THE ACTIVITY GEOMETRY OF DISCRETE GENERATIVE SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Iterative refinement is a key capability of diffusion-based generation and has motivated growing efforts to bring similar behavior to discrete diffusion and flow models. A growing line of work enables such refinement by allowing previously generated tokens to be revised during sampling, thereby shaping the coupling between successive sampling states. Despite its growing role in sampler design, this coupling freedom remains insufficiently characterized. We introduce **Activity Geometry** to study these finite-step couplings through expected transition cost, which we call activity. We show that activity decomposes into the minimum required by the induced marginal transport and an excess term arising from the coupling. Using this minimum as a reference, we propose Minimal-Activity Finite-Time Coupling (MAFTC) to realize prescribed marginals with minimum activity without retraining or additional model evaluations. In the continuous-time limit, the minimum activity per unit time converges to the minimum expected jump rate compatible with the prescribed path. For metric-induced Gibbs paths, we further show that the activity of the kinetic-optimal flux admits a geometric decomposition. We find that total activity alone can obscure coupling excess: a sampler may approach the minimum activity required by the prescribed marginal transport while still retaining substantial coupling excess. In language generation, as the number of model evaluations increases, refinement-based samplers accumulate additional trajectory activity with limited diversity gains, whereas MAFTC reduces trajectory activity while maintaining comparable generation quality and higher lexical diversity.

1 INTRODUCTION

Discrete diffusion (Austin et al., 2021; Lou et al., 2024) and flow (Gat et al., 2024; Shaul et al., 2025) models provide frameworks for iterative refinement, with samplers (Campbell et al., 2022) performing successive token updates. At each step, the sampler uses model predictions to specify the probabilities of retaining current token values or replacing them with new ones (Gat et al., 2024; Shaul et al., 2025). In particular, sampling rules can be designed to allow previously generated tokens to be revised, providing a mechanism for iterative refinement. Currently, efforts to enable iterative refinement span sampling methods for masked and uniform-state diffusion, as well as probability-path design. In masked diffusion (Shi et al., 2024), the ancestral sampler used by MDLM (Sahoo et al., 2024) keeps tokens fixed once unmasked; ReMDM (Wang et al., 2025a) then enables these tokens to be remasked and resampled, narrowing the sample quality gap with autoregressive models in text generation. Meanwhile, uniform-state diffusion (Hoogeboom et al., 2021; Austin et al., 2021; Sahoo et al., 2025) allows generated tokens to transition directly to other token values. Ψ-samplers (Deschenaux et al., 2026) further extend predictor–corrector sampling to arbitrary noise priors and, in uniform-state models, continue to improve sample quality as the number of sampling steps increases. Beyond that, probability-path design provides a complementary route: FUDOKI (Wang et al., 2025b) uses metric-induced paths and kinetic-optimal velocities (Shaul et al., 2025) for iterative refinement in a unified model for visual understanding and image generation.

Given the input distribution, each sampling rule induces a *coupling*, the joint distribution of successive sampling states. These revision mechanisms provide additional coupling choices, creating opportunities for correction while also permitting repeated token changes. Yuan et al. (2026) show that stochastic transitions can preserve the prescribed time marginals while contracting sampling errors.

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.

At the same time, Xiang et al. (2026) identify flip-flop oscillations in which tokens are remasked and later restored to their previous values, and show that these oscillatory revisions can be reduced while preserving generation quality. However, how coupling choices shape token movement during sampling has not been studied systematically.

To this end, we introduce **Activity Geometry**, a framework for analyzing finite-step couplings through expected transition cost, which we call *activity*. Our main contributions are as follows:

(i) Finite-time Activity Geometry. We establish a minimum-activity reference for transport between categorical distributions under the discrete mismatch cost. We then decompose a sampler's activity into the minimum required by its induced marginal transport and the excess introduced by its coupling. We quantify marginal mismatch between the induced and prescribed output marginals in total variation. Using this minimum-activity reference, we propose **Minimal-Activity Finite-Time Coupling (MAFTC)**, which exactly transports prescribed categorical marginals with minimum activity. We extend this optimality to transitions between factorized conditional marginals under a weighted Hamming cost and obtain a sampler for pretrained discrete generative models without retraining or additional model evaluations.

(ii) Continuous-time consistency and kinetic localization. For a differentiable probability path on a finite state space, we establish that the minimum finite-time activity per unit time converges to the minimum expected jump rate compatible with the path as the step size tends to zero. We further give an exact geometric localization of the excess jump activity of the kinetic-optimal flow on metric-induced Gibbs paths.

(iii) Trajectory activity and generation quality. We find in pretrained image and language models that comparable generation quality can be achieved with substantially fewer token changes. In our text experiments, MAFTC's trajectory activity decreases as the number of sampling steps increases, whereas the two baseline samplers, Duo+DCD and Ψ, accumulate more token changes. At similar generative perplexity (Gen-PPL), MAFTC also achieves higher corpus-level diversity than both baselines.

2 Background: Discrete Flow Matching

Let $\mathcal{T} = [V] \triangleq \{1, \dots, V\}$ be our vocabulary. The state space is given by $\mathcal{X} = \mathcal{T}^D$ where $D \in \mathbb{N}$ is sequence length. We consider discrete generative modeling over sequences $x = (x^1, \dots, x^D) \in \mathcal{X}$, where $x^i \in \mathcal{T}$ denotes the token at position i.

Probability paths. We denote the source and data probability mass functions (PMFs) on $\mathcal{X}$ by p_0 and p_1, respectively, and consider a factorized source distribution $p_0(x) = \prod_{i=1}^{D} p_0^i(x^i)$. Following discrete flow matching (Gat et al., 2024; Shaul et al., 2025), we consider a *probability path* $(p_t)_{t\in[0,1]}$ interpolating between p_0 and p_1. We construct this path by introducing a data endpoint $A \sim p_1$ and marginalizing *conditional probability paths* over its realizations $a \in \mathcal{X}$, with

$$p_t(x) = \sum_{a\in\mathcal{X}} p_t(x \mid a) p_1(a), \qquad p_t(x \mid a) = \prod_{i=1}^{D} p_t^i(x^i \mid a^i). \tag{1}$$

We refer to $(p_t)_{t\in[0,1]}$ as the *marginal probability path*. For each coordinate i, the conditional probability path $p_t^i(\cdot \mid a^i)$ interpolates between the source marginal p_0^i and the point mass at a^i, satisfying the boundary conditions $p_0^i(\cdot \mid a^i) = p_0^i$ and $p_1^i(\cdot \mid a^i) = \delta_{a^i}$, where $\delta_b(z) = \mathbf{1}\{z = b\}$ denotes the point mass at b.

Probability velocities and fluxes. We model the generative dynamics by a time-inhomogeneous continuous-time Markov chain (CTMC) $(X_t)_{t\in[0,1]}$ whose time marginals follow the prescribed probability path,

$$X_t \sim p_t, \qquad t \in [0, 1]. \tag{2}$$

Its infinitesimal dynamics are governed by a *probability velocity* u_t, also known as a rate matrix (Gat et al., 2024; Shaul et al., 2025). Using the destination-first convention, $u_t(y, x)$ denotes the instantaneous transition rate from state x to state y, so that for small $h > 0$,

$$\Pr(X_{t+h} = y \mid X_t = x) = \delta_x(y) + h\, u_t(y, x) + o(h), \qquad h \downarrow 0, \tag{3}$$

where $o(h)/h \to 0$, $u_t(y, x) \geq 0$ for $y \neq x$, and $\sum_y u_t(y, x) = 0$. We say that u_t generates the probability path $(p_t)_{t\in[0,1]}$ when the associated CTMC has time marginal p_t at every t.

We restrict the CTMC to transitions that change at most one token at a time. For each coordinate i, we denote by $u_t^i(y^i, x)$ the transition rate from x to the state obtained by replacing x^i with $y^i \neq x^i$ while leaving the remaining coordinates unchanged. We set the diagonal entry to $u_t^i(x^i, x) := -\sum_{z\in\mathcal{T}\setminus\{x^i\}} u_t^i(z, x)$. Combining these coordinatewise rates gives the probability velocity on $\mathcal{X}$:

$$u_t(y, x) = \sum_{i=1}^{D} u_t^i(y^i, x) \prod_{j\neq i} \delta_{x^j}(y^j). \tag{4}$$

Thus, for $x \neq y$, the rate $u_t(y, x)$ is zero when x and y differ in more than one coordinate, and equals $u_t^i(y^i, x)$ when they differ only at coordinate i.

We define the *probability flux* $j_t(y, x) := p_t(x)u_t(y, x)$, whose off-diagonal entries quantify the probability mass moving from state x to state y per unit time. The marginal evolution is governed by the continuity equation, also known as the Kolmogorov forward equation,

$$\dot{p}_t(y) = \sum_{x\neq y} j_t(y, x) - \sum_{x\neq y} j_t(x, y), \qquad \dot{p}_t(y) := \frac{d}{dt}p_t(y). \tag{5}$$

For a fixed probability path, this equation determines only the net flux at each state, not the individual pairwise fluxes $j_t(y, x)$. Consequently, different probability velocities can generate the same probability path (Shaul et al., 2025).

Posterior prediction and marginal velocities. Following the marginalization construction of Gat et al. (2024); Shaul et al. (2025), we obtain marginal velocities by marginalizing conditional velocities with respect to the endpoint posterior. For each coordinate i, we let $u_t^i(y^i, x^i \mid a^i)$ denote a conditional velocity generating $p_t^i(\cdot \mid a^i)$, and the corresponding marginal velocity is given by

$$u_t^i(y^i, x) = \sum_{a^i\in\mathcal{T}} u_t^i(y^i, x^i \mid a^i)\, p_{1|t}^i(a^i \mid x). \tag{6}$$

We then combine these marginal velocities through Eq. (4) to obtain a probability velocity u_t generating the marginal probability path. Here $p_{1|t}^i(\cdot \mid x)$ is the endpoint posterior for the i-th token,

$$p_{1|t}^i(a^i \mid x) := \Pr(A^i = a^i \mid X_t = x) = \sum_{a'\in\mathcal{X}} \delta_{a^i}(a'^i)\frac{p_t(x \mid a')p_1(a')}{p_t(x)}, \qquad p_t(x) > 0. \tag{7}$$

Because these endpoint posteriors depend on the unknown data distribution p_1, practical discrete flow models approximate them with a learned factorized posterior parameterized by θ,

$$q_\theta(a \mid x, t) = \prod_{i=1}^{D} q_{\theta,i}(a^i \mid x, t), \tag{8}$$

where each $q_{\theta,i}(\cdot \mid x, t)$ conditions on the full current state x and approximates $p_{1|t}^i(\cdot \mid x)$.

Sampling on a time grid. For a PMF ν and Markov kernel K on $\mathcal{X}$, we write $(\nu K)(y) := \sum_x \nu(x)K(y \mid x)$ for the marginal obtained by applying K to ν. For $0 \leq t < s \leq 1$, we denote the CTMC transition kernel by $P_{s|t}(y \mid x) := \Pr(X_s = y \mid X_t = x)$, which satisfies $p_t P_{s|t} = p_s$. Practical samplers operate on a grid of N sampling intervals, $0 = t_0 < \cdots < t_N = 1$, and approximate $P_{t_{n+1}|t_n}$ with a tractable transition kernel $K_{t_n,t_{n+1}}$ constructed from model predictions. Euler-type schemes (Campbell et al., 2022; Shaul et al., 2025) freeze the transition rates at the beginning of each interval, whereas corrected schemes (Wan et al., 2026; Fu & Fan, 2026) incorporate within-interval corrections to this approximation. Applying $K_{t_n,t_{n+1}}$ to the prescribed time marginal p_{t_n} gives the *induced output marginal*

$$\widetilde{p}_{t_{n+1}} = p_{t_n} K_{t_n,t_{n+1}}. \tag{9}$$

During rollout, however, the same kernel acts on the sampler's current time marginal, which may already differ from p_{t_n}. We use the number of function evaluations (NFE) to measure the sampling budget; Appendix E specifies the counting convention under classifier-free guidance.

3 FINITE-TIME ACTIVITY GEOMETRY

We develop Activity Geometry by relating the expected cost of a finite-time transition to its input and output marginals. We first separate the minimum activity required by the induced marginal transport from coupling excess and relate activity to marginal mismatch. We then construct MAFTC, extend it to factorized conditional probability paths, and define the trajectory quantities used in our experiments.

3.1 TRANSITION KERNELS AND A MINIMUM-ACTIVITY REFERENCE

We first consider a single categorical variable with state space $\mathcal{T}$, and extend the construction to sequences in $\mathcal{X} = \mathcal{T}^D$ in Section 3.4. Let p and q denote the input and prescribed output marginals, respectively. We consider a Markov kernel K on $\mathcal{T}$, where $K(y \mid x) \geq 0$ for all $x, y \in \mathcal{T}$ and $\sum_{y \in \mathcal{T}} K(y \mid x) = 1$ for every $x \in \mathcal{T}$.

Definition 3.1 (Finite-time transition and activity)**.** For $X \sim p$ and $Y \mid X \sim K(\cdot \mid X)$, we define the *induced coupling* π_K and induced output marginal $\widetilde{q}$ by

$$\pi_K(x, y) = p(x)K(y \mid x), \qquad \widetilde{q}(y) = (pK)(y) = \sum_x p(x)K(y \mid x).$$

For a nonnegative transition cost $d : \mathcal{T} \times \mathcal{T} \to [0, \infty)$, the *activity* of K under input marginal p is $A_d(K) := \mathbb{E}_{\pi_K}[d(X, Y)]$. For the discrete mismatch cost $d(x, y) = \mathbf{1}\{x \neq y\}$, we abbreviate this to $A(K) = \Pr_{\pi_K}(X \neq Y)$.

Let $\Gamma(p, q)$ denote the set of couplings of p and q. By the classical maximal-coupling identity (Griffeath, 1975; Thorisson, 2000), the *minimum activity* required to transport p to q is

$$A_{\min}(p, q) := \min_{\pi \in \Gamma(p,q)} \Pr_{\pi}(X \neq Y) = \mathrm{TV}(p, q) = \frac{1}{2} \sum_x |p(x) - q(x)|, \tag{10}$$

where TV denotes the total variation distance. A maximal coupling assigns mass $\min\{p(x), q(x)\}$ to $X = Y = x$ for each state x, so the remaining probability mass corresponds to state changes. Appendix A gives the derivation and characterizes the minimizing couplings.

3.2 INDUCED MARGINAL, MARGINAL MISMATCH, AND COUPLING EXCESS

A Markov kernel need not realize the prescribed output marginal q. With p and q fixed, we distinguish the minimum activity needed to reach the induced marginal $\widetilde{q} = pK$ from that needed to reach q:

$$A_{\mathrm{ind}}(K) := A_{\min}(p, \widetilde{q}), \qquad A_{\mathrm{pres}} := A_{\min}(p, q), \qquad \epsilon(K) := \mathrm{TV}(\widetilde{q}, q). \tag{11}$$

The *marginal mismatch* $\epsilon(K)$ measures the total variation distance between the induced output marginal $\widetilde{q}$ and the prescribed output marginal q. Since $\pi_K \in \Gamma(p, \widetilde{q})$, Eq. (10) gives $A(K) = \Pr_{\pi_K}(X \neq Y) \geq A_{\min}(p, \widetilde{q}) = A_{\mathrm{ind}}(K)$. We define the *coupling-excess activity* as $\chi(K) := A(K) - A_{\mathrm{ind}}(K)$, obtaining

$$A(K) = A_{\mathrm{ind}}(K) + \chi(K), \qquad \chi(K) \geq 0. \tag{12}$$

The decomposition separates the minimum activity needed to reach $\widetilde{q}$ from additional changes induced by the coupling; $\epsilon(K)$ measures marginal mismatch separately. In particular, activity below A_{pres} implies nonzero marginal mismatch. We use $[z]_+ := \max\{z, 0\}$ to denote the positive part of a scalar z.

