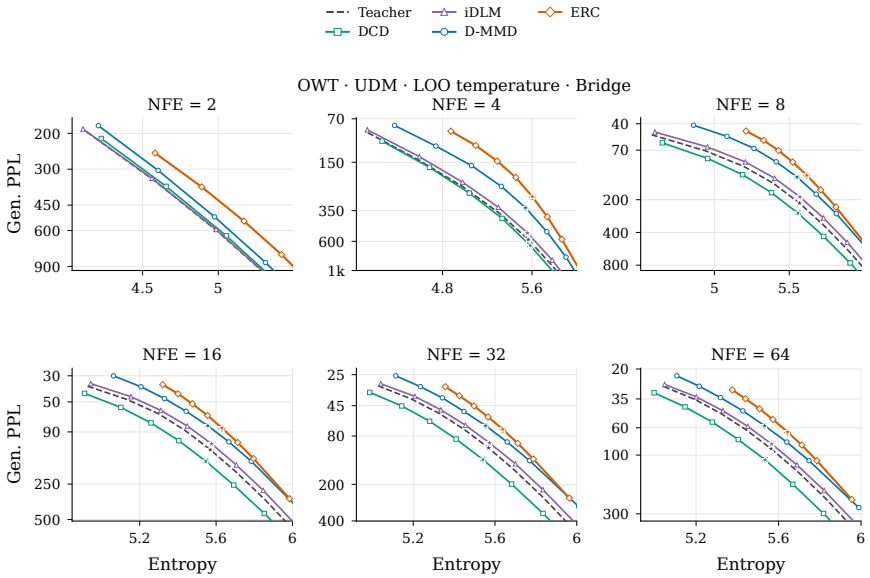

Figure 35: OWT UDM entropy–perplexity curves with LOO temperature under bridge sampling.

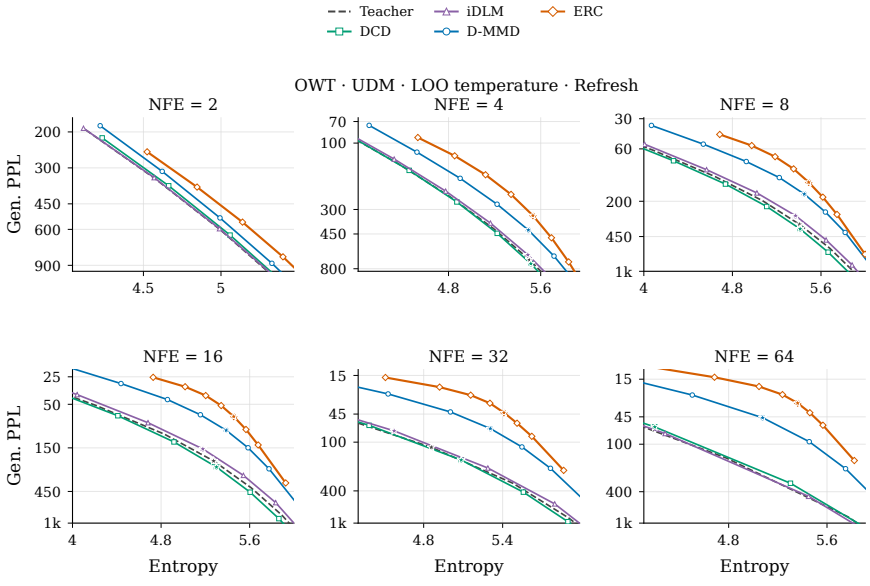

Figure 36: The same OWT UDM LOO-temperature comparison under refresh sampling.

H.10 GUMBEL CONDITIONING

We compare ERC with and without conditioning on the Gumbel variables $-\log E_t$ after 10,000 student updates. The matched MDM runs use a fixed training gap; the UDM runs increase the gap over four rounds. In both, the target and ERF transport denoiser are refreshed together. Removing Gumbel inputs does not change the ERF transport, including its net-flux rates for UDM. TinyGSM uses both samplers and averages sampling seeds 101 and 303; OWT uses bridge sampling with seed 101. No matched OWT refresh comparison is available.

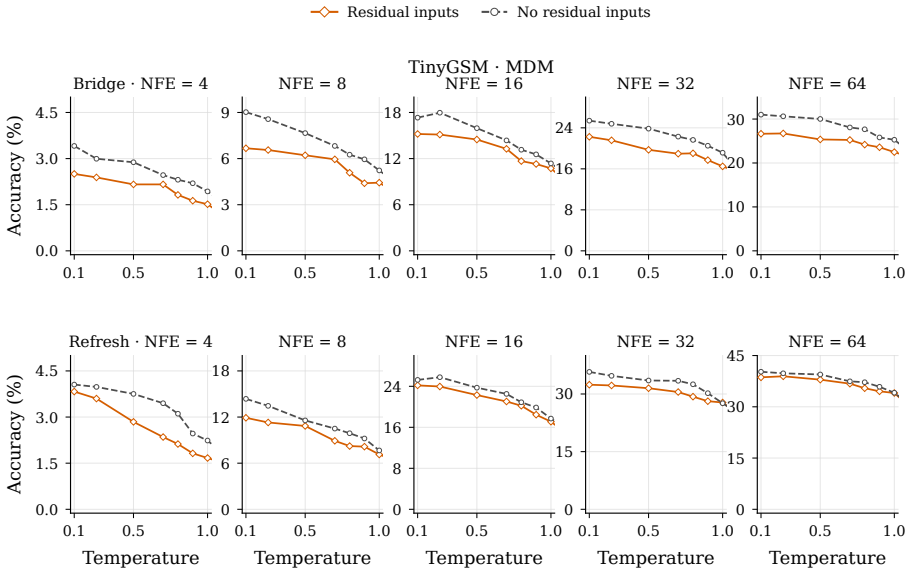

Figure 37: TinyGSM MDM with and without Gumbel inputs under bridge and refresh sampling. Points average two sampling seeds.

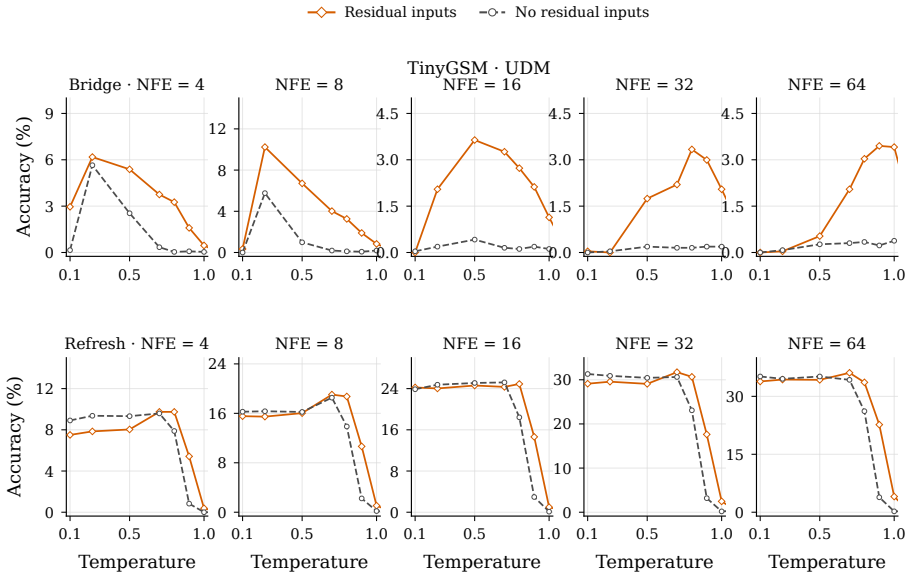

Figure 38: TinyGSM UDM under the same Gumbel-input comparison, with an increasing training gap.

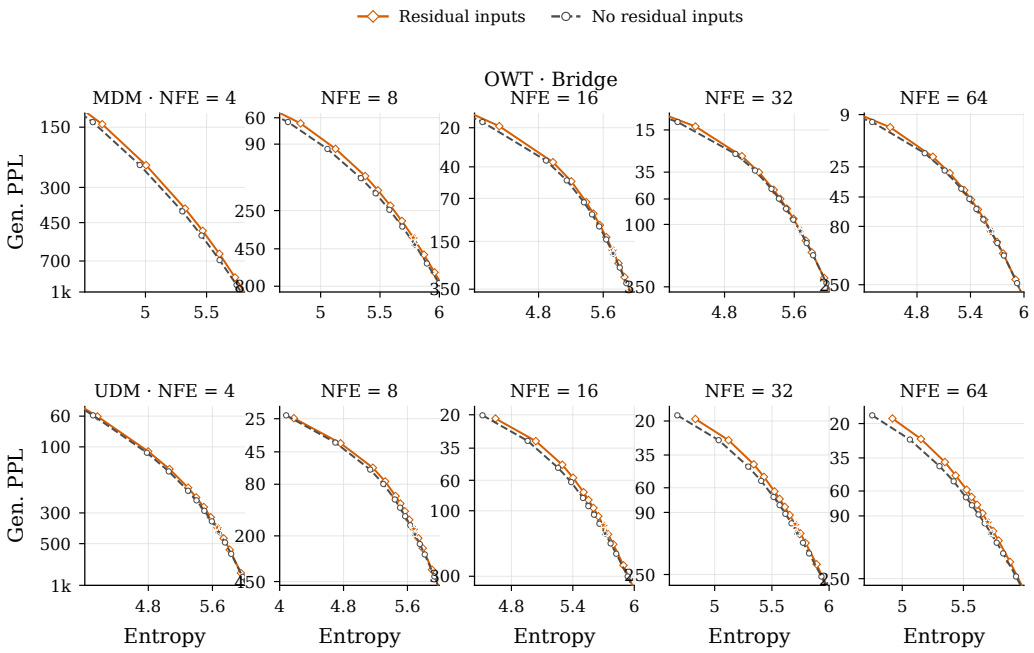

Figure 39: OWT Gumbel-input comparison for MDM and UDM under bridge sampling at five NFEs and fourteen temperatures.

I DiffusionGemma experiment

We evaluate the instruction-tuned 26B-parameter, 4B-active DiffusionGemma checkpoint (Google DeepMind, 2026) on the MMLU-Pro test split (Wang et al., 2024). We select ten questions from each of its fourteen categories by a stable hash with seed 20260820, yielding the same 140 questions for run seeds 1, 2, and 3. Each prompt requests a concise justification ending with `The answer is (X).`, and generation uses one canvas block of at most 256 tokens. The strict evaluator accepts only this final-answer form and counts malformed or truncated completions as incorrect.