Proposition 3.2 (Marginal mismatch bound)**.** *For PMFs p, q on $\mathcal{T}$ and any Markov kernel K on $\mathcal{T}$,*

$$A(K) + \epsilon(K) \geq A_{\mathrm{pres}}, \qquad \epsilon(K) \geq [A_{\mathrm{pres}} - A(K)]_+. \tag{13}$$

For every $b \in [0, A_{\mathrm{pres}}]$, there is a Markov kernel K_b with $A(K_b) = b$ and $\epsilon(K_b) = A_{\mathrm{pres}} - b$.

Total activity close to A_{pres} can still contain substantial coupling excess when the induced marginal transition requires less activity. Appendix A proves the bound and its sharpness.

3.3 MINIMAL-ACTIVITY FINITE-TIME COUPLING

We now construct a kernel that maps p to q with activity A_{pres}, so that both $\epsilon(K)$ and $\chi(K)$ vanish. At each state x, we retain the common mass $c(x) := \min\{p(x), q(x)\}$ and define the *loss and gain residuals* by $r^-(x) := [p(x) - q(x)]_+$ and $r^+(x) := [q(x) - p(x)]_+$. These are the masses to remove from and add to state x, respectively. Since p and q both sum to one, the total mass removed equals the total mass added: $\tau := \sum_x r^-(x) = \sum_x r^+(x) = \mathrm{TV}(p, q)$. For $\tau > 0$, we retain $c(x)$ at each state and distribute the mass $r^-(x)$ from each loss state x among gain states y in proportion to $r^+(y)$. The resulting coupling π^{MA}, with MA denoting minimum activity, is

$$\pi^{\mathrm{MA}}(x, x) = c(x), \qquad \pi^{\mathrm{MA}}(x, y) = \frac{r^-(x) r^+(y)}{\tau} \quad (x \neq y). \tag{14}$$

We call this construction *Minimal-Activity Finite-Time Coupling (MAFTC)*. For $\tau > 0$ and $p(x) > 0$, normalizing each row yields the Markov kernel

$$K^{\mathrm{MA}}(y \mid x) = \begin{cases} \min\{p(x), q(x)\}/p(x), & y = x, \\ r^-(x) r^+(y)/(\tau p(x)), & y \neq x. \end{cases} \tag{15}$$

If $\tau = 0$, then $p = q$, and we use the diagonal coupling $\pi^{\mathrm{MA}}(x, y) = p(x)\delta_x(y)$ and identity kernel $K^{\mathrm{MA}}(y \mid x) = \delta_x(y)$. For $\tau > 0$, we also set $K^{\mathrm{MA}}(\cdot \mid x) = \delta_x$ on rows with $p(x) = 0$, which do not affect the induced coupling.

Proposition 3.3 (Marginal exactness and activity optimality of MAFTC)**.** *For PMFs p, q on $\mathcal{T}$, the MAFTC coupling has marginals p, q and mismatch probability $\mathrm{TV}(p, q)$. Its kernel satisfies $pK^{\mathrm{MA}} = q$ and $\epsilon(K^{\mathrm{MA}}) = \chi(K^{\mathrm{MA}}) = 0$.*

For $p(x) > 0$, the kernel retains x with probability $\min\{1, q(x)/p(x)\}$. When a move occurs, it draws the destination from r^+/τ. This rule requires $O(V)$ work to form and sample from the residual distribution, without constructing a $V \times V$ coupling matrix. Appendix A proves the proposition.

3.4 TOKEN SEQUENCES AND CONDITIONAL SAMPLING

We extend the categorical construction to sequences in $\mathcal{X} = \mathcal{T}^D$ using an additive cost. For positive coordinate weights $w_1, \ldots, w_D$, define the *weighted Hamming distance* $d_{H,w}(x, y) := \sum_{i=1}^D w_i \mathbf{1}\{x^i \neq y^i\}$.

Proposition 3.4 (Minimum weighted Hamming activity)**.** *Let $p = \bigotimes_{i=1}^D p^i$ and $q = \bigotimes_{i=1}^D q^i$ be PMFs on $\mathcal{X}$, where p^i, q^i are their coordinate marginals. Every coupling $\Pi \in \Gamma(p, q)$ satisfies*

$$\mathbb{E}_\Pi[d_{H,w}(X, Y)] \geq \sum_{i=1}^D w_i \,\mathrm{TV}(p^i, q^i). \tag{16}$$

The product of coordinatewise MAFTC couplings attains equality and is therefore optimal over all sequence couplings, including those with dependence across coordinates.

We apply this result to the factorized conditional probability path in Eq. (1). Fix a data endpoint $a \in \mathcal{X}$ and sampling times $0 \leq t < s \leq 1$. For each coordinate i, construct K_i^{MA} from the time marginals $p_t^i(\cdot \mid a^i)$ and $p_s^i(\cdot \mid a^i)$. Their product, $K_{t,s}^a(y \mid x) := \prod_{i=1}^D K_i^{\mathrm{MA}}(y^i \mid x^i)$, then satisfies

$$p_t(\cdot \mid a) K_{t,s}^a = p_s(\cdot \mid a), \tag{17}$$

and minimizes expected weighted Hamming distance between these conditional time marginals. Appendix A gives the proof and a coordinatewise activity–error bound that does not require factorized marginals.

The minimum-activity guarantee above applies to each fixed-endpoint transition. To recover the marginal path, we average $K_{t,s}^a$ with the exact joint *endpoint posterior* $p_{1|t}(a \mid x) := \Pr(A = a \mid X_t = x)$. For $p_t(x) > 0$, Bayes' rule gives $p_t(x) p_{1|t}(a \mid x) = p_1(a) p_t(x \mid a)$; hence the kernel $K_{t,s}^\star(y \mid x) := \sum_a p_{1|t}(a \mid x) K_{t,s}^a(y \mid x)$ satisfies $p_t K_{t,s}^\star = p_s$. Algorithm 1 instead samples endpoint tokens independently from the learned factorized posterior q_θ and applies the conditional kernels in parallel, using one model evaluation per interval. We denote the sampler state at t_n by $\widehat{X}_n$ and an optional conditioning input by ξ. Appendix C gives the resulting kernel and proves exactness of the joint-posterior construction; Appendix D describes the implementation.

Algorithm 1: MAFTC sampling for endpoint-conditioned discrete flows

Input: conditioning input ξ, source p_0, grid $0 = t_0 < \cdots < t_N = 1$, learned token posteriors $q_{\theta,i}$, conditional probability paths $p_t^i(\cdot \mid a^i)$

Sample $\widehat{X}_0 \sim p_0$.
for $n = 0, \ldots, N-1$ **do**
  Obtain $q_{\theta,i}(\cdot \mid \widehat{X}_n, t_n, \xi)$ for all i with one model evaluation.
  Independently sample $a^i \sim q_{\theta,i}(\cdot \mid \widehat{X}_n, t_n, \xi)$ for all i; set $a = (a^1, \ldots, a^D)$.
  for $i = 1, \ldots, D$ **in parallel do**
    Set $p \leftarrow p_{t_n}^i(\cdot \mid a^i)$, $q \leftarrow p_{t_{n+1}}^i(\cdot \mid a^i)$, and $x \leftarrow \widehat{X}_n^i$.
    Set $\tau \leftarrow \mathrm{TV}(p, q)$; **if** $\tau = 0$ or $p(x) = 0$, set $\widehat{X}_{n+1}^i \leftarrow x$ and **continue**.
    Set $\kappa \leftarrow \min\{1, q(x)/p(x)\}$ and draw $U \sim \mathrm{Unif}[0, 1]$.
    if $U < \kappa$, set $\widehat{X}_{n+1}^i \leftarrow x$; **otherwise**, sample $\widehat{X}_{n+1}^i \sim [q - p]_+/\tau$.
  end for
end for
return $\widehat{X}_N$.

3.5 From one-step activity to sampled trajectories

To describe realized token changes over a rollout, we use the unweighted Hamming distance $d_H(x, y) := \sum_{i=1}^D \mathbf{1}\{x^i \neq y^i\}$. We define the *trajectory activity* L_{act}, *net displacement* D_{net}, and *trajectory excess* R_{traj} by

$$L_{\mathrm{act}} = \sum_{n=0}^{N-1} d_H(\widehat{X}_n, \widehat{X}_{n+1}), \qquad D_{\mathrm{net}} = d_H(\widehat{X}_0, \widehat{X}_N), \qquad R_{\mathrm{traj}} = L_{\mathrm{act}} - D_{\mathrm{net}} \geq 0. \tag{18}$$

The triangle inequality makes R_{traj} nonnegative. It measures Hamming path length beyond the initial-to-final distance and can be positive even when each transition minimizes activity, because a token can change multiple times over the trajectory. These quantities describe individual sampled trajectories, whereas $A(K)$ and $\chi(K)$ are expectations under a specified input marginal and kernel. Averages of these trajectory quantities describe rollouts under the sampler's current time marginals, which may differ from the prescribed time marginals. NFE counts model evaluations, while L_{act} counts realized token changes.

4 Continuous-Time Limit and Kinetic Localization

We connect the finite-time minimum to the minimum expected jump rate along a differentiable probability path. We then specialize to metric-induced Gibbs paths and identify the transitions that account for the excess jump activity of the kinetic-optimal flow.

4.1 The minimum expected jump rate

For a CTMC, the expected jump rate is often called *dynamical activity* or *traffic* (Shiraishi et al., 2018; Maes, 2020; Van Vu & Saito, 2023). We consider a differentiable probability path $(p_t)_{t\in[0,1]}$ on $\mathcal{T}$ and fix an interior time t. Following the flux formulation of Shaul et al. (2025), we call an off-diagonal probability flux $j_t(y, x)$ *admissible* if it is nonnegative, vanishes whenever $p_t(x) = 0$, and satisfies the continuity equation. Its *instantaneous activity* $\mathcal{A}_t(j)$ is the total probability flux per unit time, so the constraint and objective are

$$\dot{p}_t(y) = \sum_{x \neq y} j_t(y, x) - \sum_{x \neq y} j_t(x, y), \qquad \mathcal{A}_t(j) := \sum_{x \neq y} j_t(y, x). \tag{19}$$

We denote the rates of probability gain and loss at state x by $g_t(x) := [\dot{p}_t(x)]_+$ and $\ell_t(x) := [-\dot{p}_t(x)]_+$. Because $\sum_x \dot{p}_t(x) = 0$, they have equal total mass per unit time, $M_t := \sum_x g_t(x) = \sum_x \ell_t(x) = \frac{1}{2}\|\dot{p}_t\|_1$, where $\|\dot{p}_t\|_1 = \sum_x |\dot{p}_t(x)|$.

Proposition 4.1 (Instantaneous minimum activity). *Every admissible flux for $\dot{p}_t$ satisfies*

$$\mathcal{A}_t(j) \geq M_t = \tfrac{1}{2}\|\dot{p}_t\|_1. \tag{20}$$

For $M_t > 0$, equality is attained by

$$j_t^{\mathrm{MA}}(y, x) = \ell_t(x) g_t(y)/M_t, \qquad x \neq y. \tag{21}$$

Equality holds exactly when all nonzero flux moves from states with $\dot{p}_t < 0$ to states with $\dot{p}_t > 0$, with statewise outflow ℓ_t and inflow g_t. If $M_t = 0$, the zero flux is the unique minimizer.

This instantaneous minimum is the limit of the finite-time reference. Differentiability on the finite state space gives

$$\lim_{h \downarrow 0} \frac{\mathrm{TV}(p_t, p_{t+h})}{h} = \tfrac{1}{2}\|\dot{p}_t\|_1, \tag{22}$$

so the minimum finite-time activity per unit time converges to the minimum expected jump rate. The MAFTC construction also has a flux limit. Let $\pi^{\mathrm{MA}}_{t,t+h}$ be its coupling of p_t and p_{t+h}. For $M_t > 0$, the loss and gain residuals divided by h converge to ℓ_t and g_t, giving

$$\frac{\pi^{\mathrm{MA}}_{t,t+h}(x, y)}{h} \longrightarrow j_t^{\mathrm{MA}}(y, x), \qquad x \neq y.$$

Appendix A proves the proposition and these limits.

4.2 Localization on metric-induced Gibbs paths

We specialize Eq. (1) to the metric-induced Gibbs paths of Shaul et al. (2025), which are also used in FUDOKI (Wang et al., 2025b). Suppressing the coordinate index, we fix an endpoint token $a \in \mathcal{T}$ and a finite nonnegative endpoint-distance score $\rho(x, a)$, with $\rho(a, a) = 0 < \rho(x, a)$ for $x \neq a$. Using inverse temperature $\beta \geq 0$ as the path parameter, we write

$$p_\beta(x \mid a) = \frac{e^{-\beta\rho(x,a)}}{Z_a(\beta)}, \qquad Z_a(\beta) := \sum_{z \in \mathcal{T}} e^{-\beta\rho(z,a)}. \tag{23}$$

The path starts at the uniform source when $\beta = 0$ and converges to δ_a as $\beta \to \infty$. For a fixed finite β, we abbreviate $p_\beta(\cdot \mid a)$ to p_β and define the current mean endpoint distance as $\mu := \sum_x p_\beta(x)\rho(x, a)$. The identity $\partial_\beta p_\beta(x) = (\mu - \rho(x, a))p_\beta(x)$ shows how this mean partitions the states: those above μ lose probability mass, those below it gain mass, and those at the mean have zero probability derivative.

The kinetic-optimal (KO) probability velocity, which minimizes the symmetric kinetic-energy objective of Shaul et al. (2025), induces the flux

$$j_\beta^{\mathrm{KO}}(y, x) = p_\beta(x)p_\beta(y)[\rho(x, a) - \rho(y, a)]_+, \qquad x \neq y. \tag{24}$$

We denote its activity per unit β by $\mathcal{A}_{\mathrm{KO}}$ and the minimum over fluxes admissible for $\partial_\beta p_\beta$ by $\mathcal{A}_{\mathrm{MA}}$. To express both in terms of endpoint distance, take independent token-valued $X, X' \sim p_\beta$ and write $R := \rho(X, a)$ and $R' := \rho(X', a)$. The flux sum and Proposition 4.1 give, respectively,

$$\mathcal{A}_{\mathrm{KO}} = \tfrac{1}{2}\mathbb{E}|R - R'|, \qquad \mathcal{A}_{\mathrm{MA}} = \tfrac{1}{2}\mathbb{E}|R - \mu|. \tag{25}$$

To locate their difference, we partition the state space into $H := \{x : \rho(x, a) > \mu\}$ and $L := \{x : \rho(x, a) \leq \mu\}$, placing states at the mean in L.

Theorem 4.2 (Localization of kinetic-optimal excess activity). *For the Gibbs path in Eq. (23), the total kinetic-optimal flux from H to L equals the minimum activity:*

$$\sum_{x \in H,\, y \in L} j_\beta^{\mathrm{KO}}(y, x) = \mathcal{A}_{\mathrm{MA}}. \tag{26}$$

The excess activity is exactly the flux within the two groups:

$$\mathcal{A}_{\mathrm{KO}} - \mathcal{A}_{\mathrm{MA}} = \sum_{\substack{x,y \in H \\ x \neq y}} j_\beta^{\mathrm{KO}}(y, x) + \sum_{\substack{x,y \in L \\ x \neq y}} j_\beta^{\mathrm{KO}}(y, x). \tag{27}$$

The within-group flux accounts for the excess over the minimum expected jump rate while contributing to the prescribed statewise probability derivatives. At times with finite β_t, a differentiable nondecreasing schedule β_t multiplies the flux and both activities by $\dot{\beta}_t \geq 0$, giving the corresponding quantities per unit model time and preserving the decomposition. Appendix B gives the proof and the finite-interval threshold separating loss and gain residuals.

5 EXPERIMENTS

We evaluate activity decomposition and kinetic localization, then examine how finite-step coupling choices affect terminal accuracy and generation performance. Appendix E provides full protocols and results.

Setup. On frozen FUDOKI (Wang et al., 2025b), we compare Euler, time-corrected (TC) sampling (Wan et al., 2026), and MAFTC. On a Duo checkpoint obtained by Discrete Consistency Distillation (DCD) (Sahoo et al., 2025), we compare ancestral sampling (Duo+DCD), Ψ (Deschenaux et al., 2026), and MAFTC. Within each model, methods share the checkpoint, path, grid, truncation, and decoding settings at matched NFE. Each configuration uses all 553 GenEval prompts (Ghosh et al., 2023) with four images each, or 2,048 length-1,024 text sequences; we record every trajectory. We report GenEval for compositional image quality, GPT-2-large Gen-PPL for text quality, and pooled Dist-4 for corpus four-gram coverage.

5.1 ACTIVITY DECOMPOSITION AND KINETIC LOCALIZATION

Figure 1(a) shows that Euler's conditional activity nearly matches the prescribed minimum, yet 24.7% is coupling excess. Its induced marginal transition requires substantially less activity than the prescribed one, but coupling excess brings the total close to the prescribed minimum. For the corresponding continuous-time kinetic-optimal flow, every jump decreases endpoint distance, yet flux within the mean-distance groups accounts for 24.3% of its integrated activity (Theorem 4.2; Appendix E.3). Excess activity thus arises even without backtracking.

5.2 TERMINAL ACCURACY AND GENERATION PERFORMANCE

Exact-posterior sampling. We propagate the complete distribution on $\{0, 1, 2, 3\}^3$ using exact token posteriors and parallel updates, measuring terminal joint TV against the target distribution. Figure 1(b) compares accuracy and activity on Gibbs paths with dependent targets (Appendix E.7). Coarse Euler steps can make fewer changes while incurring greater terminal error. On finer grids, MAFTC achieves lower terminal TV than Euler and TC with fewer token changes.

Pretrained generators. Table 1 compares quality and cumulative token changes at matched NFE. MAFTC improves GenEval over Euler on coarse grids and retains competitive generation quality with fewer changes on finer grids. At 32 NFE, MAFTC reduces text trajectory activity by 74.2% and 75.0% relative to Duo+DCD and Ψ, respectively, while maintaining similar Gen-PPL and comparable pooled Dist-4.

Table 1: **Generation quality and token changes at matched NFE.** Higher is better for GenEval (↑) and pooled Dist-4 (↑); lower is better for Gen-PPL (↓). NTC denotes the mean cumulative number of token changes per sample (L_{act}), rounded to the nearest integer. Shading identifies MAFTC.

	FUDOKI (Image)						Duo (Text)								
	Euler		TC		MAFTC		Duo+DCD			Ψ			MAFTC		
NFE	GenEval	NTC	GenEval	NTC	GenEval	NTC	Gen-PPL	NTC	Dist-4	Gen-PPL	NTC	Dist-4	Gen-PPL	NTC	Dist-4
8	0.690	2146	0.722	2277	0.712	1880	59.65	2761	0.8624	59.37	2783	0.8622	60.64	1114	0.8657
32	0.753	2408	0.745	2450	0.744	1883	34.07	4044	0.8497	33.51	4177	0.8493	33.59	1042	0.8515
128	0.727	2618	0.739	2649	0.737	2067	29.77	5269	0.8461	27.77	5849	0.8390	28.39	1029	0.8437

Figure 1(c,d) separates cumulative token changes into net displacement and trajectory excess. Net displacement varies little with NFE, so baseline activity growth mainly reflects trajectory excess. Table 1 and Figure 1(c,d) show that, from 8 to 128 NFE, Duo+DCD and Ψ roughly double their cumulative token changes, whereas MAFTC's text activity decreases by about 8%. MAFTC's text trajectory excess falls by about 62%, bringing cumulative token changes closer to net displacement. MAFTC's image activity grows more slowly and remains below Euler and TC.

The comparisons at similar Gen-PPL in Appendix E.6 show higher pooled Dist-4 for MAFTC than for Duo+DCD and Ψ, while within-sequence repetition varies across comparisons.

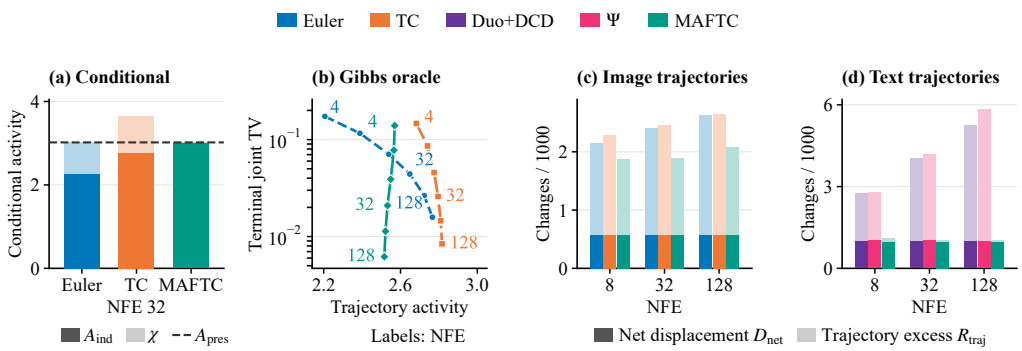


Figure 1: **Activity decomposition, terminal accuracy, and sampling trajectories.** (a) Conditional FUDOKI activity at 32 NFE: induced minimum A_{ind} and coupling excess χ, with prescribed minimum A_{pres} dashed. Activity quantities are summed over intervals and averaged over endpoints. (b) Gibbs terminal joint TV versus cumulative changes; points average eight targets using exact token posteriors at 4–128 steps. (c,d) Image and text changes at 8, 32, and 128 NFE, split into net displacement D_{net} and trajectory excess R_{traj}.

6 Related Work

Jump objectives and finite-time couplings. Campbell et al. (2024) construct jump-minimizing generators for uniform and masking conditional paths, whereas Haxholli et al. (2026) optimize source–target training couplings under a jump-cost objective. For prescribed paths, General Discrete Paths (Shaul et al., 2025) minimizes symmetric kinetic energy; we localize kinetic-optimal excess activity on metric-induced Gibbs paths. At finite steps, maximal coupling appears in learned noising processes (Bartosh et al., 2026) and uniform-diffusion bridges (Song et al., 2021; Gourevitch et al., 2026). MAFTC applies this principle to prescribed conditional marginals. Our framework decomposes activity into the minimum required by induced marginal transport and coupling excess, while measuring marginal mismatch separately.

Revision and error correction. ReMDM (Wang et al., 2025a) and Ψ-samplers (Deschenaux et al., 2026) enable revision through predictor–corrector sampling. Yuan et al. (2026) analyze marginal-preserving stochastic transitions that contract sampling errors, whereas COVER (Xiang et al., 2026) reduces remask–restore oscillations. We quantify changes relative to a marginal-transport minimum, providing a distributional reference for interpreting revision activity.

Sampling accuracy. Corrected samplers (Wan et al., 2026; Fu & Fan, 2026) improve finite-step approximation, while evaluations with exact token posteriors (Tang et al., 2026) isolate sampler error from learned prediction error. We complement these studies by examining sampling accuracy and generation quality alongside activity.

7 Conclusion

Activity Geometry separates the minimum activity required by an induced marginal transition from coupling excess, while measuring marginal mismatch separately. MAFTC attains this minimum for categorical marginals and the weighted Hamming optimum for factorized conditional marginals. For differentiable probability paths, the finite-time minimum per unit time converges to the minimum expected jump rate. On metric-induced Gibbs paths, kinetic-optimal excess activity is exactly the flux within the two groups separated by the current conditional mean distance to the endpoint. Conditional-kernel experiments reveal coupling excess hidden by total activity near the prescribed minimum. In pretrained image and language models, MAFTC achieves comparable generation quality on finer grids with substantially fewer token changes. Trajectory activity thus complements terminal quality in characterizing discrete sampling.

AI USE STATEMENT

Generative AI tools, including OpenAI ChatGPT, assisted with literature organization, brainstorming, mathematical consistency checks, language editing, and code and experiment analysis. The authors verified mathematical claims through explicit derivations, checked literature claims against the cited sources, and reviewed and tested code changes before use. All reported experiments and numerical results were produced from author-controlled code and experiment logs. The authors reviewed the complete submission and take responsibility for its content.

REPRODUCIBILITY STATEMENT

Appendices A and B give complete proofs of the main claims and state their finite-state assumptions. Appendix D describes the vectorized MAFTC update, numerical stability, and computational complexity. Appendix E records the frozen checkpoints, sampling configurations, image and language metrics, activity definitions, and paired-bootstrap protocol. The accompanying code and experiment artifacts contain the manifests, evaluation scripts, and result files used to produce every table and figure.

REFERENCES

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

Grigory Bartosh, Teodora Pandeva, Sushrut Karmalkar, and Javier Zazo. Forward-learned discrete diffusion: Learning how to noise to denoise faster. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2605.18204.

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

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5453–5512. PMLR, 2024. URL https://proceedings.mlr.press/v235/campbell24a.html.

Justin Deschenaux, Caglar Gulcehre, and Subham Sekhar Sahoo. The diffusion duality, chapter II: Ψ-samplers and efficient curriculum. In *International Conference on Learning Representations*, 2026.

Feiyang Fu and Hehe Fan. A time-reparameterized cumulative intensity extrapolation sampler for discrete flow matching. In *International Conference on Machine Learning*, 2026.

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.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, volume 36, pp. 52132–52152, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

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

David Griffeath. A maximal coupling for Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 31(2):95–106, 1975. doi: 10.1007/BF00539434.

Etrit Haxholli, Yeti Z. Gürbüz, Oğul Can, and Eli Waxman. Minibatch optimal transport and perplexity bound estimation in discrete flow matching. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=A8rmJlSET9.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12454–12465, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/67d96d458abdef21792e6d8e590244e7-Paper.pdf.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119. Association for Computational Linguistics, 2016. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 762–770. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.86. URL https://aclanthology.org/2022.acl-short.86/.

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

Christian Maes. Frenesy: Time-symmetric dynamical activity in nonequilibria. *Physics Reports*, 850: 1–33, 2020. doi: 10.1016/j.physrep.2020.01.002.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T. Chiu, and Volodymyr Kuleshov. The diffusion duality. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52584–52619, 2025. URL https://proceedings.mlr.press/v267/sahoo25a.html.

Neta Shaul, Itai Gat, Marton Havasi, Daniel Severo, Anuroop Sriram, Peter Holderrieth, Brian Karrer, Yaron Lipman, and Ricky T. Q. Chen. Flow matching with general discrete paths: A kinetic-optimal perspective. In *International Conference on Learning Representations*, 2025.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103131–103167, 2024. doi: 10.52202/079017-3277. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html.

Naoto Shiraishi, Ken Funo, and Keiji Saito. Speed limit for classical stochastic processes. *Physical Review Letters*, 121(7):070601, 2018. doi: 10.1103/PhysRevLett.121.070601.

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

Luhan Tang, Longxuan Yu, Shaorong Zhang, and Greg Ver Steeg. Is your diffusion sampler actually correct? A sampler-centric evaluation of discrete diffusion language models. In *International Conference on Machine Learning*, 2026.

Hermann Thorisson. *Coupling, Stationarity, and Regeneration*. Springer, 2000.

Tan Van Vu and Keiji Saito. Thermodynamic unification of optimal transport: Thermodynamic uncertainty relation, minimum dissipation, and thermodynamic speed limits. *Physical Review X*, 13(1):011013, 2023. doi: 10.1103/PhysRevX.13.011013.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

Zhengyan Wan, Yidong Ouyang, Liyan Xie, Hongyuan Zha, Fang Fang, and Guang Cheng. Corrected samplers for discrete flow models. In *International Conference on Machine Learning*, 2026.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 163488–163545, 2025a. doi: 10.52202/085713-4927.

Jin Wang, Yao Lai, Aoxue Li, Shifeng Zhang, Jiacheng Sun, Ning Kang, Chengyue Wu, Zhenguo Li, and Ping Luo. FUDOKI: Discrete flow-based unified understanding and generation via kinetic-optimal velocities. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Gian Wiher, Clara Meister, and Ryan Cotterell. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012, 2022. doi: 10.1162/tacl_a_00502. URL https://aclanthology.org/2022.tacl-1.58/.

Yanzheng Xiang, Lan Wei, Yizhen Yao, Qinglin Zhu, Hanqi Yan, Chen Jin, Philip Alexander Teare, Dandan Zhang, Lin Gui, Amrutha Saseendran, and Yulan He. Stop the flip-flop: Context-preserving verification for fast revocable diffusion decoding. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.06161.

William Yuan, Sungwon Jeong, and Amirali Aghazadeh. On the error-correcting effects of stochasticity in discrete diffusion. *arXiv preprint arXiv:2605.26582*, 2026. URL https://arxiv.org/abs/2605.26582.

CONTENTS

Abstract **1**

1 Introduction **1**

2 Background: Discrete Flow Matching **2**

3 Finite-Time Activity Geometry **4**
3.1 Transition kernels and a minimum-activity reference 4
3.2 Induced marginal, marginal mismatch, and coupling excess 4
3.3 Minimal-Activity Finite-Time Coupling 5
3.4 Token sequences and conditional sampling 5
3.5 From one-step activity to sampled trajectories 6

4 Continuous-Time Limit and Kinetic Localization **6**
4.1 The minimum expected jump rate 6
4.2 Localization on metric-induced Gibbs paths 7

5 Experiments **8**
5.1 Activity decomposition and kinetic localization 8
5.2 Terminal accuracy and generation performance 8

6 Related Work **9**

7 Conclusion **9**

AI use statement **10**

Reproducibility statement **10**

References **10**

A Proofs of Main Results **14**
A.1 Classical maximal coupling and residual pairing 14
A.2 Proof of Proposition 3.2 and sharpness 15
A.3 MAFTC construction and proof of Proposition 3.3 15
A.4 Proof of Proposition 3.4 15
A.5 Proof of Proposition 4.1 16

B Metric-Induced Paths and Kinetic Localization **17**
B.1 Proof of Lemma B.1 18
B.2 Proof of Theorem 4.2 19

C Conditional and Posterior-Marginalized Updates **19**

D Implementation and Numerical Stability **20**
D.1 Vectorized MAFTC 20
D.2 Complexity 20

E Experimental Details and Full Results **20**
E.1 Evaluation quantities 20
E.2 Full-state evaluation of conditional kernels 21
E.3 Integrated expected jump rate 22
E.4 Finite-step baseline definitions 22
E.5 FUDOKI 23
E.6 Language generation with Duo 24
E.7 Exact-posterior evaluation of complete sampling distributions 27

A Proofs of Main Results

We establish the coupling lower bound, then prove the marginal mismatch bound, MAFTC optimality, the sequence result, and the continuous-time limit.