We compare the official native generation procedure with UDLM and the unbounded ERF Uniform and ERF U-NetFlux samplers. The custom samplers use 48 NFEs, temperature 1, and temperature scaling of the raw denoiser logits. The accuracy intervals in Table 1 use 50,000 category-stratified bootstrap replicates over question clusters while retaining all three paired seeds. The native procedure terminates adaptively after 11.1 forward evaluations on average, whereas each custom sampler uses exactly 48; the reported timings therefore describe the runs but do not constitute a compute-matched comparison. This MMLU-Pro evaluation is not directly comparable to evaluations on other benchmarks reported with the model release.

J Proofs for the ERF sampler

J.1 Intuition on the ODEs (8)

To motivate the form of this ODE, we derive a continuity equation on the extended state space $\mathsf{X} \times [0,\infty)^{L\times K}$ such that X_t follows the Kolmogorov equation with $(Q_t)_{t\in[0,1]}$. This will impose a velocity on the exponential clocks. Rewriting the Kolmogorov equation (5) and using (6), we have for any x and t:

$$\partial_t p_t(x) = p_t(x) \sum_{\ell=1}^{L} \big(b_t^\ell(x) - \lambda_t^\ell(x^\ell)\big) \ , \tag{33}$$

where $b_t^\ell(x) := \sum_{j \in \mathsf{V}\setminus\{x^\ell\}} \bar{Q}_t^\ell(j|x)$ is the reverse exit rate at position ℓ, and $\lambda_t^\ell(x^\ell) := \sum_{j \in \mathsf{V}\setminus\{x^\ell\}} Q_t^\ell(j|x^\ell)$ is the forward exit rate. Therefore, introducing the vector field (depending on the discrete state): $v_t(x) \in \mathbb{R}^{L\times K}$ with components

$$v_t^{\ell j}(x) = \begin{cases} \bar{Q}_t^\ell(j|x), & j \neq x^\ell, \\ -\lambda_t^\ell(x^\ell), & j = x^\ell \,, \end{cases} \tag{34}$$

and the extended density $\mu_t(x,e) = p_t(x)\exp(-\sum_{\ell=1}^L \sum_{j=1}^K e^{\ell j})$ on $\mathsf{X} \times [0,\infty)^{L\times K}$, we can rewrite the Kolmogorov equation (33) as

$$\partial_t \mu_t(x,e) = -\nabla_e \cdot (v_t(x)\mu_t(x,e)) \,, \tag{35}$$

where $\nabla_e \cdot$ denotes the divergence operator with respect to e. This equation can be interpreted as a continuity equation on the extended state space, and suggests the dynamics we have imposed on E_t, namely (8).

J.2 Setting

J.2.1 Reference rates

Fix integers $L \geq 1$ and $K \geq 2$. In this proof, we use the one-hot representatives of the token indices and write $\mathsf{V} = \{e_1, \dots, e_K\}$, $\mathsf{X} = \mathsf{V}^L$, and $D = LK$. Vectors in $\mathbb{R}^D$ are indexed by $[\![1,L]\!] \times [\![1,K]\!]$ in lexicographic order. Empty products equal one. An integral with no free coordinates means evaluation; in nonnegative integrands, $0\cdot(+\infty)$ is defined to be zero. Position indices are ℓ, h, token indices are j, k, and n, m count jumps or partial sums.

The notation $Q_v(y|x)$ denotes a rate from x to y. Only changes at one position are allowed. If $y^{-\ell} = x^{-\ell}$ and $y^\ell = e_j \neq x^\ell$, then

$$Q_v(y|x) = Q_v^\ell(e_j|x^\ell). \tag{36}$$

All other off-diagonal entries are zero. Set

$$\lambda_v^\ell(x^\ell) = \sum_{j: e_j \neq x^\ell} Q_v^\ell(e_j|x^\ell), \qquad \lambda_v(x) = \sum_{\ell=1}^L \lambda_v^\ell(x^\ell), \qquad Q_v(x|x) = -\lambda_v(x). \tag{37}$$

Assumption 1. For every position ℓ and distinct tokens $z, z' \in \mathsf{V}$, the map $v \mapsto Q_v^\ell(z'|z)$ is nonnegative and continuous on $[0,1]$. For every $x \in \mathsf{X}$, the map $v \mapsto p_v(x)$ is C^1 on $[0,1]$. For each $v \in [0,1]$, $p_v(x) > 0$ for every x, $\sum_x p_v(x) = 1$, and

$$\partial_v p_v(x) = \sum_{\ell=1}^L \sum_{\substack{y^{-\ell} = x^{-\ell} \\ y^\ell \neq x^\ell}} p_v(y) Q_v^\ell(x^\ell|y^\ell) - p_v(x)\lambda_v(x). \tag{38}$$

The derivatives at 0 and 1 are one-sided.

For $v \in [0,1]$ and distinct $x, y \in \mathsf{X}$, define the reverse rates by

$$\bar{Q}_v(y|x) = \frac{p_v(y)}{p_v(x)} Q_v(x|y). \tag{39}$$

In particular, if $y^{-\ell} = x^{-\ell}$ and $y^\ell = e_j \neq x^\ell$, write

$$\bar{Q}_v^\ell(e_j|x) = \bar{Q}_v(y|x) = \frac{p_v(y)}{p_v(x)} Q_v^\ell(x^\ell|e_j). \tag{40}$$

Write $\bar{\lambda}_v(x) = \sum_{\ell=1}^L \sum_{j: e_j \neq x^\ell} \bar{Q}_v^\ell(e_j|x)$ and $\bar{Q}_v(x|x) = -\bar{\lambda}_v(x)$. Individual rates and position exit rates may vanish. Reverse rates may depend on the entire configuration, and no product assumption is imposed on p_v.

J.2.2 Residual dynamics

The augmented state is $(X_v, E_v) \in \mathsf{X} \times [0,\infty)^D$, with $E_v = (E_v^{\ell j})_{\ell,j}$. Time runs backwards. For $x \in \mathsf{X}$ and $v \in [0,1]$, define the signed velocities in elapsed time by

$$b_v^{\ell j}(x) = \begin{cases} -\bar{Q}_v^{\ell}(e_j|x), & e_j \neq x^{\ell}, \\ \lambda_v^{\ell}(x^{\ell}), & e_j = x^{\ell}. \end{cases} \tag{41}$$

For $0 \le u \le v \le 1$, put

$$\Delta_{u|v}(x) = \int_u^v b_\tau(x)\,\mathrm{d}\tau. \tag{42}$$

If the state remains x from v down to u, then

$$E_u = E_v + \Delta_{u|v}(x). \tag{43}$$

A residual (ℓ, j) is active at x if $e_j \neq x^{\ell}$. When it first hits zero, the state changes to the configuration y with $y^{-\ell} = x^{-\ell}$ and $y^{\ell} = e_j$, and every residual remains continuous. The triggering residual is not resampled: it becomes passive with value zero.

Here is a finite recursive definition. Set $T_0 = 1$, $\Xi_0 = X_1$, and $W_0 = E_1$. Suppose n steps have been performed, with $(T_n, \Xi_n, W_n) = (v, x, \eta)$, $v > 0$, and all active components of η positive. Define

$$\theta_{\ell,j}(v,x,\eta) = \sup\left\{u \in [0,v] : \int_u^v \bar{Q}_\tau^{\ell}(e_j|x)\,\mathrm{d}\tau \ge \eta^{\ell j}\right\}, \qquad e_j \neq x^{\ell}, \tag{44}$$

with $\sup \varnothing = -\infty$, and let Θ_{n+1} be the largest candidate. If $\Theta_{n+1} = -\infty$, extend the path to 0 without another change. If

$$0 < \Theta_{n+1} < v, \qquad \operatorname*{arg\,max}_{\ell,j:e_j \neq x^{\ell}} \theta_{\ell,j}(v,x,\eta) = \{(h,k)\}, \qquad \bar{Q}^h_{\Theta_{n+1}}(e_k|x) > 0, \tag{45}$$

set

$$\begin{aligned} T_{n+1} &= \Theta_{n+1}, \qquad C_{n+1} = (h,k), \qquad Y_{n+1} = x, \\ W_{n+1} &= \eta + \Delta_{T_{n+1}|v}(x), \qquad \Xi_{n+1} = y, \quad y^{-h} = x^{-h}, \quad y^h = e_k. \end{aligned} \tag{46}$$

Here Y_{n+1} is the state before the change and C_{n+1} is the residual index. At any other finite hit, stop and record a failure, including a hit at time 0. On the segment before the next hit, set

$$X_u = \Xi_n, \qquad E_u = W_n + \Delta_{u|T_n}(\Xi_n), \qquad \Theta_{n+1} < u \le T_n. \tag{47}$$

At T_{n+1} the state is Ξ_{n+1} and the residual vector is W_{n+1}, so X is right-continuous in elapsed time $1-u$. If a successful step does not exist, set its time and every subsequent step time to $-\infty$; all other variables after stopping receive fixed values. Absence of a hit and failure are recorded separately. We shall prove that failures and infinite sequences of hits have probability zero under the initialization below.

J.3 Theorem

Let $\mathcal{O} = (0,\infty)^D$. For $r \in [0,\infty)^D$ and $u \in [0,1]$, put

$$\gamma(r) = \exp\left(-\sum_{\ell,j} r^{\ell j}\right), \qquad F_u(x,r) = p_u(x)\gamma(r). \tag{48}$$

Lebesgue measure on $\mathcal{O}$ is denoted by $\mathrm{d}r$.