A.1 Classical maximal coupling and residual pairing

Let p and q be PMFs on the state space $\mathcal{T}$. The coupling set has the standard marginal constraints (Villani, 2009),

$$\Gamma(p,q) = \left\{ \pi \geq 0 : \begin{array}{ll} \sum_{y\in\mathcal{T}} \pi(x,y) = p(x), & x \in \mathcal{T}, \\ \sum_{x\in\mathcal{T}} \pi(x,y) = q(y), & y \in \mathcal{T} \end{array} \right\}.$$

We first separate the mass shared by p and q from the mass that must move. Let

$$c(x) := \min\{p(x), q(x)\}, \qquad r^-(x) := [p(x) - q(x)]_+, \qquad r^+(x) := [q(x) - p(x)]_+.$$

The loss residual r^- is the mass to remove from the input, and the gain residual r^+ is the mass to add to reach the output; they satisfy

$$p = c + r^-, \qquad q = c + r^+, \qquad q - p = r^+ - r^-.$$

For every state x, at least one of the two residuals is zero, giving $r^-(x)r^+(x) = 0$ and $|p(x)-q(x)| = r^-(x) + r^+(x)$. Moreover, since $\sum_x \big(q(x) - p(x)\big) = 0$, we have

$$\tau := \sum_x r^-(x) = \sum_x r^+(x) = 1 - \sum_x c(x) = \frac{1}{2}\sum_x |p(x) - q(x)| = \mathrm{TV}(p,q), \tag{28}$$

where the last equality is the standard expression for total variation distance on a discrete state space (Griffeath, 1975; Thorisson, 2000).

Lemma A.1 (Coupling lower bound and equality conditions)**.** *For every $\pi \in \Gamma(p,q)$, $\Pr_\pi(X \neq Y) \geq \mathrm{TV}(p,q)$. Equality holds exactly when $\pi(x,x) = \min\{p(x), q(x)\}$ for every x.*

Proof. Since π has marginals p and q, we have $\pi(x,x) \leq c(x)$ for every x. Hence, by Eq. (28),

$$\Pr_\pi(X \neq Y) - \mathrm{TV}(p,q) = \left(1 - \sum_x \pi(x,x)\right) - \left(1 - \sum_x c(x)\right) = \sum_x \big[c(x) - \pi(x,x)\big] \geq 0. \tag{29}$$

Thus $\Pr_\pi(X \neq Y) \geq \mathrm{TV}(p,q)$. Since every summand in Eq. (29) is nonnegative, equality holds if and only if $\pi(x,x) = c(x)$ for every x.

We next characterize the couplings attaining equality. Suppose $\tau > 0$, and let D_c be the diagonal array with $[D_c]_{xy} := c(x)\mathbf{1}\{x = y\}$. For any coupling π attaining equality, define

$$\gamma(x,y) := \frac{\pi(x,y) - [D_c]_{xy}}{\tau}. \tag{30}$$

The equality condition gives $\gamma(x,x) = 0$, while for $x \neq y$, $\gamma(x,y) = \pi(x,y)/\tau \geq 0$. Moreover, its row and column marginals are

$$\sum_y \gamma(x,y) = \frac{r^-(x)}{\tau}, \qquad \sum_x \gamma(x,y) = \frac{r^+(y)}{\tau}. \tag{31}$$

Therefore every coupling attaining equality with $\tau > 0$ has the form

$$\pi = D_c + \tau\gamma, \qquad \gamma \in \Gamma(r^-/\tau, r^+/\tau). \tag{32}$$

Conversely, let $\gamma \in \Gamma(r^-/\tau, r^+/\tau)$ and define $\pi := D_c + \tau\gamma$. Since $r^-(x)r^+(x) = 0$ for every x, at least one of the row or column marginals of γ at x is zero, so $\gamma(x,x) = 0$. Hence π has marginals p and q and diagonal entries $\pi(x,x) = c(x)$. By the equality condition above, it therefore attains $\Pr_\pi(X \neq Y) = \mathrm{TV}(p,q)$.

Finally, if $\tau = 0$, then $p = q$, and the equality condition uniquely gives the diagonal coupling $\pi(x,y) = p(x)\mathbf{1}\{x = y\}$. □

We also use the correspondence between couplings and Markov kernels. For any $\pi \in \Gamma(p,q)$, define

$$K_\pi(y \mid x) = \begin{cases} \pi(x,y)/p(x), & p(x) > 0, \\ \delta_x(y), & p(x) = 0. \end{cases} \tag{33}$$

Each row sums to one. A zero row marginal forces $\pi(x,y) = 0$ for all y, so $p(x)K_\pi(y \mid x) = \pi(x,y)$ also holds when $p(x) = 0$. Summing over x gives $pK_\pi = q$. Any probability distribution can replace the identity row on these zero-probability states without affecting the induced coupling.

A.2 Proof of Proposition 3.2 and sharpness

Proof. The induced law $\pi_K(x, y) = p(x)K(y \mid x)$ couples p and $\widetilde{q} = pK$. Lemma A.1 gives $A(K) \geq \mathrm{TV}(p, \widetilde{q})$, hence $\chi(K) \geq 0$. The triangle inequality gives

$$\mathrm{TV}(p, q) \leq \mathrm{TV}(p, \widetilde{q}) + \mathrm{TV}(\widetilde{q}, q) \leq A(K) + \epsilon(K).$$

Rearranging and using $\epsilon(K) \geq 0$ proves Eq. (13).
For sharpness, let $\tau = \mathrm{TV}(p, q) > 0$ and choose an activity level $b \in [0, \tau]$. Set $\lambda = b/\tau$ and $\widetilde{q} = (1 - \lambda)p + \lambda q$. Since $\widetilde{q} - p = \lambda(q - p)$ and $q - \widetilde{q} = (1 - \lambda)(q - p)$, homogeneity of the ℓ^1 norm gives

$$\mathrm{TV}(p, \widetilde{q}) = \lambda\,\mathrm{TV}(p, q) = b, \qquad \mathrm{TV}(\widetilde{q}, q) = (1 - \lambda)\,\mathrm{TV}(p, q) = \tau - b.$$

By Eq. (33), a maximal coupling of p and $\widetilde{q}$ therefore induces a kernel with $A(K) = b$ and $\epsilon(K) = \tau - b$. If $p = q$, the identity kernel attains zero activity and error. □

Corollary A.2 (Coordinatewise activity–error bound)**.** *Let Π couple sequence-valued input and output states with coordinate marginals p^i and $\widetilde{q}^i$, respectively. Fix prescribed output coordinate marginals q^i and positive weights $w_1, \ldots, w_D$. Then*

$$\mathbb{E}_\Pi[d_{H,w}(X, Y)] + \sum_i w_i\,\mathrm{TV}(\widetilde{q}^i, q^i) \geq \sum_i w_i\,\mathrm{TV}(p^i, q^i).$$

Proof of Corollary A.2. Project Π onto coordinate i and write $A_i := \Pr_\Pi(X^i \neq Y^i)$. This mismatch probability and the induced output marginal $\widetilde{q}^i$ satisfy Proposition 3.2, so

$$A_i + \mathrm{TV}(\widetilde{q}^i, q^i) \geq \mathrm{TV}(p^i, q^i).$$

Multiplication by $w_i > 0$ preserves the inequality. Summing over coordinates and using $\mathbb{E}_\Pi[d_{H,w}(X, Y)] = \sum_i w_i A_i$ gives the stated bound. □

A.3 MAFTC construction and proof of Proposition 3.3

Proof. Use the common mass and residuals from Appendix A.1. If $\tau = 0$, the identity kernel realizes $p = q$ with zero activity. For $\tau > 0$, choose the independent residual coupling

$$\gamma = (r^-/\tau) \otimes (r^+/\tau), \qquad \gamma(x, y) = \frac{r^-(x)r^+(y)}{\tau^2},$$

where $\otimes$ denotes the product of the two probability distributions. It belongs to $\Gamma(r^-/\tau, r^+/\tau)$, so Eq. (32) gives

$$\pi^{\mathrm{MA}} = D_c + \tau\gamma.$$

This coupling has marginals p, q and activity τ. Applying Eq. (33) gives the kernel in Eq. (15), with $pK^{\mathrm{MA}} = q$. Hence $\epsilon(K^{\mathrm{MA}}) = 0$ and $\chi(K^{\mathrm{MA}}) = \tau - \mathrm{TV}(p, q) = 0$.
For a loss state x, $p(x) > q(x)$, so its retention and move probabilities are $q(x)/p(x)$ and $r^-(x)/p(x)$, respectively. Conditional on a move,

$$\Pr(Y = y \mid X = x,\, Y \neq X) = \frac{K^{\mathrm{MA}}(y \mid x)}{r^-(x)/p(x)} = \frac{r^+(y)}{\tau}, \qquad y \neq x.$$

For a gain or neutral state with $p(x) > 0$, $r^-(x) = 0$ and the kernel retains the state. Zero-probability input states use the identity-row convention in Eq. (33). □

A.4 Proof of Proposition 3.4

Proof. Let Π be any coupling of the product distributions $p(x) = \prod_i p^i(x^i)$ and $q(y) = \prod_i q^i(y^i)$, written $p = \otimes_i p^i$ and $q = \otimes_i q^i$. We do not assume that Π factorizes. The joint law Π_i of (X^i, Y^i) belongs to $\Gamma(p^i, q^i)$ because its marginals are p^i and q^i. Therefore the one-coordinate result gives

$$\mathbb{E}_\Pi \mathbf{1}\{X^i \neq Y^i\} = \Pr_{\Pi_i}(X^i \neq Y^i) \geq \mathrm{TV}(p^i, q^i).$$

Multiplying by $w_i > 0$ and summing yields

$$\mathbb{E}_\Pi[d_{H,w}(X, Y)] = \sum_i w_i \Pr_{\Pi_i}(X^i \neq Y^i) \geq \sum_i w_i\,\mathrm{TV}(p^i, q^i). \tag{34}$$

For achievability, choose any maximal coupling $\pi_i^\star \in \Gamma(p^i, q^i)$ for every coordinate and form the product coupling

$$\Pi^\star((x^1, \dots, x^D), (y^1, \dots, y^D)) = \prod_{i=1}^{D} \pi_i^\star(x^i, y^i).$$

Its first marginal is $\otimes_i p^i = p$ and its second is $\otimes_i q^i = q$. Each projected pair has mismatch probability $\mathrm{TV}(p^i, q^i)$, so $\Pi^\star$ attains equality in Eq. (34). This proves Eq. (16).
The product structure ensures that this construction has full marginals p and q. For general sequence marginals, the same lower bound holds, and equality requires a full-sequence coupling whose coordinate projections are all maximal couplings. Moreover, the proposition minimizes expected weighted Hamming distance rather than whole-sequence mismatch $\Pr(X \neq Y)$. The minimum of the latter is $\mathrm{TV}(p, q)$, which generally differs from the coordinatewise sum. □

Conditional MAFTC exactness. Fix an endpoint a and abbreviate $p^i = p_t^i(\cdot \mid a^i)$ and $q^i = p_s^i(\cdot \mid a^i)$. Form the product kernel $K^a = \otimes_i K_i^{\mathrm{MA}}$, meaning $K^a(y \mid x) = \prod_i K_i^{\mathrm{MA}}(y^i \mid x^i)$. For every $y \in \mathcal{T}^D$, finite-sum factorization gives

$$\begin{aligned}(p_t(\cdot \mid a)K^a)(y) &= \sum_{x \in \mathcal{T}^D} \prod_i p^i(x^i) K_i^{\mathrm{MA}}(y^i \mid x^i) \\ &= \prod_i \sum_{x^i \in \mathcal{T}} p^i(x^i) K_i^{\mathrm{MA}}(y^i \mid x^i) \\ &= \prod_i q^i(y^i) = p_s(y \mid a).\end{aligned} \tag{35}$$

Thus the conditional output marginal equals $p_s(\cdot \mid a)$ for any interval length. Each induced coordinate coupling is maximal, so the product coupling attains the lower bound in Proposition 3.4 for every choice of positive weights. This proves the conditional sequence claim in Section 3.4. □

A.5 Proof of Proposition 4.1

Proof. Fix an interior time $t \in (0, 1)$ and suppress the subscript. Let $g = [\dot{p}]_+$ and $\ell = [-\dot{p}]_+$. Differentiating $\sum_x p_t(x) = 1$ gives $\sum_x \dot{p}(x) = 0$, and therefore

$$\sum_x g(x) = \sum_x \ell(x) = \frac{1}{2} \sum_x |\dot{p}(x)| =: M. \tag{36}$$

For any admissible flux j, define the total inflow and outflow of state y by

$$I(y) = \sum_{x \neq y} j(y, x), \qquad O(y) = \sum_{x \neq y} j(x, y).$$

Our indexing convention is that $j(y, x)$ denotes flux from source x to destination y, so the continuity equation is $\dot{p}(y) = I(y) - O(y)$. On a gain state, $\dot{p}(y) > 0$ and hence $I(y) = \dot{p}(y) + O(y) \geq \dot{p}(y)$. Summing only over gain states gives

$$\mathcal{A}(j) = \sum_y I(y) \geq \sum_{y : \dot{p}(y) > 0} I(y) \geq \sum_{y : \dot{p}(y) > 0} \dot{p}(y) = M. \tag{37}$$

For $M > 0$, define

$$j^{\mathrm{MA}}(y, x) = \frac{\ell(x) g(y)}{M}.$$

The supports of g and ℓ are disjoint, so the flux is off diagonal. Summing over destinations gives outgoing flux

$$\sum_{y \neq x} j^{\mathrm{MA}}(y, x) = \ell(x) \frac{\sum_y g(y)}{M} = \ell(x),$$

and summing over sources gives incoming flux $g(y)$. Hence the induced derivative is $g(y) - \ell(y) = \dot{p}(y)$ and the total activity is $\sum_x \ell(x) = M$. At the interior time under consideration, $p_t(x) = 0$ implies $\dot{p}_t(x) = 0$ because $p_t(x)$ is nonnegative and differentiable. Thus $\ell(x) = 0$ on zero-probability states, and the constructed flux satisfies the support condition for admissibility. This attains Eq. (37). If $M = 0$, then $\dot{p} = 0$ and the zero flux is admissible and optimal.