Theorem J.1. *Assume Assumption 1, define the reverse rates by (39)–(40), and use the dynamics (41)–(47). Suppose*

$$\mathcal{L}(X_1, E_1) = p_1 \otimes \mathrm{Exp}(1)^{\otimes D}. \tag{49}$$

Then the dynamics is almost surely well defined on $[0,1]$, with finitely many jumps. Every jump satisfies (45), and the probability of a jump at any fixed $u \in [0,1]$ is zero. For every deterministic

$u \in [0,1]$,

$$\mathcal{L}(X_u, E_u) = p_u \otimes \mathrm{Exp}(1)^{\otimes D}. \tag{50}$$

In particular, for every $x \in \mathsf{X}$ *and* $\nu \in [0,\infty)^D$,

$$\mathbb{P}\big(X_u = x,\ E_u^{\ell j} \le \nu^{\ell j} \text{ for all } (\ell, j)\big) = p_u(x) \prod_{\ell, j} \left(1 - \exp(-\nu^{\ell j})\right). \tag{51}$$

For $n \ge 0$ and $u \in [0,1)$, define

$$H_u^{(n)} = \{\text{the first } n \text{ steps satisfy (45)},\ T_n > u,\ \Theta_{n+1} < u\}. \tag{52}$$

For $n = 0$, the first condition is vacuous. The candidate Θ_{n+1} is computed before testing (45), so (52) requires only a finite construction. An integrand restricted by an event indicator is evaluated on that event and defined to be zero elsewhere. For each Borel $B \subseteq [0,\infty)^D$, set

$$\mu_u^{(n)}(\{x\} \times B) = \mathbb{E}\Big[\mathbb{1}_{H_u^{(n)}} \mathbb{1}_{\{\Xi_n = x\}} \mathbb{1}_{\{W_n + \Delta_{u|T_n}(\Xi_n) \in B\}}\Big]. \tag{53}$$

We prove the theorem by constructing densities $f_u^{(n)}$ of these measures and summing them. Initialization (49) is not conditioning on a fixed state. The conclusion concerns deterministic times; it imposes no independence at jump times and does not assert that X is Markovian.

J.4 Preliminaries

For $x \in \mathsf{X}$, $\ell \in [\![1, L]\!]$, $e_j \ne x^\ell$, and $0 \le u \le v \le 1$, define

$$A_{u|v}^{\ell}(x) = \int_u^v \lambda_\tau^\ell(x^\ell)\, \mathrm{d}\tau, \qquad B_{u|v}^{\ell j}(x) = \int_u^v \bar{Q}_\tau^\ell(e_j|x)\, \mathrm{d}\tau. \tag{54}$$

Consequently,

$$\Delta_{u|v}^{\ell j}(x) = \begin{cases} A_{u|v}^{\ell}(x), & e_j = x^\ell, \\ -B_{u|v}^{\ell j}(x), & e_j \ne x^\ell. \end{cases} \tag{55}$$

Lemma 1. *Under Assumption 1, all forward and reverse off-diagonal rates and all exit rates are continuous and bounded on* $[0,1]$. *For* $v \in [0,1]$, $x \in \mathsf{X}$, *a position* ℓ, *and* y *with* $y^{-\ell} = x^{-\ell}$ *and* $y^\ell \ne x^\ell$,

$$\partial_v \log p_v(x) = \bar{\lambda}_v(x) - \lambda_v(x), \tag{56}$$

$$p_v(y)\bar{Q}_v^\ell(x^\ell|y) = p_v(x) Q_v^\ell(y^\ell|x^\ell), \tag{57}$$

$$\sum_{\substack{y^{-\ell} = x^{-\ell} \\ y^\ell \ne x^\ell}} p_v(y)\bar{Q}_v^\ell(x^\ell|y) = p_v(x)\lambda_v^\ell(x^\ell). \tag{58}$$

In (56), *endpoint derivatives are one-sided. In particular,*

$$\bar{Q}_v^\ell(x^\ell|y) > 0 \quad \Longrightarrow \quad \lambda_v^\ell(x^\ell) > 0. \tag{59}$$

For $0 \le u \le v \le 1$, $x \in \mathsf{X}$, *and* $r, w \in [0,\infty)^D$ *with* $r = w + \Delta_{u|v}(x)$,

$$p_v(x)\gamma(w) = p_u(x)\gamma(r). \tag{60}$$

Proof. Since X is finite and $p_v(x)$ is continuous and positive on $[0,1]$, the number $p_* = \min_{v \in [0,1], x \in \mathsf{X}} p_v(x)$ is positive. Boundedness and continuity of the forward rates, of (40), and of their finite sums follow. Applying (40) to the transition from y into x gives

$$\bar{Q}_v^\ell(x^\ell|y) = \frac{p_v(x)}{p_v(y)} Q_v^\ell(y^\ell|x^\ell).$$

This proves (57); summing over the predecessor states gives (58). Nonnegativity of the summands and positivity of $p_v(x)$ imply (59). Dividing (38) by $p_v(x)$ and using (40) proves (56).

For $0 \le u \le v \le 1$, equations (41) and (56) give

$$\begin{aligned}\sum_{\ell,j} \Delta_{u|v}^{\ell j}(x) &= \int_u^v \big(\lambda_\tau(x) - \bar{\lambda}_\tau(x)\big)\,\mathrm{d}\tau \\ &= -\int_u^v \partial_\tau \log p_\tau(x)\,\mathrm{d}\tau = \log \frac{p_u(x)}{p_v(x)}.\end{aligned} \tag{61}$$

Since $w = r - \Delta_{u|v}(x)$, substitution in (48) yields

$$p_v(x)\gamma(w) = p_v(x)\gamma(r)\exp\left(\sum_{\ell,j} \Delta_{u|v}^{\ell j}(x)\right) = p_u(x)\gamma(r).$$

□

Lemma 2. *Let $\alpha < \beta$ be finite, let $\omega : [\alpha, \beta] \to [0, \infty)$ be continuous, and set $\Psi(v) = \int_\alpha^v \omega(\tau)\,\mathrm{d}\tau$ and $\rho(\zeta) = \inf\{v \in [\alpha, \beta] : \Psi(v) \ge \zeta\}$ for $0 < \zeta < \Psi(\beta)$.*

(i) *The set $\Psi(\{v \in [\alpha, \beta] : \omega(v) = 0\})$ has one-dimensional Lebesgue measure zero. For almost every $\zeta \in (0, \Psi(\beta))$, the equation $\Psi(v) = \zeta$ has exactly one solution; it lies in (α, β) and has $\omega(v) > 0$. More generally, for every $v_0 \in (\alpha, \beta)$ with $\omega(v_0) > 0$, one has $\Psi^{-1}(\{\Psi(v_0)\}) = \{v_0\}$.*

(ii) *For every nonnegative Borel χ on $[0, \Psi(\beta)]$,*

$$\int_\alpha^\beta \chi(\Psi(v))\omega(v)\,\mathrm{d}v = \int_0^{\Psi(\beta)} \chi(\zeta)\,\mathrm{d}\zeta. \tag{62}$$

For every nonnegative Borel φ on $[\alpha, \beta]$,

$$\int_\alpha^\beta \varphi(v)\omega(v)\,\mathrm{d}v = \int_0^{\Psi(\beta)} \varphi(\rho(\zeta))\,\mathrm{d}\zeta. \tag{63}$$

(iii) *If $v_1, v_2 : [\alpha, \beta] \to \mathbb{R}$ are Lipschitz, then $\{(v_1(v), v_2(v)) : v \in [\alpha, \beta]\}$ has two-dimensional Lebesgue measure zero. An integral over $(0, \Psi(\beta))$ is zero when $\Psi(\beta) = 0$.*

Proof. Fix $\epsilon > 0$. The relatively open set $O_\epsilon = \{v \in [\alpha, \beta] : \omega(v) < \epsilon\}$ is a countable disjoint union of intervals. On each component interval J, monotonicity and the integral expression for Ψ imply $\mathrm{length}(\Psi(J)) \le \epsilon\,\mathrm{length}(J)$. Their images, together with at most their endpoint images, cover the image of the zero set of ω. Countable subadditivity bounds its outer measure by $\epsilon(\beta - \alpha)$. Letting ϵ tend to zero proves nullity. Exclude these image values and $0, \Psi(\beta)$. The intermediate value theorem gives a solution $v \in (\alpha, \beta)$ of $\Psi(v) = \zeta$, and this solution has $\omega(v) > 0$. If $v_1 < v_2$ had the same image, monotonicity would make Ψ constant on $[v_1, v_2]$; continuity and nonnegativity would give $\omega(v_1) = \omega(v_2) = 0$, a contradiction. The same argument proves the last assertion whenever $\omega(v_0) > 0$, without an exceptional set. This proves (i).

For continuous χ, let $\Phi(\zeta) = \int_0^\zeta \chi(\zeta')\,\mathrm{d}\zeta'$. The chain rule gives $(\Phi \circ \Psi)'(v) = \chi(\Psi(v))\omega(v)$. Integration proves (62). Equality against continuous functions identifies the two finite measures, so the identity extends to nonnegative Borel χ. For $\alpha < v < \beta$ with $\omega(v) > 0$, monotonicity gives $\rho(\Psi(v)) = v$. Applying (62) to $\chi(\zeta) = \varphi(\rho(\zeta))$, with arbitrary values at the endpoints and exceptional ζ, proves (63).

For (iii), let Λ bound both Lipschitz constants. Divide $[\alpha, \beta]$ into m intervals of length $(\beta - \alpha)/m$. Each image lies in a square of side at most $2\Lambda(\beta - \alpha)/m$. The total area of these m squares is at most $4\Lambda^2(\beta - \alpha)^2/m$. Letting $m \to \infty$ proves the assertion. □

For a numerical pair $c \in [\![1, L]\!] \times [\![1, K]\!]$, define

$$\mathcal{B}_c = \{r \in [0, \infty)^D : r^c = 0,\ r^d > 0 \text{ for every } d \ne c\}. \tag{64}$$

Let $\iota_c : (0,\infty)^{D-1} \to \mathcal{B}_c$ insert zero in coordinate c, using lexicographic order for the others. The measure σ_c on $\mathcal{B}_c$ is defined by

$$\int_{\mathcal{B}_c} \varphi(w)\, \sigma_c(\mathrm{d}w) = \int_{(0,\infty)^{D-1}} \varphi(\iota_c(\zeta))\, \mathrm{d}\zeta \tag{65}$$

for nonnegative Borel φ; equivalently $\sigma_c(\mathrm{d}w) = \delta_0(\mathrm{d}w^c) \prod_{d \neq c} \mathrm{d}w^d$ on $\mathcal{B}_c$. For $x \in \mathsf{X}$, let

$$c_\ell(x) = \text{the unique pair } (\ell, j) \text{ with } e_j = x^\ell, \qquad \mathcal{O}_x = \mathcal{O} \cup \bigcup_{\ell, j : e_j \neq x^\ell} \mathcal{B}_{(\ell,j)}. \tag{66}$$