It remains to characterize equality. The first inequality in Eq. (37) is an equality exactly when $I(y) = 0$ for every loss or neutral state. The second is an equality exactly when $O(y) = 0$ for every gain state, because $I(y) = \dot{p}(y) + O(y)$. The continuity equation then gives

$$O(x) = \ell(x) \quad (x \in \{\dot{p} < 0\}), \qquad I(y) = g(y) \quad (y \in \{\dot{p} > 0\}),$$

and all other inflows and outflows vanish. Thus every minimizing flux is supported from the loss set to the gain set and has statewise outflow ℓ and inflow g. For $M > 0$, the array $\gamma(x, y) := j(y, x)/M$ has row marginal ℓ/M and column marginal g/M. This identifies the minimizing fluxes with the couplings in $\Gamma(\ell/M, g/M)$. The flux in Eq. (21) corresponds to the independent coupling.
Finally, differentiability on the finite state space gives

$$\left\| \frac{p_{t+h} - p_t}{h} - \dot{p}_t \right\|_1 \longrightarrow 0 \qquad (h \downarrow 0).$$

Continuity of the norm then yields

$$\frac{\mathrm{TV}(p_t, p_{t+h})}{h} = \frac{1}{2} \left\| \frac{p_{t+h} - p_t}{h} \right\|_1 \longrightarrow \frac{1}{2} \|\dot{p}_t\|_1 = M. \tag{38}$$

To obtain the coupling limit, write $r_h^- = [p_t - p_{t+h}]_+$, $r_h^+ = [p_{t+h} - p_t]_+$, and $\tau_h = \mathrm{TV}(p_t, p_{t+h})$. The positive-part map is 1-Lipschitz, since $|[a]_+ - [b]_+| \le |a - b|$. Applying this inequality pointwise gives

$$\left\| \frac{r_h^-}{h} - \ell \right\|_1 \longrightarrow 0, \qquad \left\| \frac{r_h^+}{h} - g \right\|_1 \longrightarrow 0.$$

When $M > 0$, $\tau_h/h \to M$ ensures that $\tau_h > 0$ for sufficiently small h, and hence

$$\frac{r_h^-}{\tau_h} \longrightarrow \frac{\ell}{M}, \qquad \frac{r_h^+}{\tau_h} \longrightarrow \frac{g}{M} \quad \text{in } \ell^1.$$

For $x \neq y$, Eq. (14) now gives

$$\frac{\pi^{\mathrm{MA}}_{t,t+h}(x, y)}{h} = \frac{[r_h^-(x)/h][r_h^+(y)/h]}{\tau_h/h} \longrightarrow \frac{\ell(x) g(y)}{M} = j^{\mathrm{MA}}(y, x).$$

Since there are finitely many off-diagonal pairs, the sum of their absolute errors also vanishes:

$$\sum_{x \neq y} \left| \frac{\pi^{\mathrm{MA}}_{t,t+h}(x, y)}{h} - j^{\mathrm{MA}}(y, x) \right| \longrightarrow 0.$$

This is the claimed ℓ^1 convergence of the off-diagonal arrays.

□

B Metric-Induced Paths and Kinetic Localization

We identify the finite-interval loss and gain threshold for a Gibbs path, then prove the localization of kinetic-optimal excess activity. Fix an endpoint token $a \in \mathcal{T}$ and a finite nonnegative endpoint-distance score $\rho(\cdot, a)$. For inverse temperature $\beta \ge 0$, let $Z_a(\beta)$ be the normalizing constant of the conditional path

$$p_\beta(x \mid a) = \frac{e^{-\beta \rho(x,a)}}{Z_a(\beta)}, \qquad Z_a(\beta) := \sum_{z \in \mathcal{T}} e^{-\beta \rho(z,a)}. \tag{39}$$

The calculations below use finite β, for which every state has positive probability. We suppress the coordinate index and abbreviate $p_\beta(\cdot \mid a)$ to p_β.
For inverse temperatures $0 \le \beta_t < \beta_s < \infty$ at two sampling times, let $m(\beta) := \sum_x p_\beta(x) \rho(x, a)$ be the mean endpoint distance under the conditional marginal at parameter β. The loss and gain sets are separated by the threshold

$$\rho^*_{t,s} = \frac{\log Z_a(\beta_t) - \log Z_a(\beta_s)}{\beta_s - \beta_t} = \frac{1}{\beta_s - \beta_t} \int_{\beta_t}^{\beta_s} m(\beta)\, d\beta. \tag{40}$$

Lemma B.1 characterizes the gain and loss sets and proves the shrinking-interval limit

$$\lim_{\beta_s \downarrow \beta_t} \rho^*_{t,s} = m(\beta_t). \tag{41}$$

Kinetic activity around the mean-distance threshold. At a fixed finite β, write $\rho_x := \rho(x, a)$ and $\mu := m(\beta) = \sum_x p_\beta(x)\rho_x$. Since $\partial_\beta \log Z_a(\beta) = -\mu$, differentiating the Gibbs path gives

$$\partial_\beta p_\beta(x) = p_\beta(x)(\mu - \rho_x). \tag{42}$$

For the kinetic-optimal construction in Shaul et al. (2025, Eq. (26)), also used by FUDOKI (Wang et al., 2025b), the flux in our destination-first convention is therefore

$$j_\beta^{\mathrm{KO}}(y, x) = \big[p_\beta(x)\partial_\beta p_\beta(y) - p_\beta(y)\partial_\beta p_\beta(x)\big]_+ = p_\beta(x)p_\beta(y)[\rho_x - \rho_y]_+, \qquad x \neq y.$$

For a differentiable nondecreasing schedule with finite β_t, returning to model time t multiplies this flux and both activities below by $\dot\beta_t \geq 0$, leaving their decomposition unchanged. Let $R = \rho(X, a)$ and $R' = \rho(X', a)$ for independent $X, X' \sim p_\beta$, with $\mu = \mathbb{E}R$. The activities of the kinetic-optimal and minimum-activity fluxes are

$$\mathcal{A}_{\mathrm{KO}} = \tfrac{1}{2}\mathbb{E}|R - R'|, \qquad \mathcal{A}_{\mathrm{MA}} = \tfrac{1}{2}\mathbb{E}|R - \mu|. \tag{43}$$

Both fluxes induce the same derivative of the Gibbs probability path.
Theorem 4.2 states the localization result in the main text. In the notation above, its within-group identity is equivalently

$$\mathcal{A}_{\mathrm{KO}} - \mathcal{A}_{\mathrm{MA}} = \mathbb{E}[(R - R')_+\mathbf{1}\{R' > \mu\}] + \mathbb{E}[(R - R')_+\mathbf{1}\{R \leq \mu\}]. \tag{44}$$

The first term is the flux among states whose distances both exceed μ; the second is the flux among states at or below μ. The proof below also establishes the cross-threshold identity.

Lemma B.1 (Threshold structure of Gibbs residuals)**.** *Let $p_\beta(x \mid a) \propto e^{-\beta\rho(x,a)}$ and $0 \leq \beta_t < \beta_s < \infty$. With $\rho^*_{t,s}$ defined in Eq. (40),*

$$\begin{aligned} p_{\beta_s}(x \mid a) > p_{\beta_t}(x \mid a) &\iff \rho(x, a) < \rho^*_{t,s}, \\ p_{\beta_s}(x \mid a) = p_{\beta_t}(x \mid a) &\iff \rho(x, a) = \rho^*_{t,s}, \\ p_{\beta_s}(x \mid a) < p_{\beta_t}(x \mid a) &\iff \rho(x, a) > \rho^*_{t,s}. \end{aligned}$$

Moreover, $\rho^*_{t,s}$ is the average of $m(\beta)$ over $[\beta_t, \beta_s]$.

B.1 Proof of Lemma B.1

Proof. Let $\beta_s > \beta_t$ and write $\Delta\beta = \beta_s - \beta_t$. The ratio of the two Gibbs probabilities satisfies

$$\log \frac{p_{\beta_s}(x \mid a)}{p_{\beta_t}(x \mid a)} = -\Delta\beta\, \rho(x, a) + \log Z_a(\beta_t) - \log Z_a(\beta_s). \tag{45}$$

Therefore $p_{\beta_s}(x \mid a) > p_{\beta_t}(x \mid a)$ exactly when $\rho(x, a) < \rho^*_{t,s}$, equality holds at the threshold, and states with distance above the threshold lose mass. This proves the single-threshold separation in Eq. (40).
To obtain its integral form, differentiate the finite sum defining the partition function:

$$\partial_\beta Z_a(\beta) = -\sum_x \rho(x, a)e^{-\beta\rho(x,a)}, \qquad \partial_\beta \log Z_a(\beta) = \frac{\partial_\beta Z_a(\beta)}{Z_a(\beta)} = -m(\beta).$$

The fundamental theorem of calculus then gives

$$\log Z_a(\beta_t) - \log Z_a(\beta_s) = \int_{\beta_t}^{\beta_s} m(\beta)\, d\beta. \tag{46}$$

Dividing by $\Delta\beta$ yields Eq. (40). Continuity of $m(\beta)$ on the finite discrete state space gives $\rho^*_{t,s} \to m(\beta_t)$ as $\beta_s \downarrow \beta_t$. By the residual-coupling characterization in Eq. (32), every minimum-activity off-diagonal transition moves mass from above the finite-interval threshold to below it and strictly decreases distance to a. □

For the generative endpoint, assume $\rho(a, a) = 0$ and $\rho(x, a) > 0$ for $x \neq a$. Then

$$p_\beta(a \mid a) = \frac{1}{1 + \sum_{x \neq a} e^{-\beta\rho(x,a)}} \longrightarrow 1 \qquad (\beta \to \infty).$$

We use this limit to define the terminal marginal $p_\infty(\cdot \mid a) = \delta_a$.

B.2 Proof of Theorem 4.2

Proof. Fix a finite β and use $\rho_x := \rho(x, a)$ and $\mu := \sum_x p_\beta(x)\rho_x$. Let $X \sim p_\beta$ and set $R := \rho(X, a)$. By Eq. (42), states below the mean distance gain mass, those above it lose mass, and those at the mean have zero probability derivative. Proposition 4.1 therefore gives the minimum expected jump rate with respect to β:

$$\mathcal{A}_{\mathrm{MA}} = \sum_{\rho_x > \mu} p_\beta(x)(\rho_x - \mu) = \mathbb{E}[(R-\mu)_+] = \tfrac{1}{2}\mathbb{E}|R-\mu|, \tag{47}$$

where the last equality uses $\mathbb{E}(R-\mu) = 0$.

We next verify that the kinetic-optimal flux induces the same derivative. For each state y, let $I_{\mathrm{KO}}(y) := \sum_{x \neq y} j^{\mathrm{KO}}_\beta(y, x)$ and $O_{\mathrm{KO}}(y) := \sum_{x \neq y} j^{\mathrm{KO}}_\beta(x, y)$ denote its inflow and outflow. Their difference is

$$\begin{aligned} I_{\mathrm{KO}}(y) - O_{\mathrm{KO}}(y) &= p_\beta(y) \sum_x p_\beta(x)\big([\rho_x - \rho_y]_+ - [\rho_y - \rho_x]_+\big) \\ &= p_\beta(y) \sum_x p_\beta(x)(\rho_x - \rho_y) \\ &= p_\beta(y)(\mu - \rho_y) = \partial_\beta p_\beta(y), \end{aligned}$$

where the identity $[u]_+ - [-u]_+ = u$ is used in the second line and Eq. (42) in the last. Hence j^{KO}_β is an admissible flux for the Gibbs path.

Let $X' \sim p_\beta$ be independent of X and set $R' := \rho(X', a)$. Regard X as the source and X' as the destination. Summing the flux over ordered pairs gives

$$\mathcal{A}_{\mathrm{KO}} = \mathbb{E}[(R - R')_+].$$

Exchangeability implies that $(R-R')_+$ and $(R'-R)_+$ have the same expectation and sum to $|R-R'|$, proving the first identity in Eq. (43).

Now partition the event $R > R'$ into three disjoint classes:

$$\{R > \mu \geq R'\}, \qquad \{R > R' > \mu\}, \qquad \{\mu \geq R > R'\}. \tag{48}$$

The first class is precisely the flux crossing the mean-distance threshold from its upper side to its lower side. If $R = \mu$ almost surely, both activities vanish and the result is immediate. Otherwise let $\omega_H = \Pr(R > \mu)$, $\omega_L = \Pr(R \leq \mu)$, $m_+ = \mathbb{E}[R \mid R > \mu]$, and $m_- = \mathbb{E}[R \mid R \leq \mu]$. Then $\omega_H, \omega_L > 0$, $\omega_H + \omega_L = 1$, and independence gives

$$\begin{aligned} \mathbb{E}[(R-R')\mathbf{1}\{R > \mu \geq R'\}] &= \omega_H \omega_L (m_+ - m_-) \\ &= \omega_H(m_+ - \mu) \\ &= \mathbb{E}[(R-\mu)_+], \end{aligned} \tag{49}$$

because $\mu = \omega_H m_+ + \omega_L m_-$. Equations (47) and (49) prove that the total mean-crossing kinetic flux equals the minimum activity. The other two classes contribute

$$\mathbb{E}[(R-R')_+\mathbf{1}\{R' > \mu\}] \quad \text{and} \quad \mathbb{E}[(R-R')_+\mathbf{1}\{R \leq \mu\}],$$

respectively. Summing the three disjoint contributions from Eq. (48) and subtracting the mean-crossing term proves Eq. (27). The first remainder lies entirely above the mean; the second lies at or below the mean. □

C Conditional and Posterior-Marginalized Updates

Let $A \sim p_1$ be the data endpoint and fix $0 \leq t < s \leq 1$. Suppose $p_t(\cdot \mid a)K^a_{t,s} = p_s(\cdot \mid a)$ for every a. For $p_t(x) > 0$, the exact endpoint posterior is $p_{1|t}(a \mid x) = p_1(a)p_t(x \mid a)/p_t(x)$. On states with $p_t(x) = 0$, choose any probability distribution over endpoints. Define

$$K^\star_{t,s}(y \mid x) = \sum_a p_{1|t}(a \mid x) K^a_{t,s}(y \mid x).$$

The mixture identity $p_t(x) = \sum_a p_1(a)p_t(x \mid a)$ implies that every summand is zero when $p_t(x) = 0$. Thus Bayes' identity $p_t(x)p_{1|t}(a \mid x) = p_1(a)p_t(x \mid a)$ holds on all input states and yields

$$\begin{aligned} (p_t K^\star_{t,s})(y) &= \sum_{x,a} p_1(a)p_t(x \mid a)K^a_{t,s}(y \mid x) \\ &= \sum_a p_1(a)p_s(y \mid a) = p_s(y). \end{aligned} \tag{50}$$

For coordinatewise MAFTC, write $K^a_{t,s}(y \mid x) = \prod_i K^{i,a^i}_{t,s}(y^i \mid x^i)$, where $K^{i,a^i}_{t,s}$ is the conditional kernel for coordinate i. Algorithm 1 samples endpoint tokens independently from q_θ, inducing the parallel transition kernel

$$K^\theta_{t,s}(y \mid x) = \prod_{i=1}^{D} \sum_{a^i \in \mathcal{T}} q_{\theta,i}(a^i \mid x, t) K^{i,a^i}_{t,s}(y^i \mid x^i). \tag{51}$$

Appendix E.7 evaluates this parallel construction with exact token posterior marginals to isolate its finite-step sampling behavior.

D Implementation and Numerical Stability

D.1 Vectorized MAFTC

Each stage evaluates the frozen model once, obtains the learned endpoint posterior at each position, and samples endpoint tokens independently. Model-specific conversion and decoding truncation are described in Appendix E. The sampled endpoint and the probability path determine the adjacent marginals p, q.