Thus $r^{c_\ell(x)} = \langle x^\ell, r^\ell \rangle$. For $u \in [0,1)$, $x \in \mathsf{X}$, and $r \in \mathcal{O}_x$, define

$$\mathcal{N}_u(x) = \{ r \in \mathcal{O}_x : \langle x^h, r^h \rangle > A^h_{u|1}(x) \text{ for every } h \}, \tag{67}$$

$$\rho_\ell(u, x, r) = \inf\{ v \in [u,1] : A^\ell_{u|v}(x) \geq \langle x^\ell, r^\ell \rangle \}, \qquad \inf \varnothing = +\infty. \tag{68}$$

Define also

$$\begin{aligned} \mathcal{R}^\ell_u(x) = \big\{ r \in \mathcal{O}_x : \exists v \in (u,1) \text{ such that } \quad & \lambda^\ell_v(x^\ell) > 0, \\ & \langle x^\ell, r^\ell \rangle = A^\ell_{u|v}(x), \\ & \langle x^h, r^h \rangle > A^h_{u|v}(x) \quad (h \neq \ell), \\ & r^d - \Delta^d_{u|v}(x) > 0 \quad (d \neq c_\ell(x)) \big\}, \end{aligned} \tag{69}$$

and

$$\mathcal{R}_u(x) = \bigcup_{\ell=1}^{L} \mathcal{R}^\ell_u(x). \tag{70}$$

For $r \in \mathcal{R}^\ell_u(x)$, use the notation

$$v = \rho_\ell(u, x, r), \qquad c = c_\ell(x), \qquad w = r - \Delta_{u|v}(x) \in \mathcal{B}_c. \tag{71}$$

Lemma 3. *Assume Assumption 1.*

(i) *The maps in (68) are Borel in (u, r) for fixed x. The regions in (67) and (69), as sets of (u, r), are Borel. The sets $\mathcal{R}^\ell_u(x)$ are pairwise disjoint and disjoint from $\mathcal{N}_u(x)$.*

(ii) *For each fixed $u \in [0,1)$ and $x \in \mathsf{X}$,*

$$\mathrm{Leb}_D\big(\mathcal{O} \setminus (\mathcal{N}_u(x) \cup \mathcal{R}_u(x))\big) = 0. \tag{72}$$

For each such u, x and $d = (h, j)$ with $e_j \neq x^h$ and $\bar{Q}^h_u(e_j|x) > 0$,

$$\sigma_d\big(\mathcal{B}_d \setminus (\mathcal{N}_u(x) \cup \mathcal{R}_u(x))\big) = 0. \tag{73}$$

Equivalently, the exceptional set is null for Lebesgue measure in the $D-1$ coordinates other than d. In particular, for each fixed x and active $d = (h, j)$,

$$\int_0^1 \int_{\mathcal{B}_d} \mathbb{1}_{\{r \notin \mathcal{N}_u(x) \cup \mathcal{R}_u(x)\}} \bar{Q}^h_u(e_j|x)\, \sigma_d(\mathrm{d}r)\, \mathrm{d}u = 0. \tag{74}$$

(iii) *Fix $0 \leq u < v < 1$, $x \in \mathsf{X}$, a position ℓ, and $w \in \mathcal{B}_{c_\ell(x)}$ with $\lambda^\ell_v(x^\ell) > 0$. Set $r = w + \Delta_{u|v}(x)$. If $r \in \mathcal{O}$, there is no active hit on $[u, v)$, and $r \in \mathcal{R}^\ell_u(x)$ with $\rho_\ell(u, x, r) = v$. The same reconstruction conclusion holds if $r \in \mathcal{B}_d$ for an active $d = (h, j)$ with $\bar{Q}^h_u(e_j|x) > 0$; in this case the first active hit after leaving v is at u through d.*

(iv) *Conversely, if $r \in \mathcal{R}^\ell_u(x)$, then (71) gives the only pair (v, ℓ) for which a segment in state x, starting with only residual $c_\ell(x)$ zero and positive initial passive speed, can end at (u, x, r) without an intervening active hit. For $r \in \mathcal{B}_d$, this assertion is restricted to $\bar{Q}^h_u(e_j|x) > 0$, where $d = (h, j)$.*

Proof. For $\tau \in [0,1]$, on the domain $r \in \mathcal{O}_x$,

$$\{(u, r) : \rho_\ell(u, x, r) \leq \tau\} = \{(u, r) : u \leq \tau,\ \langle x^\ell, r^\ell \rangle \leq A^\ell_{u|\tau}(x)\}.$$

The right-hand side is Borel because the integrated rate is continuous. If a witness v exists in (69), positivity of $\lambda_v^\ell(x^\ell)$ and Lemma 2(i) imply that it is unique and equals $\rho_\ell(u,x,r)$. The other passive inequalities imply $\rho_h(u,x,r) > v$ for $h \neq \ell$. Membership in the region is therefore tested on the Borel set $\{u < \rho_\ell(u,x,r) < 1\}$ by substituting this Borel function into the remaining equalities and inequalities. This proves the Borel assertions and disjointness across positions. A finite root is incompatible with all the inequalities defining $\mathcal{N}_u(x)$.

Fix u, x. For each ℓ, exclude the values

$$\{0, A_{u|1}^\ell(x)\} \cup \{A_{u|v}^\ell(x) : v \in [u,1],\ \lambda_v^\ell(x^\ell) = 0\}$$

for the coordinate $\langle x^\ell, r^\ell\rangle$. These sets have one-dimensional measure zero by Lemma 2. For $\ell \neq h$, also exclude pairs

$$\big(\langle x^\ell, r^\ell\rangle, \langle x^h, r^h\rangle\big) \in \{(A_{u|v}^\ell(x), A_{u|v}^h(x)) : v \in [u,1]\}.$$

Their planar measure is zero by Lemma 2(iii). Fubini and the finite number of indices show that the excluded set is null in $\mathcal{O}$. If no root is finite, the point belongs to $\mathcal{N}_u(x)$. Otherwise the least root v has a unique position ℓ, belongs to $(u,1)$, and has $\lambda_v^\ell(x^\ell) > 0$. For $h \neq \ell$, strict minimality gives $\langle x^h, r^h\rangle - A_{u|v}^h(x) > 0$. For an active $d = (h,j)$,

$$r^d - \Delta_{u|v}^d(x) = r^d + B_{u|v}^{hj}(x) > 0.$$

The remaining equality at $c_\ell(x)$ follows from continuity at the first root. Thus the point belongs to $\mathcal{R}_u^\ell(x)$, proving (72).

On $\mathcal{B}_d$, with d active at x, all L passive coordinates remain free under σ_d. The excluded sets are still null by Fubini in these coordinates. The only additional zero coordinate is $r^d = 0$. If $d = (h,j)$ and $\bar{Q}_u^h(e_j|x) > 0$, continuity yields $B_{u|v}^{hj}(x) > 0$ for every $v > u$. Consequently $w^d > 0$, and the same argument proves (73). Joint Borel measurability and Tonelli, restricted first to bounded coordinate boxes if necessary, give (74).

For (iii), $\langle x^\ell, r^\ell\rangle = A_{u|v}^\ell(x)$. Since $\lambda_v^\ell(x^\ell) > 0$, this integral is strictly smaller for any upper limit below v, so $\rho_\ell(u,x,r) = v$. For $h \neq \ell$, positivity of $\langle x^h, w^h\rangle$ gives $\langle x^h, r^h\rangle > A_{u|v}^h(x)$ and hence $\rho_h(u,x,r) > v$. If r is interior, every active residual on $[u,v]$ is at least its positive final value. Residual $c_\ell(x)$ increases from zero and is positive at every time below v, since its speed is positive immediately below v. All other passive residuals increase from positive values. This proves absence of a hit. If $r^d = 0$ and $\bar{Q}_u^h(e_j|x) > 0$, residual d has value $B_{u|\tau}^{hj}(x) > 0$ at every $\tau \in (u,v)$, while the other active residuals have positive final values. Its first hit is therefore exactly at u.

For (iv), the definitions give the stated w, its positivity, and the specified speed. The preceding monotonicity arguments give the required segment. By (iii), another such segment would have to produce the same unique minimum among the ρ_h. Its time, position, and initial residual vector are therefore those in (71). □

Lemma 4. *Under Assumption 1, every finite construction in (44)–(47), including its success and failure events, is Borel measurable in its initial condition. After a hit satisfying (45) at which only its triggering residual is zero, every residual is strictly positive for a sufficiently short interval in decreasing time after the update, and no immediate further hit occurs.*

Proof. For fixed x, active (ℓ,j), and $u \in [0,1]$,

$$\{(v,\eta) : \theta_{\ell,j}(v,x,\eta) \geq u\} = \left\{(v,\eta) : v \geq u,\ \int_u^v \bar{Q}_\tau^\ell(e_j|x)\,\mathrm{d}\tau \geq \eta^{\ell j}\right\}$$

on $\eta^{\ell j} > 0$. The right-hand side is Borel. Finite maxima, tests of uniqueness and positivity, and integrations of continuous rates are Borel operations. Induction over these finitely many operations proves the first assertion, including measurability of (52)–(53).

Suppose the state changes from y to x at time v, with $y^{-\ell} = x^{-\ell}$, $y^\ell \neq x^\ell$, and $\bar{Q}_v^\ell(x^\ell|y) > 0$. Equation (59) gives $\lambda_v^\ell(x^\ell) > 0$. By continuity, $\lambda_\tau^\ell(x^\ell) > 0$ throughout a sufficiently short interval immediately below v, so the new passive residual increases strictly from zero. Every other residual is positive at the hit and has bounded speed, so all remain positive on a sufficiently short common interval. This proves the second assertion. □

J.5 CHANGE OF VARIABLES

Lemma 5. *Assume Assumption 1 and (49). Define, for every $u \in [0,1)$, $x \in \mathsf{X}$, and $r \in \mathcal{O}_x$,*

$$f_u^{(0)}(x,r) := F_u(x,r)\mathbb{1}_{\mathcal{N}_u(x)}(r). \tag{75}$$

Then, for every $u \in [0,1)$, every $x \in \mathsf{X}$, and every Borel $B \subseteq \mathcal{O}$,

$$\mu_u^{(0)}(\{x\} \times B) = \int_B f_u^{(0)}(x,r)\,\mathrm{d}r. \tag{76}$$

The measure $\mu_u^{(0)}$ gives zero mass to $\mathsf{X} \times ([0,\infty)^D \setminus \mathcal{O})$.