For FUDOKI, the implementation forms normalized log probabilities $\ell_p = \log p$ and $\ell_q = \log q$. At the realized source token x with $p(x) > 0$, it computes

$$\kappa_x = \exp\big(\min\{0, \ell_q(x) - \ell_p(x)\}\big).$$

For rows selected to move, the gain residual is evaluated on states with $\ell_q(y) > \ell_p(y)$ as

$$g(y) = e^{\ell_q(y)}\big[-\operatorname{expm1}(\ell_p(y) - \ell_q(y))\big],$$

and set to zero elsewhere, where $\operatorname{expm1}(z) = e^z - 1$ is evaluated accurately near zero. Categorical sampling normalizes these nonnegative weights. This evaluates $[q - p]_+$ without subtracting nearly equal probabilities. Zero-probability source rows use the identity convention; if $p = q$, every state is retained. At the terminal point-mass target, the update directly returns the sampled endpoint.

For text, let $\alpha \in [0, 1]$ parameterize the uniform-source conditional path $p_\alpha(x \mid a) = \alpha\delta_a(x) + (1 - \alpha)/V$, and let α_t and α_s denote its values at successive sampling times. For $0 \le \alpha_t < \alpha_s \le 1$, the gain residual is supported only at a. For $x \neq a$, the MAFTC retention probability is

$$\frac{p_{\alpha_s}(x \mid a)}{p_{\alpha_t}(x \mid a)} = \frac{(1 - \alpha_s)/V}{(1 - \alpha_t)/V} = \frac{1 - \alpha_s}{1 - \alpha_t}.$$

A move therefore goes directly to a, while a token already equal to a remains fixed. Equal interpolation parameters give the identity update.

D.2 Complexity

For batch size B, sequence length D, and vocabulary size V, dense MAFTC uses $O(BDV)$ arithmetic and storage per stage, including the residual tensors, without constructing a $V \times V$ coupling matrix. The uniform-source conditional update reduces to $O(BD)$ once endpoint tokens have been sampled. Both implementations use one neural evaluation per stage.

E Experimental Details and Full Results

E.1 Evaluation quantities

We report sample means of L_{act}, D_{net}, and R_{traj} from Eq. (18). Complete trajectory statistics are included in the accompanying result files.

For each prescribed conditional pair (p, q) and Markov kernel K on the discrete state space, we evaluate

$$\widetilde{q}(y) = \sum_x p(x) K(y \mid x), \qquad A(K) = 1 - \sum_x p(x) K(x \mid x).$$

The resulting A_{pres}, A_{ind}, χ, and ϵ are defined in Eqs. (11)–(12). We sum activity and its components over intervals and average over endpoints. Marginal mismatch is reported as the mean over both endpoints and intervals, denoted $\bar{\epsilon}$. Each conditional evaluation uses the prescribed input distribution p, so $\bar{\epsilon}$ measures one-step marginal mismatch under prescribed input marginals. The conditional activity-to-minimum ratio divides summed expected activity by the summed prescribed minimum. Terminal sample quality is evaluated separately on full rollouts.

For text, the external evaluator p_{eval} is GPT-2 large (Radford et al., 2019).[1] We decode generated tokens with the GPT-2 tokenizer, remove special tokens, and retokenize with the evaluator's tokenizer without adding special tokens. Each sequence is truncated to at most 1,024 evaluator tokens. For a retokenized sequence z_j of length L_j, we score positions $2, \dots, L_j$ conditioned on their prefixes, excluding padding. We report

$$\text{Gen-PPL} = \exp\left[-\frac{1}{M}\sum_j \sum_{m=2}^{L_j} \log p_{\text{eval}}(z_{j,m} \mid z_{j,<m})\right], \qquad M = \sum_j (L_j - 1).$$

Thus corpus negative log-likelihood is weighted by the number of scored tokens, rather than averaged equally across sequences. Lexical metrics use the original generated token IDs with padding removed and EOS retained. We distinguish corpus-level from within-sequence diversity, following the set-level and string-level distinction in Wiher et al. (2022). Pooled unigram entropy is $-\sum_{v=1}^{V} \widehat{f}_v \log \widehat{f}_v$, where $\widehat{f}_v$ is the empirical token frequency across all sequences. Mean sequence unigram entropy averages the corresponding entropy computed separately for each sequence. We use natural logarithms and the convention $0 \log 0 = 0$. Following the Distinct family of measures (Li et al., 2016), pooled Dist-n divides the number of distinct token n-grams across the corpus by the number of within-sequence n-gram positions; no n-grams cross sequence boundaries. Rep-4 averages the within-sequence fraction of repeated four-gram occurrences. Since Distinct scores depend on text length (Liu et al., 2022), we compare corpora with the same number of generated sequences and the same prescribed sequence length.

For the text uncertainty intervals below, we resample shared source indices 10,000 times and report the 2.5th and 97.5th percentiles of paired Gen-PPL differences. Methods share initial states and random-number streams for each source index.

E.2 Full-state evaluation of conditional kernels

For FUDOKI, we select 32 endpoint token IDs uniformly without replacement from the 16,384-entry codebook, using a PyTorch CPU random permutation with seed 1208. We hold this set fixed across methods and evaluate all codebook states on uniform grids with 8, 16, or 32 steps. For endpoint a and interval $[t_k, t_{k+1}]$, the prescribed pair is $p = p_{t_k}(\cdot \mid a)$ and $q = p_{t_{k+1}}(\cdot \mid a)$. The path uses $\beta_t = 3(t/(1-t))^{0.9}$ and squared cosine-distance geometry $\rho(x, a) = (2 - 2\langle \bar{e}_x, \bar{e}_a \rangle)^2$, where $\bar{e}$ denotes a normalized codebook embedding. We evaluate each implemented conditional kernel over the full state space in double precision using the original codebook distances. This directly yields the induced output marginal and the activity decomposition, without sampled histograms. All methods use the same fixed endpoint token for each conditional pair and reveal it directly in the final interval.

Table 2: **FUDOKI conditional activity decomposition and marginal mismatch.** Activity columns sum over intervals and average over 32 endpoints. The final column averages one-step marginal mismatch in total variation over both intervals and endpoints. Thus $\sum_k A_k = \sum_k A_{\text{ind},k} + \sum_k \chi_k$, while $\bar{\epsilon}$ is measured separately. MAFTC's zero entries are rounded numerical values.

NFE	Method	$\sum_k A_{\text{pres},k}$	$\sum_k A_k$	$\sum_k A_{\text{ind},k}$	$\sum_k \chi_k$	$\bar{\epsilon}$
8	Euler	2.8126	2.2937	1.9113	0.3824	0.11830
8	TC	2.8126	3.1372	2.5497	0.5875	0.03775
8	MAFTC	2.8126	2.8126	2.8126	0.0000	0.00000
16	Euler	2.9849	2.8166	2.2531	0.5635	0.04810
16	TC	2.9849	3.4888	2.7557	0.7331	0.01595
16	MAFTC	2.9849	2.9849	2.9849	0.0000	0.00000
32	Euler	3.0186	3.0314	2.2823	0.7491	0.02419
32	TC	3.0186	3.6474	2.7613	0.8860	0.00865
32	MAFTC	3.0186	3.0186	3.0186	0.0000	0.00000

At 32 NFE, Euler's summed activity is 3.031, close to the prescribed minimum of 3.019. Its induced marginal transport requires only 2.282 changes, leaving 0.749 in coupling excess. The excess nearly

[1] https://huggingface.co/openai-community/gpt2-large

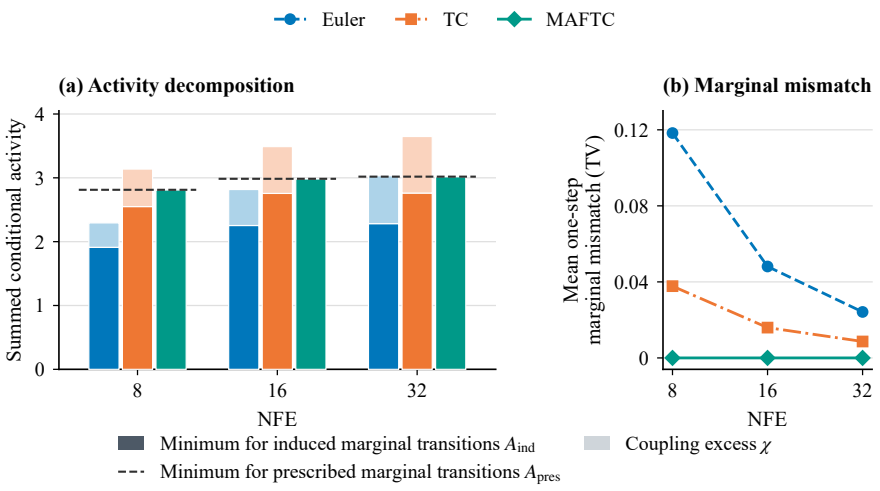


Figure 2: **Conditional activity and marginal mismatch in FUDOKI.** (a) Activity decomposition into the minimum for the induced marginal transition A_{ind} and coupling excess χ, compared with the minimum for the prescribed marginal transition A_{pres}. Each activity component is summed over sampling intervals and averaged over the 32 endpoints. (b) Marginal mismatch averaged over both intervals and endpoints, $\bar{\epsilon}$, across methods and NFEs. Values are reported in Table 2.

offsets the gap between the induced and prescribed minima: total activity alone obscures the different contributions. TC reduces one-step marginal mismatch relative to Euler at all three evaluated NFEs, but has greater coupling excess.

For Duo, we use the uniform-source conditional path and closed-form ancestral transition specified below. Duo+DCD's conditional activity-to-minimum ratios are 2.58590, 3.29904, and 3.98325 at 8, 16, and 32 NFE, respectively; MAFTC's are one, and both kernels realize $pK = q$.

E.3 Integrated expected jump rate

For the same 32 FUDOKI endpoints, we integrate the minimum and kinetic-optimal activities in Eq. (43) along each full conditional path, with inverse temperature β ranging from zero to infinity. Their endpoint-averaged integrals are 3.0381 and 4.0124, respectively. By Theorem 4.2, the difference, 0.9743, is the integrated within-group flux. Its ratio to the integrated kinetic-optimal activity is 24.3%, computed from endpoint-averaged integrals. We use 2,048-interval composite Simpson quadrature in $\log(1+\beta)$ up to $\beta \approx 1964.98$; the remaining tail is bounded by 6.2×10^{-15}.

To bound the omitted tail, fix an endpoint a, let $m_a := \min_x \rho(x, a)$, and write $\widetilde{\rho}_x := \rho(x, a) - m_a$ and $\widetilde{Z}_a(\beta) := \sum_x e^{-\beta \widetilde{\rho}_x}$. Subtracting m_a leaves the Gibbs probabilities unchanged. For $X, X' \sim p_\beta$ independently, nonnegativity gives

$$\mathcal{A}_{\text{KO}}(\beta) = \tfrac{1}{2}\mathbb{E}|\widetilde{\rho}_X - \widetilde{\rho}_{X'}| \le \mathbb{E}\widetilde{\rho}_X, \qquad \mathcal{A}_{\text{MA}}(\beta) = \tfrac{1}{2}\mathbb{E}|\widetilde{\rho}_X - \mathbb{E}\widetilde{\rho}_X| \le \mathbb{E}\widetilde{\rho}_X.$$

The evaluated paths have a unique minimum-distance state, so $\widetilde{Z}_a(\beta) \to 1$ as $\beta \to \infty$. Since $-\partial_\beta \log \widetilde{Z}_a(\beta) = \mathbb{E}\widetilde{\rho}_X$, either activity satisfies

$$\int_{\beta_{\max}}^{\infty} \mathcal{A}_\bullet(\beta)\, d\beta \le \int_{\beta_{\max}}^{\infty} \mathbb{E}\widetilde{\rho}_X\, d\beta = \log \widetilde{Z}_a(\beta_{\max}), \qquad \bullet \in \{\text{MA}, \text{KO}\}.$$

Across the 32 endpoints, the maximum omitted-tail bound is 6.16×10^{-15}. The maximum change in the retained-interval integral between the final two nested Simpson grids is 3.26×10^{-6}, which we use to assess numerical convergence.

E.4 Finite-step baseline definitions

The Euler sampler in the released FUDOKI implementation (Wang et al., 2025b) uses a coordinatewise exponential update with frozen rates. After drawing endpoint token a, it freezes the outgoing kinetic-optimal rates $u_{t_k}(\cdot, x \mid a)$ over a step of width $h = t_{k+1} - t_k$ and samples at most one coordinate

jump:

$$\Pr(Y \neq x \mid x, a) = 1 - e^{-h\lambda_k}, \qquad \lambda_k = \sum_{z \neq x} u_{t_k}(z, x \mid a),$$

$$\Pr(Y = y \mid Y \neq x, x, a) = \frac{u_{t_k}(y, x \mid a)}{\lambda_k}.$$

The destination distribution is used only when $\lambda_k > 0$; if $\lambda_k = 0$, the update retains x. The time-corrected (TC) sampler uses the same endpoint-conditioned rate geometry but integrates its schedule dependence across the interval while keeping the source state fixed (Wan et al., 2026). Our implementation uses four left-endpoint quadrature nodes to approximate the first-jump distribution, with at most one jump per coordinate per step. Euler and TC both use the implementation's convention $\dot{\beta}_0 = 0$. TC corrects the frozen-time approximation without an additional model evaluation.

For Duo+DCD (Sahoo et al., 2025), we use the uniform-source conditional path $p_\alpha(z \mid a) = \alpha\delta_a(z) + (1-\alpha)/V$. Let $0 \leq \alpha_k < \alpha_{k+1} \leq 1$ denote adjacent interpolation values and set $r_k = \alpha_k/\alpha_{k+1}$. For one coordinate, the conditional ancestral bridge is

$$B_k^a(y \mid x) = \frac{p_{\alpha_{k+1}}(y \mid a)}{p_{\alpha_k}(x \mid a)} \left[r_k \mathbf{1}\{x = y\} + \frac{1 - r_k}{V} \right]. \tag{52}$$

For the full current sequence x, the update averages this bridge over the learned endpoint posterior at each coordinate:

$$K_k^{\mathrm{Duo+DCD},i}(y^i \mid x) = \sum_{a^i \in \mathcal{T}} q_{\theta,i}(a^i \mid x, \alpha_k) B_k^{a^i}(y^i \mid x^i).$$

We parameterize model predictions by the interpolation value α_k at time t_k; Appendix E.6 specifies its grid. For both Duo+DCD and MAFTC, the endpoint probabilities are obtained from the truncated denoiser probabilities $b_{\theta,i}$ by

$$q_{\theta,i}(a^i \mid x, \alpha) \propto b_{\theta,i}(a^i \mid x, \alpha) \left[\alpha \mathbf{1}\{a^i = x^i\} + \frac{1 - \alpha}{V} \right],$$

with normalization over $a^i \in \mathcal{T}$. This is the learned endpoint posterior for the i-th token used in Algorithm 1.

For the Ψ-sampler baseline, we use the posterior mixture of Deschenaux et al. (2026, Eq. (12)). The implemented update for coordinate i is

$$K_k^{\Psi,i}(y^i \mid x) = \omega_k K_k^{\mathrm{Duo+DCD},i}(y^i \mid x) + (1 - \omega_k) \left[\alpha_{k+1} q_{\theta,i}(y^i \mid x, \alpha_k) + \frac{1 - \alpha_{k+1}}{V} \right].$$

We fix the implementation's rescale parameter to $\eta = 0.05$. For nonterminal steps, an auxiliary correction parameter σ_k determines the mixture weight ω_k through

$$\sigma_k = \eta \min\left\{ \eta, \frac{1 - \alpha_{k+1}}{\alpha_k} \right\}, \qquad \omega_k = \max\left\{ 0, \min\left\{ 1, 1 - \frac{\sigma_k}{1 - \alpha_{k+1}} \right\} \right\}.$$

The final step uses $\omega_{N-1} = 1$. All methods use the same α grid specified below and one model prediction per step; no Ψ hyperparameters are tuned for this evaluation.