For every $x \in \mathsf{X}$, $d = (h,j) \in [\![1,L]\!] \times [\![1,K]\!]$ with $e_j \neq x^h$, and nonnegative Borel $\psi : (0,1) \times \mathcal{B}_d \to [0,\infty]$,

$$\begin{aligned}&\mathbb{E}\big[\mathbb{1}_{\{T_1\in(0,1),\,Y_1=x,\,C_1=d\}}\psi(T_1,W_1)\big]\\ &\quad= \int_0^1 \int_{\mathcal{B}_d} \psi(v,w)\bar{Q}_v^h(e_j|x) f_v^{(0)}(x,w)\,\sigma_d(\mathrm{d}w)\,\mathrm{d}v.\end{aligned} \tag{77}$$

The first hit, when present on $[0,1]$, satisfies (45) almost surely and has only its firing residual equal to zero. For each fixed $s \in [0,1]$, the probability of a first hit at s is zero.

Proof. Fix u, x. On $H_u^{(0)} \cap \{X_u = x\}$, the initial residuals corresponding to the endpoint r are

$$\eta = r - \Delta_{u|1}(x).$$

For a passive index $c_\ell(x)$, positivity of $\eta^{c_\ell(x)}$ is equivalent to $\langle x^\ell, r^\ell\rangle > A_{u|1}^\ell(x)$. For an active index (ℓ, j),

$$\eta^{\ell j} = r^{\ell j} + B_{u|1}^{\ell j}(x).$$

If r is interior, the active residual remains positive on the whole segment: at $\tau \in [u,1]$ it equals $r^{\ell j} + B_{u|\tau}^{\ell j}(x)$. These conditions are also necessary. The endpoint map is a translation with derivative matrix I_D. Its density is therefore

$$p_1(x)\gamma(r - \Delta_{u|1}(x))\mathbb{1}_{\mathcal{N}_u(x)}(r) = F_u(x,r)\mathbb{1}_{\mathcal{N}_u(x)}(r),$$

by (60). Initial residuals are strictly positive, and no hit down to u implies positive final active residuals and positive passive residuals. This proves (76) and the assertion about boundary mass.

We first exclude the exceptional first hits. Condition on $X_1 = x$. For an active $d = (h,j)$, a finite candidate hitting time v satisfies

$$E_1^d = B_{v|1}^{hj}(x).$$

As a function of v, the right-hand side is C^1, nonincreasing, and has derivative $-\bar{Q}_v^h(e_j|x)$. Lemma 2, applied after reversing the time parameter, shows that the initial values yielding a zero-speed candidate form a one-dimensional null set. An exponential variable avoids that set almost surely. A candidate at a prescribed $s \in [0,1]$ requires $E_1^d = B_{s|1}^{hj}(x)$, which is one fixed value and has probability zero. A tie for two distinct active indices $d = (h,j)$ and $a = (\ell,k)$ requires

$$(E_1^d, E_1^a) \in \{(B_{v|1}^{hj}(x), B_{v|1}^{\ell k}(x)) : v \in [0,1]\}.$$

This set is planar-null by Lemma 2(iii), whereas the pair of independent exponentials has a Lebesgue density. A finite union over indices and states excludes all these cases. The passive residuals increase from positive values, so all residuals other than the unique firing residual are positive at this hit.

Now fix $x, d = (h,j)$ and write $w = \iota_d(\zeta)$. Consider the open set

$$\begin{aligned}\mathcal{U}_{x,d} := \{(v,\zeta) \in (0,1)\times(0,\infty)^{D-1} : \bar{Q}_v^h(e_j|x) > 0,\\ \langle x^\ell, w^\ell\rangle > A_{v|1}^\ell(x) \text{ for every } \ell \in [\![1,L]\!]\}.\end{aligned} \tag{78}$$

Define $U : \mathcal{U}_{x,d} \to (0,\infty)^D$ by

$$U(v,\zeta) = w - \Delta_{v|1}(x). \tag{79}$$

In particular, its d-th coordinate is $B^{hj}_{v|1}(x)$. Its other active coordinates are $w^a + B^a_{v|1}(x)$, where $B^a_{v|1} = B^{\ell k}_{v|1}$ for $a = (\ell, k)$, and its passive coordinates are $\langle x^\ell, w^\ell\rangle - A^\ell_{v|1}(x)$. All of them are positive on (78).

Order the target coordinates as $(d, -d)$ and the source coordinates as (v, ζ). Since $\partial_v \Delta_{v|1}(x) = -b_v(x)$,

$$DU(v,\zeta) = \begin{pmatrix} -\bar{Q}^h_v(e_j|x) & 0_{1\times(D-1)} \\ b_v^{-d}(x) & I_{D-1} \end{pmatrix}, \qquad |\det DU(v,\zeta)| = \bar{Q}^h_v(e_j|x). \tag{80}$$

Here b_v^{-d} is the vector with its d-th component removed. Thus every non-firing column is a column of the identity, and expansion of the determinant in the first row gives the displayed value.

The map U is injective. Indeed, its d-th coordinate determines v uniquely among points where $\bar{Q}^h_v(e_j|x) > 0$, by monotonicity and Lemma 2(i). Once v is known, $w = \eta + \Delta_{v|1}(x)$ determines ζ. For an input in its image, the firing residual first reaches zero at v; the other active residuals have positive value w^a at v and did not reach zero earlier. Conversely, every first hit satisfying (45) yields one point of (78). Therefore the change-of-variables theorem applied to U gives the left-hand side of (77) as

$$\int_{\mathcal{U}_{x,d}} \psi(v, \iota_d\zeta)\, p_1(x)\gamma(U(v,\zeta))\bar{Q}^h_v(e_j|x)\,\mathrm{d}v\,\mathrm{d}\zeta.$$

The map is C^1 and has nonzero determinant on its open domain, so this change of variables is applicable to all nonnegative Borel integrands. By (60),

$$p_1(x)\gamma(w - \Delta_{v|1}(x)) \prod_{\ell=1}^{L} \mathbb{1}_{\{\langle x^\ell, w^\ell\rangle > A^\ell_{v|1}(x)\}} = f^{(0)}_v(x,w).$$

Inserting this identity, using (65), and observing that the integrand is zero when $\bar{Q}^h_v(e_j|x) = 0$ proves (77). The exceptional hits have already been excluded. □

Lemma 6. *Assume Assumption 1 and fix $x \in \mathsf{X}$. For each $\ell \in [\![1, L]\!]$, let $(v,w) \mapsto g^\ell_v(x,w)$ be a nonnegative, finite-valued Borel function on $(0,1) \times \mathcal{B}_{c_\ell(x)}$. Suppose that*

$$\sum_{\ell=1}^{L} \int_0^1 \int_{\mathcal{B}_{c_\ell(x)}} g^\ell_v(x,w)\,\sigma_{c_\ell(x)}(\mathrm{d}w)\,\mathrm{d}v < \infty, \qquad g^\ell_v(x,w) = 0 \text{ if } \lambda^\ell_v(x^\ell) = 0. \tag{81}$$

For data (v, ℓ, w) distributed according to these finite measures, start the constant-state segment at physical time v in state x, with residual vector w, and follow (43) toward 0 until its first active hit. Denote that hitting time by $\theta(v,x,w)$, its residual index by $C(v,x,w)$, and set $\theta = -\infty$ if there is no hit. Values of C on exceptional hits may be fixed arbitrarily.

For every $u \in [0,1)$ and $r \in \mathcal{O}_x$, define

$$h_u(x,r) := \begin{cases} \dfrac{g^\ell_v(x,w)}{\lambda^\ell_v(x^\ell)}, & r \in \mathcal{R}^\ell_u(x), \quad v = \rho_\ell(u,x,r), \quad w = r - \Delta_{u|v}(x), \\ 0, & r \notin \mathcal{R}_u(x). \end{cases} \tag{82}$$

This defines its values on $\mathcal{B}_d$ separately from its interpretation as an interior density. The following statements hold.

(i) For every deterministic $u \in [0,1)$ and every Borel set $B \subseteq \mathcal{O}$,

$$\int_B h_u(x,r)\,\mathrm{d}r = \sum_{\ell=1}^{L} \int_u^1 \int_{\mathcal{B}_{c_\ell(x)}} \mathbb{1}_{\{w+\Delta_{u|v}(x)\in B\}} g^\ell_v(x,w)\,\sigma_{c_\ell(x)}(\mathrm{d}w)\,\mathrm{d}v. \tag{83}$$

Only source times v with $u < v < 1$ occur in this formula. Its right-hand side is the measure of the data whose segment has no active hit down to u and ends in B.

(ii) For every active index $d = (h, j)$ with $e_j \neq x^h$ and every nonnegative Borel $\psi : (0,1) \times \mathcal{B}_d \to [0, \infty]$,

$$\sum_{\ell=1}^{L} \int_0^1 \int_{\mathcal{B}_{c_\ell(x)}} g_v^\ell(x,w) \mathbb{1}_{\{0<\theta(v,x,w)<v,\ C(v,x,w)=d\}} \times \psi\big(\theta(v,x,w), w + \Delta_{\theta(v,x,w)|v}(x)\big)\, \sigma_{c_\ell(x)}(\mathrm{d}w)\, \mathrm{d}v = \int_0^1 \int_{\mathcal{B}_d} \psi(u,r) \bar{Q}_u^h(e_j|x) h_u(x,r)\, \sigma_d(\mathrm{d}r)\, \mathrm{d}u. \tag{84}$$

Thus u on the right is the hitting time, and its source time always satisfies $0 < u < v < 1$.

(iii) With respect to the finite source measures in (81), a next hit, when present, is almost surely unique and has positive incoming speed. For every fixed $s \in [0,1]$, the set of data whose next hit occurs at s has measure zero. At a hit, only the firing residual is zero. No singular part in the interior or in the hitting measure is omitted in (83)–(84).

Proof. Fix an index ℓ, let $c = c_\ell(x)$, and consider its source measure. By (81), it is enough to use $v \in (0,1)$ with $\lambda_v^\ell(x^\ell) > 0$.

For a fixed such v, all $D-1$ coordinates w^{-c} have a density with respect to Lebesgue measure; independence is not needed. Residual c is passive and increases from zero, with positive speed immediately below v. It is therefore positive at every later elapsed time before the next state change. Every other passive residual increases from a strictly positive value. Each active residual $d = (h, j)$ starts from a positive free coordinate and has a finite candidate hitting time u only if

$$w^d = B_{u|v}^{hj}(x), \qquad 0 \le u < v. \tag{85}$$

For fixed v, Lemma 2 shows that values of w^d yielding $\bar{Q}_u^h(e_j|x) = 0$ form a null set. Two candidate hitting times coincide only if their two free coordinates lie on the Lipschitz curve of their integrated rates. This is null by Lemma 2(iii). A hit at a fixed $s < v$ imposes the single equality $w^d = B_{s|v}^{hj}(x)$ and is null as well. At $s = v$, an active hit is impossible because all active coordinates are positive initially. These conclusions hold under any density in w^{-c} by Fubini; integrating in v proves the assertions of (iii) about exceptional hits and prescribed times, including $s = 0$. Positivity of every non-firing residual at a unique hit follows from the same monotonicity and positivity of the passive residuals.