E.5 FUDOKI

We use the released `LucasJinWang/FUDOKI` checkpoint,[2] whose transformer has 24 layers, hidden width 2,048, and 16 attention heads. Its vector-quantized (VQ) image tokenizer, VQ-16, downsamples each spatial dimension by a factor of 16 and has 16,384 codes with eight-dimensional embeddings. We generate a 24×24 latent grid (576 tokens), decoded to 384×384 pixels. The model and tokenizer remain frozen. Image tokens are initialized independently from the uniform distribution, and endpoint predictions use the full vocabulary without top-k or nucleus truncation.

Sampling uses $t_k = k/N$ for $k = 0, \ldots, N$, with the Gibbs path and β_t schedule specified in Appendix E.2. We retain FUDOKI's classifier-free guidance (CFG) convention, $\ell_{\mathrm{cfg}} = (1+w)\ell_{\mathrm{cond}} - w\ell_{\mathrm{uncond}}$, where ℓ_{cond} and ℓ_{uncond} are the conditional and unconditional logits and $w = 5$ is the

[2] https://huggingface.co/LucasJinWang/FUDOKI

guidance scale. Reported NFE counts calls to this guided prediction function, so N sampling steps require N guided predictions and $2N$ backbone forward passes for every image sampler.
For each sampler and NFE, we evaluate all 553 GenEval prompts with four images per prompt and generation seed 3407, recording every image's complete token trajectory. GenEval is the macro average of its six task accuracies. Table 3 also reports mean raw image–text cosine similarity from CLIP ViT-B/32 with OpenAI weights (Radford et al., 2021), evaluated over all 2,212 images, and the activity quantities from Appendix E.1.

Table 3: **Complete FUDOKI benchmark.** GenEval and CLIP cosine similarity are higher-is-better; L_{act} and R_{traj} are mean cumulative token changes and trajectory excess. Each configuration contains 2,212 images from 553 prompts and complete per-image trajectories.

NFE	Sampler	GenEval ↑	CLIP ↑	L_{act}	R_{traj}
4	Euler	0.5106	0.3132	1486.0	910.1
	TC	0.5958	0.3245	1844.3	1268.4
	MAFTC	0.5872	0.3222	1671.8	1095.9
8	Euler	0.6903	0.3328	2146.3	1570.5
	TC	0.7221	0.3354	2276.8	1701.0
	MAFTC	0.7122	0.3337	1879.6	1304.2
16	Euler	0.7508	0.3372	2307.4	1732.0
	TC	0.7572	0.3380	2376.3	1801.0
	MAFTC	0.7429	0.3362	1893.9	1320.0
32	Euler	0.7531	0.3382	2407.8	1833.2
	TC	0.7453	0.3376	2450.4	1876.1
	MAFTC	0.7443	0.3371	1882.7	1310.6
64	Euler	0.7528	0.3367	2492.0	1917.9
	TC	0.7452	0.3366	2529.6	1955.4
	MAFTC	0.7438	0.3357	1956.1	1384.4
128	Euler	0.7268	0.3349	2618.0	2044.0
	TC	0.7392	0.3351	2648.5	2074.5
	MAFTC	0.7369	0.3344	2067.0	1495.3

Table 4 reports the six task accuracies underlying GenEval. At NFE 4, MAFTC improves all six categories over Euler. The category trends vary on finer grids: from NFE 64 to 128, MAFTC's color-attribute and counting accuracies increase, while its position and color accuracies decrease.

E.6 Language generation with Duo

We use `s-sahoo/duo-distilled`,[3] the released Duo checkpoint produced by DCD. It has 12 transformer blocks, hidden width 768, 12 attention heads, and a 1,024-token context. Its vocabulary contains the 50,257 GPT-2 byte-pair tokens and one padding token ($V = 50,258$). OpenWebText is the checkpoint's training corpus; our evaluation consists of unconditional generations from independent uniform token strings. Each cell contains 2,048 sequences of length 1,024, using seed 3407 and shared initial states across samplers. Nucleus sampling (Holtzman et al., 2020) with top-$p = 0.9$ is applied to the denoiser probabilities before the endpoint-posterior conversion in Appendix E.4.
The conditional probability path and sampling grid are

$$p_\alpha(x \mid a) = \alpha \mathbf{1}\{x = a\} + \frac{1-\alpha}{V}, \qquad \alpha_k = \delta + \frac{k}{N-1}\big[1 - \delta(1-\delta) - \delta\big] \quad (0 \le k < N),$$

with $\delta = 10^{-3}$ and a final target $\alpha_N = 1$. The model is queried at the N equally spaced entrance values from δ to $1 - \delta(1-\delta)$, using its noise input $-\log \alpha_k$; the last query supplies the terminal reveal. We initialize from uniform token strings and make the first model query at $\alpha_0 = \delta$. We use $N \in \{2, 4, 8, 16, 32, 64, 128\}$, so this protocol performs exactly N neural evaluations, each requiring one backbone forward pass. Table 5 reports generative perplexity and lexical metrics for all three samplers.

[3] https://huggingface.co/s-sahoo/duo-distilled

Table 4: **Complete FUDOKI GenEval task breakdown.** All accuracies are higher-is-better. Overall is the macro average of the six task accuracies. Each configuration contains 2,212 images from all 553 prompts under the protocol of Table 3. MAFTC rows are shaded.

NFE	Sampler	Overall	Single obj.	Two obj.	Counting	Colors	Position	Color attr.
4	Euler	0.5106	0.8563	0.5404	0.3281	0.7314	0.4425	0.1650
	TC	0.5958	0.9156	0.6364	0.3906	0.7846	0.5500	0.2975
	MAFTC	0.5872	0.8750	0.6263	0.4125	0.7846	0.5275	0.2975
8	Euler	0.6903	0.9406	0.7727	0.4500	0.8537	0.6200	0.5050
	TC	0.7221	0.9438	0.8005	0.5125	0.8883	0.6525	0.5350
	MAFTC	0.7122	0.9219	0.7854	0.4719	0.9016	0.6525	0.5400
16	Euler	0.7508	0.9594	0.8308	0.4875	0.9096	0.7050	0.6125
	TC	0.7572	0.9531	0.8788	0.4969	0.9096	0.6750	0.6300
	MAFTC	0.7429	0.9438	0.8056	0.5406	0.8750	0.6925	0.6000
32	Euler	0.7531	0.9594	0.8384	0.4969	0.9016	0.6650	0.6575
	TC	0.7453	0.9500	0.8333	0.4594	0.9043	0.6600	0.6650
	MAFTC	0.7443	0.9500	0.8561	0.4844	0.8803	0.6725	0.6225
64	Euler	0.7528	0.9250	0.8460	0.4844	0.8963	0.6700	0.6950
	TC	0.7452	0.9313	0.8561	0.4813	0.8777	0.6575	0.6675
	MAFTC	0.7438	0.9250	0.8510	0.5000	0.8670	0.6700	0.6500
128	Euler	0.7268	0.9250	0.8283	0.4531	0.8697	0.6425	0.6425
	TC	0.7392	0.9125	0.8258	0.5250	0.8670	0.6925	0.6125
	MAFTC	0.7369	0.9063	0.8434	0.5094	0.8324	0.6375	0.6925

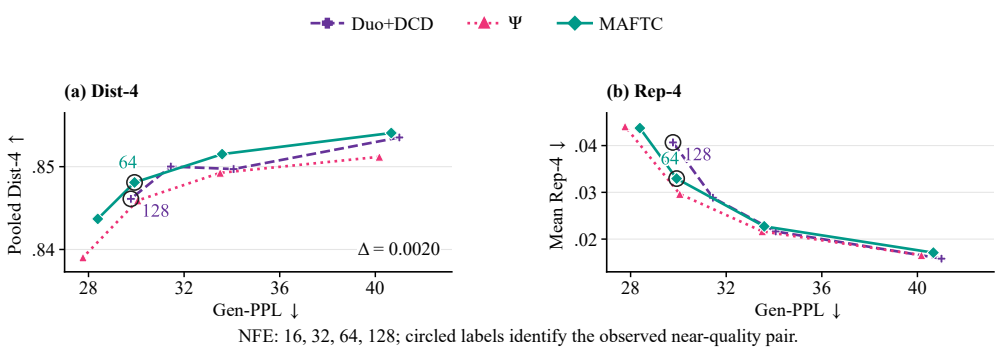
Table 5: **Full text quality and token-diversity results.** H_{pool} and H_{seq} are pooled and mean sequence unigram entropy (nats), D_n is pooled Dist-n, and R_4 is mean sequence Rep-4. Each cell contains 2,048 length-1,024 sequences at seed 3407. Lexical metrics exclude PAD but retain EOS.

NFE	Method	Gen-PPL ↓	H_{pool}	H_{seq}	D_1	D_2	D_4	R_4
2	Duo+DCD	924.20	6.48	5.39	0.00963	0.3752	0.9518	0.0018
	Ψ	922.05	6.48	5.38	0.00963	0.3748	0.9519	0.0017
	MAFTC	923.96	6.48	5.39	0.00963	0.3751	0.9517	0.0018
4	Duo+DCD	156.00	6.39	5.21	0.01613	0.2639	0.8897	0.0051
	Ψ	155.75	6.39	5.21	0.01606	0.2632	0.8894	0.0052
	MAFTC	158.25	6.41	5.23	0.01625	0.2674	0.8915	0.0049
8	Duo+DCD	59.65	6.83	5.38	0.01857	0.2771	0.8624	0.0104
	Ψ	59.37	6.83	5.38	0.01856	0.2767	0.8622	0.0108
	MAFTC	60.64	6.85	5.38	0.01874	0.2808	0.8657	0.0108
16	Duo+DCD	41.01	7.01	5.41	0.01910	0.2858	0.8535	0.0158
	Ψ	40.17	7.00	5.41	0.01898	0.2840	0.8512	0.0166
	MAFTC	40.67	7.03	5.40	0.01931	0.2879	0.8541	0.0171
32	Duo+DCD	34.07	7.12	5.40	0.01937	0.2914	0.8497	0.0216
	Ψ	33.51	7.11	5.40	0.01930	0.2891	0.8493	0.0217
	MAFTC	33.59	7.14	5.38	0.01971	0.2939	0.8515	0.0227
64	Duo+DCD	31.44	7.23	5.38	0.01968	0.2981	0.8500	0.0289
	Ψ	30.06	7.20	5.36	0.01956	0.2932	0.8459	0.0296
	MAFTC	29.93	7.23	5.32	0.01994	0.2972	0.8481	0.0329
128	Duo+DCD	29.77	7.32	5.32	0.01987	0.3027	0.8461	0.0407
	Ψ	27.77	7.31	5.29	0.01980	0.2974	0.8390	0.0441
	MAFTC	28.39	7.31	5.26	0.02005	0.3015	0.8437	0.0437

Figure 3 compares corpus-level four-gram coverage and within-sequence repetition at similar Gen-PPL. Near Gen-PPL 30, MAFTC-64 achieves pooled Dist-4 of 0.8481, compared with 0.8461 for Duo+DCD-128 and 0.8459 for Ψ-64. Relative to Duo+DCD-128, it also reduces mean Rep-4 and uses about 80% fewer token changes.

Table 6: **Complete text trajectory results.** Mean cumulative token changes L_{act} and trajectory excess R_{traj} for the same 2,048 sequences per configuration as Table 5. Net displacement is $D_{\text{net}} = L_{\text{act}} - R_{\text{traj}}$.

	Duo+DCD		Ψ		MAFTC	
NFE	L_{act}	R_{traj}	L_{act}	R_{traj}	L_{act}	R_{traj}
2	1261.5	238.3	1260.6	237.3	1260.5	237.5
4	2052.6	1029.3	2056.5	1033.2	1178.2	198.3
8	2761.2	1738.0	2782.8	1759.5	1114.3	142.8
16	3406.7	2383.6	3465.6	2442.5	1067.2	96.9
32	4043.8	3020.8	4177.3	3154.3	1042.2	71.5
64	4671.9	3649.1	4955.5	3932.5	1031.8	59.4
128	5269.4	4246.7	5849.2	4826.2	1028.8	54.2



NFE: 16, 32, 64, 128; circled labels identify the observed near-quality pair.

Figure 3: **Corpus coverage and within-sequence repetition at similar Gen-PPL.** Pooled Dist-4 and mean sequence Rep-4 versus Gen-PPL at NFE 16, 32, 64, and 128. Lines connect measured configurations, and labels give NFE. Open circles identify MAFTC-64 and Duo+DCD-128 at similar Gen-PPL; their pooled Dist-4 difference is approximately 0.0020.

Four-gram decomposition. For M equal-length sequences, let u_s be the number of distinct four-grams in sequence s, U the number of distinct four-grams across all sequences, and T the total number of four-gram positions. Pooled Dist-4 is $D_4 = U/T$, mean sequence Rep-4 is $R_4 = 1 - \sum_s u_s/T$, and cross-sequence reuse is $C_4 = (\sum_s u_s - U)/T$. Hence $1 - D_4 = R_4 + C_4$. For a comparison with either baseline, let Δ denote the MAFTC-minus-baseline difference. The change in corpus coverage then separates as

$$\Delta D_4 = -\Delta R_4 - \Delta C_4.$$

Figure 4 compares these contributions at similar Gen-PPL, following Figure 3. At three observed Gen-PPL levels near 41, 34, and 30, we pair each MAFTC configuration with the closest measured Gen-PPL of each baseline among NFE 16–128. All six pairs differ in Gen-PPL by less than 1.5%; the figure reports both values and sampling budgets. Near Gen-PPL 41 and 34, MAFTC's higher pooled Dist-4 comes from less cross-sequence reuse despite slightly more within-sequence repetition. Near Gen-PPL 34, these differences coexist with about 75% fewer token changes than either baseline (Table 6).

Near Gen-PPL 30, the source of MAFTC's higher pooled Dist-4 depends on the baseline. Compared with Duo+DCD-128, MAFTC-64 has less within-sequence repetition but more cross-sequence reuse; compared with Ψ-64, the directions are reversed. For the MAFTC-64/Duo+DCD-128 pair, the differences $(\Delta D_4, \Delta R_4, \Delta C_4)$ in percentage points are $(+0.198, -0.773, +0.575)$ at seed 3407 and $(+0.314, -0.823, +0.510)$ at the independent seed 3408; the corresponding Gen-PPL differences are $+0.153$ and $+0.176$. Thus higher corpus four-gram coverage at similar Gen-PPL can arise from different changes in its two components. Trajectory activity records token changes during sampling, whereas R_4 and C_4 describe repetition within and across completed texts. At similar Gen-PPL, lower trajectory activity coexists with different patterns of within- and cross-sequence repetition.

At NFE 8, the MAFTC-minus-Duo+DCD Gen-PPL difference is 0.986 with a 95% paired-bootstrap interval of $[0.331, 1.631]$; at NFE 16, it is -0.340 with interval $[-0.859, 0.161]$.