We prove (i). Fix $u \in [0,1)$, and write $w = \iota_c(\zeta)$ with $\zeta \in (0,\infty)^{D-1}$. On the open set

$$\{(v,\zeta) : u < v < 1,\ \lambda_v^\ell(x^\ell) > 0, \quad \zeta > 0,\ \iota_c(\zeta) + \Delta_{u|v}(x) \in \mathcal{O}\}, \tag{86}$$

consider

$$V_u(v,\zeta) = \iota_c(\zeta) + \Delta_{u|v}(x).$$

Order the target coordinates as $(c, -c)$. Its derivative matrix is

$$DV_u(v,\zeta) = \begin{pmatrix} \lambda_v^\ell(x^\ell) & 0_{1\times(D-1)} \\ b_v^{-c}(x) & I_{D-1} \end{pmatrix}, \qquad |\det DV_u| = \lambda_v^\ell(x^\ell). \tag{87}$$

In particular, the component $r^c = A_{u|v}^\ell(x)$ has derivative $\lambda_v^\ell(x^\ell)$ in v and no derivative in the free coordinates. Each other component has derivative $b_v^a(x)$ in v and the identity derivatives in those free coordinates. This proves the determinant formula by expansion in the first row.

Lemma 3(iii)–(iv) shows that V_u is injective on (86), with image $\mathcal{R}_u^\ell(x) \cap \mathcal{O}$ and inverse (71). Images for distinct ℓ are disjoint. Positivity of the final active coordinates is equivalent to absence of a hit down to u; all passive coordinates remain positive after leaving v. Applying the change-of-variables theorem, with determinant (87), and then summing over ℓ proves (83). The source measure is zero where $\lambda_v^\ell(x^\ell) = 0$, and the map has nonzero determinant everywhere else on its domain. Thus its interior pushforward has no omitted singular part.

We next prove (ii). Fix an active $d = (h, j)$, which is necessarily different from c. Use variables

$$0 < u < v < 1, \qquad \zeta \in (0, \infty)^{D-2},$$

indexed in the last factor by $([\![1, L]\!] \times [\![1, K]\!]) \setminus \{c, d\}$. Define r and w by

$$\begin{aligned} r^d &= 0, \qquad r^c = A^{\ell}_{u|v}(x), \qquad r^a = \zeta^a \quad (a \notin \{c, d\}), \\ w^c &= 0, \qquad w^d = B^{hj}_{u|v}(x), \qquad w^a = \zeta^a - \Delta^a_{u|v}(x) \quad (a \notin \{c, d\}). \end{aligned} \tag{88}$$

Restrict these variables to the open set on which

$$\lambda^{\ell}_v(x^{\ell}) > 0, \qquad \bar{Q}^h_u(e_j|x) > 0, \qquad w^a > 0 \quad (a \neq c). \tag{89}$$

On this set, $r \in \mathcal{B}_d$. Conversely, every next hit of the type under consideration, outside the zero-measure exceptions already excluded, determines exactly these variables. In fact, (85) determines u, and each competing active residual has the positive final value r^a. The other passive residuals are positive at both ends of the segment.

There are two changes of variables on (89). First define

$$U_1(v, u, \zeta) = (v, w^{-c}),$$

ordering the target as $(v, w^d, w^{-\{c,d\}})$. Since

$$\partial_v B^{hj}_{u|v}(x) = \bar{Q}^h_v(e_j|x), \qquad \partial_u B^{hj}_{u|v}(x) = -\bar{Q}^h_u(e_j|x),$$

and $\partial_v \Delta^a_{u|v} = b^a_v$, $\partial_u \Delta^a_{u|v} = -b^a_u$, its matrix is

$$DU_1 = \begin{pmatrix} 1 & 0 & 0 \\ \bar{Q}^h_v(e_j|x) & -\bar{Q}^h_u(e_j|x) & 0 \\ -b_v^{-\{c,d\}}(x) & b_u^{-\{c,d\}}(x) & I_{D-2} \end{pmatrix}. \tag{90}$$

Expansion in the first row and then in the first row of the remaining minor gives

$$|\det DU_1| = \bar{Q}^h_u(e_j|x). \tag{91}$$

For a fixed image (v, w^{-c}), the equation $w^d = B^{hj}_{u|v}(x)$ and the positive derivative magnitude at u give a unique u by Lemma 2. The other equations then determine ζ, so U_1 is injective.

Second define

$$U_2(v, u, \zeta) = (u, r^{-d}),$$

ordering the target as $(u, r^c, r^{-\{c,d\}})$. Its matrix is

$$DU_2 = \begin{pmatrix} 0 & 1 & 0 \\ \lambda^{\ell}_v(x^{\ell}) & -\lambda^{\ell}_u(x^{\ell}) & 0 \\ 0 & 0 & I_{D-2} \end{pmatrix}, \qquad |\det DU_2| = \lambda^{\ell}_v(x^{\ell}). \tag{92}$$

Indeed, the determinant of the upper-left 2×2 block is $-\lambda^{\ell}_v(x^{\ell})$. Lemma 3 gives the unique inverse of U_2 on its image, by first determining v from (u, x, r) and then w. These calculations also apply when $D = 2$, with I_0 and no ζ coordinates.

Applying these two substitutions in succession transforms

$$g^{\ell}_v(x, w)\, \mathrm{d}v\, \sigma_c(\mathrm{d}w) \quad \text{on hits through } d$$

into

$$\frac{\bar{Q}^h_u(e_j|x)}{\lambda^{\ell}_v(x^{\ell})} g^{\ell}_v(x, w)\, \mathrm{d}u\, \sigma_d(\mathrm{d}r), \qquad r \in \mathcal{R}^{\ell}_u(x) \cap \mathcal{B}_d, \quad \bar{Q}^h_u(e_j|x) > 0. \tag{93}$$

Both maps are C^1 on the stated open set, injective, and have nonzero determinants there. The image of U_2 is exactly the set in (93) with its time coordinate. Images for distinct ℓ are disjoint by Lemma 3. Summing (93) over ℓ gives (84) with the pointwise definition (82). Where $\bar{Q}^h_u(e_j|x) = 0$, its right-hand side is zero. The fixed-v argument at the start of the proof excludes any source mass of hits

with this speed or with ties. Hence there is no additional singular hitting measure, which completes the proof of (ii)–(iii). □

J.6 Induction on the number of jumps

Definition 1. Assume Assumption 1. Let $f_u(x,r)$ be a nonnegative Borel function on

$$\{(u,x,r) : 0 \le u < 1,\ x \in \mathsf{X},\ r \in \mathcal{O}_x\}.$$

For each such (u,x,r), define

$$(\mathcal{J}f)_u(x,r) := \begin{cases} \dfrac{1}{\lambda_v^\ell(x^\ell)} \displaystyle\sum_{\substack{y^{-\ell}=x^{-\ell}\\ y^\ell \neq x^\ell}} \bar{Q}_v^\ell(x^\ell|y) f_v(y,w), & r \in \mathcal{R}_u^\ell(x), \\ 0, & r \notin \mathcal{R}_u(x). \end{cases} \tag{94}$$

In the first line, ℓ is the unique index for which $r \in \mathcal{R}_u^\ell(x)$, $v = \rho_\ell(u,x,r)$, and $w = r - \Delta_{u|v}(x)$. If extended nonnegative functions are used, a summand with zero coefficient is defined to be zero. For each predecessor y in the sum, the index $c = c_\ell(x)$ is active at y and $w \in \mathcal{B}_c \subset \mathcal{O}_y$; thus every evaluation in (94) is on the specified domain. All denominators in its first line are strictly positive by (69).

Proposition 1. *Assume Assumption 1 and (49). Define $f^{(0)}$ by (75), with its values on all of $\mathcal{O}_x$, and recursively set*

$$f^{(n+1)} := \mathcal{J}f^{(n)}, \qquad n \in \mathbb{N} \cup \{0\}. \tag{95}$$

Then every $f^{(n)}$ is a finite-valued nonnegative Borel function. For every $n \ge 0$, every deterministic $u \in [0,1)$, every $x \in \mathsf{X}$, and every Borel $B \subseteq \mathcal{O}$,

$$\mu_u^{(n)}(\{x\} \times B) = \int_B f_u^{(n)}(x,r)\,\mathrm{d}r, \qquad \mu_u^{(n)}\big(\mathsf{X} \times ([0,\infty)^D \setminus \mathcal{O})\big) = 0. \tag{96}$$

Moreover, for every active index $d = (h,j)$ at x and every nonnegative Borel $\psi : (0,1) \times \mathcal{B}_d \to [0,\infty]$,

$$\begin{aligned} &\mathbb{E}\big[\mathbb{1}_{\{T_{n+1}\in(0,1),\, Y_{n+1}=x,\, C_{n+1}=d\}}\psi(T_{n+1}, W_{n+1})\big] \\ &\quad = \int_0^1 \int_{\mathcal{B}_d} \psi(v,w)\bar{Q}_v^h(e_j|x) f_v^{(n)}(x,w)\,\sigma_d(\mathrm{d}w)\,\mathrm{d}v. \end{aligned} \tag{97}$$

For each fixed finite index, a hit reached by the recursion is almost surely either absent or satisfies (45); only its firing residual is zero. For every fixed $s \in [0,1]$ and $n \ge 1$, the probability that the nth hit occurs at s is zero. None of these assertions assumes nonexplosion.

Proof. The function $f^{(0)}$ is finite-valued and Borel by its explicit definition. Lemma 3(i), the finite sums in (94), and positivity of the denominator on its domain show inductively that every $f^{(n)}$ is Borel and finite-valued. We prove the probabilistic assertions simultaneously by induction.

For $n = 0$, all assertions are Lemma 5. Suppose that the assertion about $\mu^{(n)}$, the event formula (97) at level n, and regularity of the first $n+1$ hits have been proved. Fix a possible state x after the $(n+1)$st hit and a position ℓ. Write $c = c_\ell(x)$. The disjoint ways of entering x through position ℓ have pre-jump states y with $y^{-\ell} = x^{-\ell}$ and $y^\ell \neq x^\ell$, and firing residual c. The residual vector is unchanged by the jump. Applying (97) to each such pre-jump state gives their sum as the measure with density

$$g_v^{(n+1),\ell}(x,w) := \sum_{\substack{y^{-\ell}=x^{-\ell}\\ y^\ell \neq x^\ell}} \bar{Q}_v^\ell(x^\ell|y) f_v^{(n)}(y,w), \qquad v \in (0,1), \quad w \in \mathcal{B}_c, \tag{98}$$

with respect to $\mathrm{d}v\,\sigma_c(\mathrm{d}w)$. More explicitly, for every nonnegative Borel φ on $(0,1)\times\mathcal{B}_c$,

$$\mathbb{E}\big[\mathbb{1}_{\{T_{n+1}\in(0,1),\,\Xi_{n+1}=x,\,C_{n+1}=c\}}\varphi(T_{n+1},W_{n+1})\big] = \int_0^1\int_{\mathcal{B}_c}\varphi(v,w)g_v^{(n+1),\ell}(x,w)\,\sigma_c(\mathrm{d}w)\,\mathrm{d}v. \tag{99}$$

Summing (99) over ℓ with $\varphi=1$ gives

$$\sum_{\ell=1}^{L}\int_0^1\int_{\mathcal{B}_{c_\ell(x)}} g_v^{(n+1),\ell}(x,w)\,\sigma_{c_\ell(x)}(\mathrm{d}w)\,\mathrm{d}v = \mathbb{P}(T_{n+1}\in(0,1),\Xi_{n+1}=x)\le 1. \tag{100}$$

If $\lambda_v^\ell(x^\ell)=0$, then $Q_v^\ell(y^\ell|x^\ell)=0$ for every predecessor y in the sum. Equation (57) and strict positivity of $p_v(y)$ imply $\bar{Q}_v^\ell(x^\ell|y)=0$. Therefore (98) is zero at such a time. It satisfies all the hypotheses (81) of Lemma 6.

Given $(T_{n+1},\Xi_{n+1},W_{n+1})=(v,x,w)$, the segment until the next hit is exactly $w+\Delta_{\tau|v}(x)$ in state x; it is a deterministic function of this triple. Consequently, after all preceding variables have been integrated out, its distribution is obtained by transporting the measures (99). Lemma 6(i) gives the density of its endpoint at any fixed $u<v$ without another hit as

$$\frac{g_v^{(n+1),\ell}(x,w)}{\lambda_v^\ell(x^\ell)} \quad \text{when } r\in\mathcal{R}_u^\ell(x),\quad v=\rho_\ell(u,x,r),\quad w=r-\Delta_{u|v}(x),$$

and zero off the union of these sets. By (98), this is exactly (94) applied to $f^{(n)}$. Its event description is $H_u^{(n+1)}$, so (96) follows at level $n+1$, including zero boundary mass.

Lemma 6(ii) applied to the same entrance measures gives the next-hit formula

$$\bar{Q}_v^h(e_j|x)f_v^{(n+1)}(x,w)\,\mathrm{d}v\,\sigma_d(\mathrm{d}w),\qquad d=(h,j),\quad e_j\neq x^h.$$

This is (97) for the $(n+2)$nd hit. Part (iii) of that lemma proves its regularity, positivity of all other residuals, and absence of an atom at every prescribed time. Lemma 4 allows continuation after each successful hit. This completes the induction for every finite n. □

Corollary 1. *Under the hypotheses of Proposition 1, for every $n\ge 0$, $u\in[0,1)$, $x\in\mathsf{X}$, and Borel $B\subseteq\mathcal{O}$,*

$$\mu_u^{(n+1)}(\{x\}\times B) = \sum_{\ell=1}^{L}\sum_{\substack{y^{-\ell}=x^{-\ell}\\ y^\ell\neq x^\ell}}\int_u^1\int_{\mathcal{B}_{c_\ell(x)}}\mathbb{1}_{\{w+\Delta_{u|v}(x)\in B\}} \times\bar{Q}_v^\ell(x^\ell|y)f_v^{(n)}(y,w)\,\sigma_{c_\ell(x)}(\mathrm{d}w)\,\mathrm{d}v. \tag{101}$$

In this integral $v=T_{n+1}$ and $u<v<1$. For every $r\in\mathcal{R}_u^\ell(x)$, with (v,w) given by (71),

$$\frac{f_u^{(n+1)}(x,r)}{F_u(x,r)} = \sum_{\substack{y^{-\ell}=x^{-\ell}\\ y^\ell\neq x^\ell}}\frac{Q_v^\ell(y^\ell|x^\ell)}{\lambda_v^\ell(x^\ell)}\frac{f_v^{(n)}(y,w)}{F_v(y,w)}. \tag{102}$$

In particular,

$$f_u^{(1)}(x,r) = F_u(x,r)\sum_{\substack{y^{-\ell}=x^{-\ell}\\ y^\ell\neq x^\ell}}\frac{Q_v^\ell(y^\ell|x^\ell)}{\lambda_v^\ell(x^\ell)}\prod_{h=1}^{L}\mathbb{1}_{\{\langle y^h,w^h\rangle>A_{v|1}^h(y)\}}. \tag{103}$$

Off $\mathcal{R}_u(x)$, $f_u^{(1)}(x,r)=0$ by definition.

Proof. Substitute (98) into (83). Its indicator is equivalent to no further hit before u because $B\subseteq\mathcal{O}$ and the active residuals are decreasing along this segment. This gives (101). For (102), the two

identities (57) and (60) give, pointwise at the specified (u, v, x, y, r, w),

$$\begin{aligned}\bar{Q}_v^\ell(x^\ell|y)F_v(y,w) &= p_v(x)Q_v^\ell(y^\ell|x^\ell)\gamma(w)\\ &= Q_v^\ell(y^\ell|x^\ell)F_u(x,r).\end{aligned} \tag{104}$$

All factors F are positive. Divide (94) by $F_u(x,r)$ and use (104) in each summand. Finally, substitute (75) at (v,y,w), where $w \in \mathcal{O}_y$, to obtain (103). □

J.7 Bounds on the partial sums and number of jumps

Lemma 7. *Assume Assumption 1. Define F by (48) on each $\mathcal{O}_x$, $f^{(0)}$ by (75), and $f^{(n+1)}$ by (95). For every $u \in [0,1)$, $x \in \mathsf{X}$, and $r \in \mathcal{O}_x$,*

$$(\mathcal{J}F)_u(x,r) = F_u(x,r)\mathbb{1}_{\mathcal{R}_u(x)}(r), \tag{105}$$

$$f_u^{(0)}(x,r) + (\mathcal{J}F)_u(x,r) = F_u(x,r)\mathbb{1}_{\mathcal{N}_u(x)\cup\mathcal{R}_u(x)}(r) \le F_u(x,r). \tag{106}$$

For every integer $m \ge 0$, at every such (u,x,r),

$$0 \le S_u^{(m)}(x,r) := \sum_{n=0}^{m} f_u^{(n)}(x,r) \le F_u(x,r). \tag{107}$$

These are pointwise statements about the specified functions, including their values at $r \in \mathcal{B}_d$ for every active index d at x.

Proof. If $r \notin \mathcal{R}_u(x)$, the definition of $\mathcal{J}$ gives (105). If $r \in \mathcal{R}_u^\ell(x)$, use (71) and (58) to calculate

$$\begin{aligned}(\mathcal{J}F)_u(x,r) &= \frac{\gamma(w)}{\lambda_v^\ell(x^\ell)} \sum_{\substack{y^{-\ell}=x^{-\ell}\\ y^\ell \ne x^\ell}} p_v(y)\bar{Q}_v^\ell(x^\ell|y)\\ &= p_v(x)\gamma(w) = F_u(x,r).\end{aligned}$$

The last step is (60). Since $\mathcal{N}_u(x)$ and $\mathcal{R}_u(x)$ are disjoint, (75) yields (106) at every point of $\mathcal{O}_x$.

For $m = 0$, (107) follows from (75). Suppose it holds for an integer m at every time, state, and point of the specified domain. Linearity and positivity of the finite sums defining $\mathcal{J}$ imply

$$S^{(m+1)} = f^{(0)} + \mathcal{J}S^{(m)} \le f^{(0)} + \mathcal{J}F \le F.$$

Every evaluation of $S^{(m)}$ in this inequality is within its domain, as checked in Definition 1. The induction is therefore valid pointwise, including on the faces, and proves the result. □

Proposition 2. *Assume Assumption 1 and (49). For $u \in [0,1]$, set*

$$M_u := \int_u^1 \sum_{x\in\mathsf{X}} p_v(x)\bar{\lambda}_v(x)\,\mathrm{d}v. \tag{108}$$

Then $M_u < \infty$, and

$$M_u = \int_u^1 \sum_{x\in\mathsf{X}} p_v(x)\lambda_v(x)\,\mathrm{d}v. \tag{109}$$

For every $u \in [0,1)$ and every integer $m \ge 1$,

$$\sum_{q=1}^{m} \mathbb{P}(T_q \in (u,1)) \le M_u, \qquad \mathbb{P}(T_m \in (u,1)) \le \frac{M_u}{m}. \tag{110}$$

Here T_q is a successful finite step time from (46) and is set to $-\infty$ if that step does not exist. The construction almost surely has no failure and only finitely many hits on $[0,1]$. For every deterministic

$u \in [0,1]$, *its jump count*

$$N_u := \sum_{q \geq 1} \mathbb{1}_{\{T_q \in (u,1)\}} \tag{111}$$

satisfies $\mathbb{E}N_u \leq M_u$.

Proof. For each $q \in \{1, \ldots, m\}$, take $\psi(v, w) = \mathbb{1}_{\{u<v<1\}}$ in (97) at level $q-1$ and sum over the pre-jump state and active residual. Since these choices partition the possible qth jumps, Tonelli's theorem gives

$$\sum_{q=1}^{m} \mathbb{P}(T_q \in (u,1)) = \int_u^1 \sum_{x \in \mathsf{X}} \sum_{\ell=1}^{L} \sum_{j : e_j \neq x^\ell} \bar{Q}_v^\ell(e_j|x) \times \int_{\mathcal{B}_{(\ell,j)}} S_v^{(m-1)}(x, w)\, \sigma_{(\ell,j)}(\mathrm{d}w)\, \mathrm{d}v. \tag{112}$$

Lemma 7 applies pointwise to the functions on these particular faces. For each active $d = (\ell, j)$,

$$\begin{aligned} \int_{\mathcal{B}_d} F_v(x, w)\, \sigma_d(\mathrm{d}w) &= p_v(x) \int_{(0,\infty)^{D-1}} \exp\left(- \sum_{a \neq d} w^a \right) \mathrm{d}w^{-d} \\ &= p_v(x) \prod_{a \neq d} \int_0^\infty \exp(-\zeta)\, \mathrm{d}\zeta = p_v(x). \end{aligned} \tag{113}$$

Substitution in (112) proves the first inequality in (110). Lemma 1 gives a common finite bound on $\bar{\lambda}_v(x)$, which proves $M_u < \infty$.