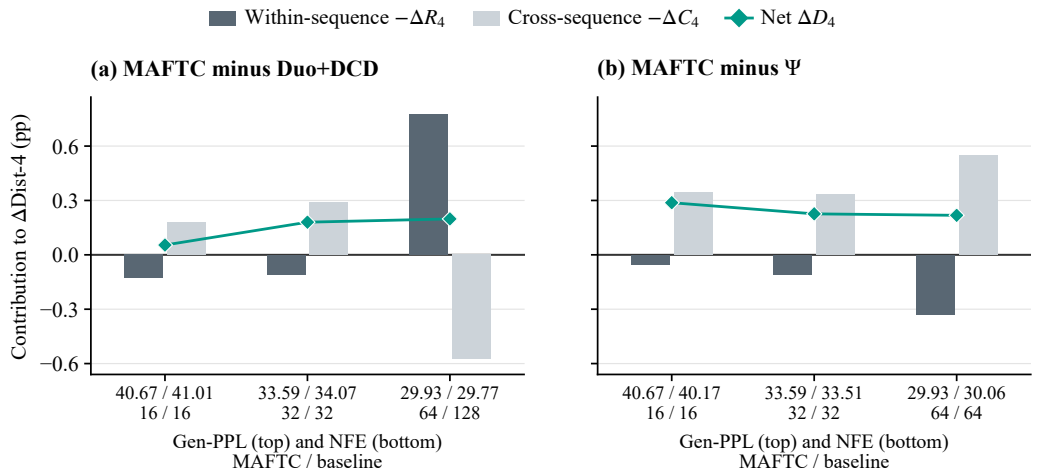


Figure 4: **Sources of pooled Dist-4 differences at similar Gen-PPL.** Differences are MAFTC minus (a) Duo+DCD and (b) Ψ. Each horizontal label gives the measured Gen-PPL pair above its NFE pair, in MAFTC/baseline order. Bars show the contributions from within-sequence repetition ($-\Delta R_4$) and cross-sequence reuse ($-\Delta C_4$); diamonds show their sum, ΔD_4. Values are in percentage points (pp). Positive contributions increase corpus four-gram coverage.

E.7 Exact-posterior evaluation of complete sampling distributions

Targets and conditional paths. We enumerate $\mathcal{X} = \{0,1,2,3\}^3$, containing 64 sequence states. For each seed in $\{0,\dots,7\}$, a NumPy random generator draws a token distribution $b \sim \mathrm{Dirichlet}(2,2,2,2)$, a row-stochastic matrix R with independent $\mathrm{Dirichlet}(1,1,1,1)$ rows, and a uniformly random permutation matrix P. We vary the dependence coefficient $\zeta \in \{0, 0.5, 0.9\}$ and define the target by

$$M_\zeta = (1-\zeta)\mathbf{1}b^\top + \zeta(0.15R + 0.85P), \qquad p_1(a) = b(a^1)M_\zeta(a^2 \mid a^1)M_\zeta(a^3 \mid a^2).$$

Rows of M_ζ index the preceding token. At $\zeta = 0$, the target factorizes as $b^{\otimes 3}$; the other settings introduce coordinate dependence. All samplers share the same target and geometry at each seed.
Both paths factorize conditional on the endpoint. For one coordinate, the uniform-source path is $p_t(x \mid a) = t\mathbf{1}\{x = a\} + (1-t)/4$. The Gibbs path is $p_t(x \mid a) \propto \exp[-\beta(t)\rho(x,a)]$, with $\beta(t) = 3t/(1-t)$. To construct its geometry, we draw four five-dimensional standard Gaussian embeddings from the same generator and normalize them to unit length. We set the diagonal distances to zero and divide $(2-2\langle e_x, e_a\rangle)^2$ by its mean over off-diagonal pairs. The controlled experiment uses this differentiable schedule at $t = 0$, whereas pretrained FUDOKI uses the schedule in Appendix E.2. Both paths begin at the uniform source and end at the sampled data endpoint.

Exact posterior and parallel updates. For each current sequence x, we compute the full endpoint posterior by Bayes' rule and then obtain its token marginals by summation. The implemented parallel kernel is

$$K_{t,s}(y \mid x) = \prod_{i=1}^{3} \sum_{a^i \in \mathcal{T}} p^i_{1|t}(a^i \mid x) K^{a^i}_{t,s}(y^i \mid x^i),$$

where $K^{a^i}_{t,s}$ is the sampler's conditional coordinate kernel. Thus the experiment removes posterior prediction error while retaining finite-step and parallel-update effects. The parallel kernel marginalizes independent coordinatewise endpoint draws from the exact token posteriors. In the final interval, each coordinate is sampled from its current token posterior.
We use uniform time grids $t_k = k/N$, with $N \in \{1,2,4,8,16,32,64,128\}$. On the uniform path, we compare the conditional ancestral bridge, the Ψ mixture with rescale parameter $\eta = 0.05$, and MAFTC. On the Gibbs path, we compare the first-event exponential Euler update, the time-corrected update, and MAFTC. The baseline constructions follow Appendix E.4, with $\alpha_k = t_k$ for the uniform path. For a nonterminal first interval at $t_0 = 0$, we set $\sigma_0 = \eta$; the ancestral and resampling kernels

coincide there, so this choice does not alter the first update. For this Gibbs schedule, $\dot{\beta}(0) = 3$. TC approximates the first-jump distribution using four quadrature nodes while keeping the source state fixed. All methods use one posterior query per grid interval, with no truncation or temperature adjustment. NFE counts posterior queries along a rollout.

Measurements and numerical checks. Starting from the uniform source, we propagate the complete sequence distribution in double precision and report $\mathrm{TV}(\widehat{p}_1, p_1)$, where $\widehat{p}_1$ is the terminal sampler law. Expected cumulative Hamming activity is summed under the evolving sampler law, and net displacement is computed by propagating the joint law of initial and current states. We verify stochastic kernel rows, nonnegative entries, mass conservation, and agreement between the propagated marginal and the initial/current joint law. The sweep contains 1,152 configurations: eight seeds, three dependence settings, two paths, eight NFEs, and three methods per path. Lines show means over the eight target/geometry seeds, and bands show interquartile ranges.

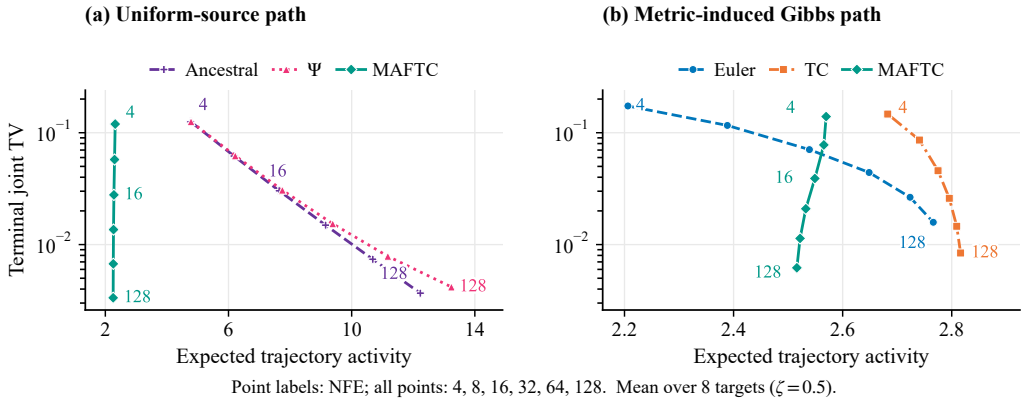


Figure 5: **Terminal accuracy versus trajectory activity with exact posteriors.** Each point averages eight targets with dependence coefficient $\zeta = 0.5$. Lines connect measured NFE 4–128 configurations; annotations indicate selected NFEs. TV compares the complete terminal joint distribution with the target.

Figure 5 summarizes the terminal accuracy–activity relation at moderate dependence. On the uniform path, MAFTC reaches similar terminal TV with much lower activity: at 128 NFE, the seed-mean activity is 2.25, versus 12.23 for ancestral sampling. On the Gibbs path, coarse Euler steps can have lower activity but greater terminal error; at 32–128 NFE, MAFTC lowers both relative to Euler and TC.

Dependence and accumulated error. Figures 6–8 report the full NFE curves. For independent targets, uniform-path ancestral sampling, Ψ, and MAFTC, as well as Gibbs-path MAFTC, attain terminal TV below 3×10^{-14} across the sweep. With dependent targets, exact token posteriors no longer make the parallel finite-step update exact, and mean terminal TV decreases under grid refinement.

Prescribed-input one-step mismatch and terminal TV can rank samplers differently. At seed 0, $\zeta = 0.9$, and two Gibbs steps, MAFTC has a larger summed local TV than TC (0.5526 versus 0.5485), yet a smaller terminal TV (0.3541 versus 0.3646). This reversal in sampler rankings shows that terminal accuracy must be assessed from the propagated output distribution.

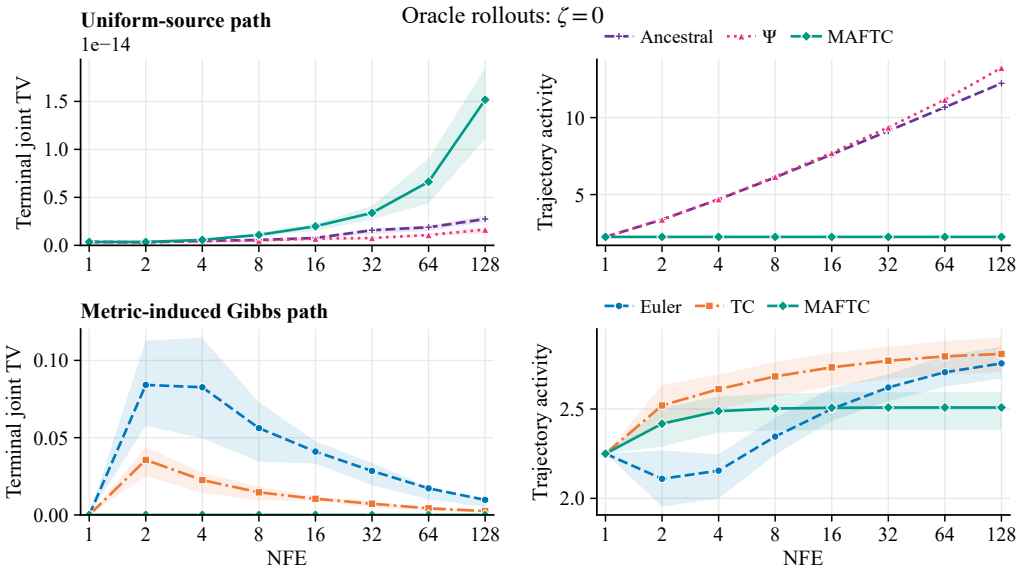


Figure 6: **Exact-posterior sampling with independent target coordinates** ($\zeta = 0$)**.** Rows show the uniform and Gibbs paths; columns show terminal joint TV and expected trajectory activity. Lines are means over eight target/geometry seeds and bands are interquartile ranges. TV uses a linear scale so numerical zeros remain visible.

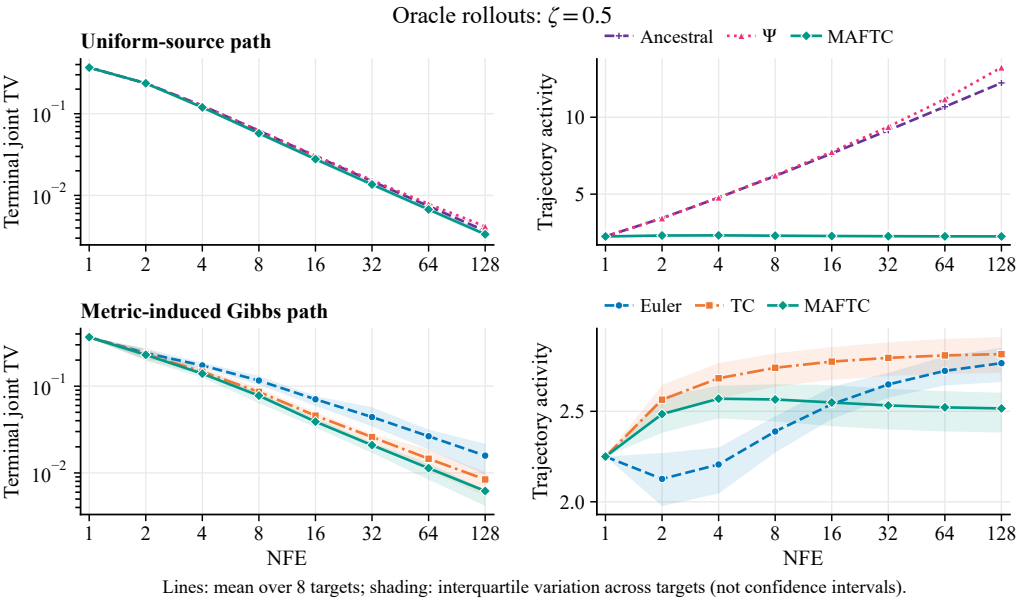


Figure 7: **Exact-posterior sampling with moderate dependence** ($\zeta = 0.5$)**.** Complete NFE curves underlying Figure 5, including NFE 1 and 2. Lines and bands show seed means and interquartile ranges; terminal joint TV uses a logarithmic scale.

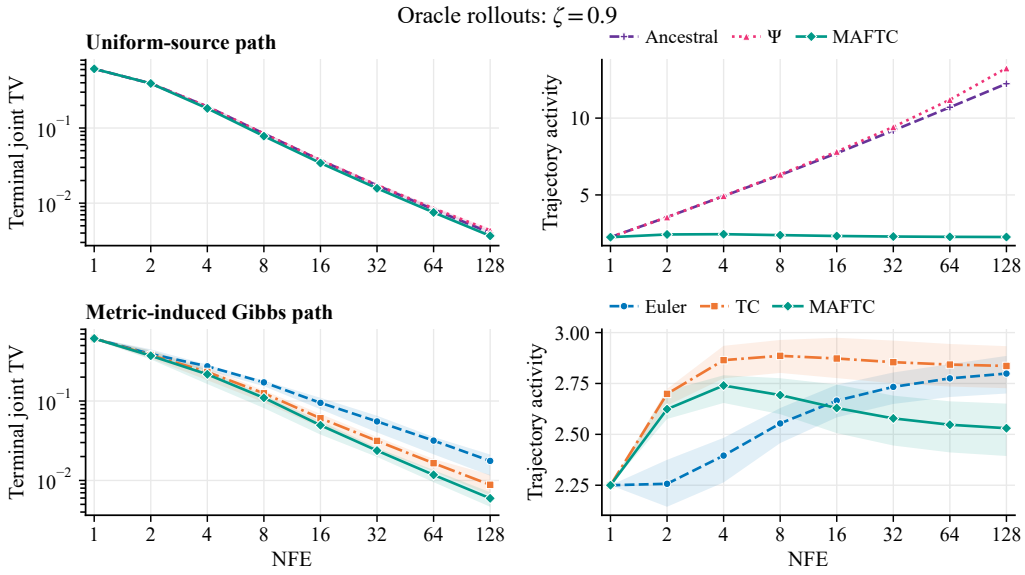


Figure 8: **Exact-posterior sampling with strong dependence** ($\zeta = 0.9$)**.** Lines and bands show means and interquartile ranges over eight targets. The comparison retains the same path-specific methods and grid as the moderate-dependence setting.