For each fixed v, the full time-reversal formula (39) gives

$$\begin{aligned} \sum_x p_v(x) \bar{\lambda}_v(x) &= \sum_x \sum_{y \neq x} p_v(x) \bar{Q}_v(y|x) \\ &= \sum_x \sum_{y \neq x} p_v(y) Q_v(x|y) = \sum_y p_v(y) \lambda_v(y). \end{aligned}$$

Integration proves (109).

If $T_m \in (u,1)$, then $T_q \in (u,1)$ for all $1 \leq q \leq m$, by their strict ordering. Hence

$$m\mathbb{P}(T_m \in (u,1)) \leq \sum_{q=1}^{m} \mathbb{P}(T_q \in (u,1)) \leq M_u,$$

which proves the second inequality.

For each finite index, Proposition 1 assigns probability zero to failure of the hitting conditions, including a hit at 0. The union of these failure events over the positive integers is null. On its complement, an infinite number of hits in $[0,1]$ would imply $T_n \in (0,1)$ for every n. These events decrease with n, and (110) at $u = 0$ implies

$$\mathbb{P}\big(T_n \in (0,1) \text{ for all } n \geq 1\big) = \lim_{n \to \infty} \mathbb{P}(T_n \in (0,1)) = 0.$$

Thus a successful finite construction eventually reaches a segment having no further hit down to 0. This yields a path on all of $[0,1]$ with finitely many jumps. Every finite construction is measurable by Lemma 4; the full path at each fixed time is defined on the countable union of these successful finite constructions and is consequently measurable. The probability of a hit at any given deterministic time is zero by Proposition 1 and a countable union over the index.

Finally, monotone convergence applied to the first inequality in (110) gives

$$\mathbb{E}N_u = \sum_{q \geq 1} \mathbb{P}(T_q \in (u,1)) \leq M_u.$$

At $u = 1$, both quantities are zero. □

J.8 Proof of Theorem J.1

Proof of Theorem J.1. Proposition 2 gives the almost-sure construction, finiteness of the number of jumps, and absence of a hit at every prescribed deterministic time.

Fix $u \in [0, 1)$. Except on the null set of failures, explosion, or a hit at u, the events $H_u^{(n)}$, $n \geq 0$, are disjoint and cover the sample space. On this set $H_u^{(n)} = \{N_u = n\}$. By (53) and Proposition 1, the law at time u is therefore supported on $\mathsf{X} \times \mathcal{O}$ and has density

$$f_u(x, r) := \sum_{n=0}^{\infty} f_u^{(n)}(x, r), \qquad x \in \mathsf{X}, \quad r \in \mathcal{O}. \tag{114}$$

Lemma 7 and monotone convergence give $0 \leq f_u(x, r) \leq F_u(x, r)$ pointwise. Tonelli's theorem and the partition by $H_u^{(n)}$ give

$$\begin{aligned} \sum_x \int_{\mathcal{O}} f_u(x, r)\, \mathrm{d}r &= \sum_{n=0}^{\infty} \sum_x \int_{\mathcal{O}} f_u^{(n)}(x, r)\, \mathrm{d}r \\ &= \sum_{n=0}^{\infty} \mathbb{P}(H_u^{(n)}) = 1. \end{aligned}$$

By the definition of F_u,

$$\sum_x \int_{\mathcal{O}} F_u(x, r)\, \mathrm{d}r = \sum_x p_u(x) \prod_{c \in [\![1,L]\!] \times [\![1,K]\!]} \int_0^{\infty} \exp(-\zeta)\, \mathrm{d}\zeta = 1.$$

Consequently,

$$\sum_x \int_{\mathcal{O}} \big(F_u(x, r) - f_u(x, r)\big)\, \mathrm{d}r = 0.$$

Its integrand is nonnegative, so $f_u(x, \cdot) = F_u(x, \cdot)$ for Lebesgue-almost every r, for each x. This proves (50) for every deterministic $u < 1$. At $u = 1$ it is (49). Integrating the product density over the coordinate box $0 < r^{\ell j} \leq \nu^{\ell j}$ gives

$$p_u(x) \prod_{(\ell,j) \in [\![1,L]\!] \times [\![1,K]\!]} \int_0^{\nu^{\ell j}} \exp(-\zeta)\, \mathrm{d}\zeta = p_u(x) \prod_{(\ell,j) \in [\![1,L]\!] \times [\![1,K]\!]} \big(1 - \exp(-\nu^{\ell j})\big),$$

which is (51). □

Corollary 2. *Under the hypotheses of Theorem J.1, for every deterministic $u \in [0, 1)$ and integer $m \geq 0$,*

$$\sum_x \int_{\mathcal{O}} \left(F_u(x, r) - \sum_{n=0}^{m} f_u^{(n)}(x, r) \right) \mathrm{d}r = \mathbb{P}(N_u \geq m + 1) \leq \frac{M_u}{m + 1}. \tag{115}$$

Proof. The integrand is nonnegative by (107). Subtract the masses of the events $H_u^{(0)}, \ldots, H_u^{(m)}$ from the total mass one in (114). This gives the equality. Except on a null set, $\{N_u \geq m + 1\} = \{T_{m+1} \in (u, 1)\}$; apply (110). □

J.9 Last-jump distribution

Proposition 3. *Assume the hypotheses of Theorem J.1. Fix $u \in [0, 1)$. On $\{N_u \geq 1\}$ define*

$$R_u := T_{N_u}, \qquad I_u := \textit{the first coordinate of } C_{N_u}.$$

For every $x \in \mathsf{X}$ and $\ell \in [\![1, L]\!]$, put $c = c_\ell(x)$. The joint measure of (R_u, E_u^{-c}), restricted to $\{N_u \geq 1, X_u = x, I_u = \ell\}$, is

$$\mathbb{P}(N_u \geq 1, X_u = x, I_u = \ell, R_u \in \mathrm{d}v, E_u^{-c} \in \mathrm{d}r^{-c})$$
$$= p_u(x)\lambda_v^\ell(x^\ell) \exp\left[-A_{u|v}^\ell(x) - \sum_{a \neq c} r^a\right] \prod_{h \neq \ell} \mathbb{1}_{\{\langle x^h, r^h\rangle > A_{u|v}^h(x)\}} \,\mathrm{d}v\,\mathrm{d}r^{-c}, \tag{116}$$

on $(u, 1) \times (0, \infty)^{D-1}$. On this event, the remaining component satisfies $E_u^c = A_{u|R_u}^\ell(x)$. The no-jump subdensity is (75).

Proof. Almost surely there are finitely many successful jumps, no jump at u, and only the firing residual is zero at a jump. On $\{N_u \geq 1, X_u = x\}$, apply Lemma 3(iii)–(iv) to the segment from its last jump $v = R_u$ to u. It shows that $E_u \in \mathcal{R}_u^{I_u}(x)$, that $R_u = \rho_{I_u}(u, x, E_u)$, and that $E_u^{c_{I_u}(x)} = A_{u|R_u}^{I_u}(x)$. If there is no jump, the endpoint instead belongs to $\mathcal{N}_u(x)$ by the proof of Lemma 5.

Theorem J.1 supplies the interior density $p_u(x)\gamma(r)$. On $\mathcal{R}_u^\ell(x) \cap \mathcal{O}$, replace the coordinate r^c by v through

$$r^c = A_{u|v}^\ell(x), \qquad r^a = r^a \quad (a \neq c).$$

The determinant in these coordinates is $\lambda_v^\ell(x^\ell)$. The conditions for ℓ to be the least reconstruction index are precisely

$$\langle x^h, r^h\rangle > A_{u|v}^h(x) \quad (h \neq \ell).$$

All active coordinates are already positive. On the set $\lambda_v^\ell(x^\ell) > 0$, the change is injective, as in Lemma 3. Lemma 2 shows that the remaining values of r^c form a null set; the corresponding right-hand side of (116) is zero where $\lambda_v^\ell(x^\ell) = 0$. Substituting $r^c = A_{u|v}^\ell(x)$ into the exponential density and multiplying by the determinant gives exactly (116). □

Corollary 3. *Under the hypotheses of Theorem J.1, suppose $L = 1$ and identify the state space with V. For every $u \in [0, 1)$, $j \in [\![1, K]\!]$, and $r \in \mathcal{O}_{e_j}$,*

$$f_u^{(0)}(e_j, r) = p_u(e_j)\gamma(r)\mathbb{1}_{\{r^{1j} > \int_u^1 \lambda_\tau^1(e_j)\,\mathrm{d}\tau\}}.$$

If $r \in \mathcal{R}_u(e_j)$, let $v \in (u, 1)$ be the unique solution with $\lambda_v^1(e_j) > 0$ of

$$r^{1j} = \int_u^v \lambda_\tau^1(e_j)\,\mathrm{d}\tau,$$

and set $w^{1j} = 0$, $w^{1k} = r^{1k} + \int_u^v \bar{Q}_\tau^1(e_k|e_j)\,\mathrm{d}\tau$ for $k \neq j$. Then, for every integer $n \geq 0$,

$$f_u^{(n+1)}(e_j, r) = \frac{1}{\lambda_v^1(e_j)} \sum_{k \in [\![1,K]\!] \setminus \{j\}} \bar{Q}_v^1(e_j|e_k) f_v^{(n)}(e_k, w). \tag{117}$$

It is zero when $r \notin \mathcal{R}_u(e_j)$.

Proof. There is one passive residual, namely $c_1(e_j) = (1, j)$. Formula (67) reduces to the single inequality in the no-jump formula. In (94), the only possible value of ℓ is 1, the predecessor state is e_k, and the reversed rate from this predecessor into e_j is $\bar{Q}_v^1(e_j|e_k)$. Equations (55) and (71) give the stated values of w. This proves (117) and the zero convention directly from the general recursion. □