NON-MARKOVIAN UNIFORM-STATE DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Uniform discrete diffusion language models (UDLMs) are typically built on a fixed Markov coupling between noise levels, which jointly determines their reverse dynamics, training objective, sampling behavior, and likelihood evaluation. Such a Markov coupling can be restrictive when different tasks require different transition behaviors. We introduce *NoMUD*, a family of non-Markovian UDLMs that makes this coupling explicitly designable through two degrees of freedom while preserving the marginals. Despite inducing different transition dynamics, all feasible couplings share the same optimal predictor, allowing adaptive coupling design during inference. We exploit this freedom in three settings: optimizing the coupling minimizes the negative evidence lower bound for likelihood evaluation; confidence-aware coupling enables adaptive retention, correction, and exploration during sampling; and learnable coupling improves the flexibility of few-step distillation. Together, these results demonstrate that explicitly designing the coupling provides a simple and effective way to enhance the flexibility of UDLMs and unlock their broader potential.

1 INTRODUCTION

Diffusion large language models (dLLMs) have become competitive alternatives to autoregressive (AR) models, owing to their potential for parallel decoding and bidirectional context modeling (Sahoo et al., 2024; Lou et al., 2024; Ou et al., 2025; Nie et al., 2025; Ye et al., 2025). Among them, uniform discrete diffusion language models (UDLMs) have attracted increasing attention due to their potential for few-step generation and iterative self-correction (Schiff et al., 2025; Sahoo et al., 2025; Deschenaux et al., 2026). A standard UDLM progressively corrupts clean tokens into random vocabulary tokens through a Markov forward process and learns a reverse process for generation (Austin et al., 2021; Schiff et al., 2025). Despite its simplicity and tractability, the Markov forward process restricts the flexibility of UDLMs and can consequently affect their performance.

The Markov forward transition determines the coupling between states at different noise levels, i.e., how a state x_s transitions to x_t. Once this coupling is fixed, the corresponding reverse transition and evidence lower bound naturally determine the training objective, sampling dynamics, and likelihood evaluation (Austin et al., 2021; Sahoo et al., 2025). However, a fixed Markov coupling need not be equally suitable for all of these purposes: For example, the value of the evidence lower bound depends on the chosen coupling (Austin et al., 2021), while effective sampling often requires design of the transition dynamics (Deschenaux et al., 2026). Continuous diffusion models provide an instructive precedent: DDIM introduces non-Markovian couplings that allow a fixed pretrained model to use different trajectories at inference time (Song et al., 2021). In UDLMs, however, how to systematically design such couplings remains less explored. This raises a natural question:

Can the coupling between states in UDLMs be designed, and if so, how should it be designed?

We answer this question with a family of **No**n-**M**arkovian **U**niform **D**iffusion Models, **NoMUD**, that explicitly exposes coupling as a design dimension orthogonal to the noise schedule. By introducing non-Markovian dependence on x_0, we show that UDLMs admit two degrees of freedom in coupling design without changing their marginals: one controls probability exchange between the clean token and the current token, while the other controls transitions among random tokens (Figure 1). Compared with previous non-Markovian UDLM special cases with one degree of freedom, this formulation

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

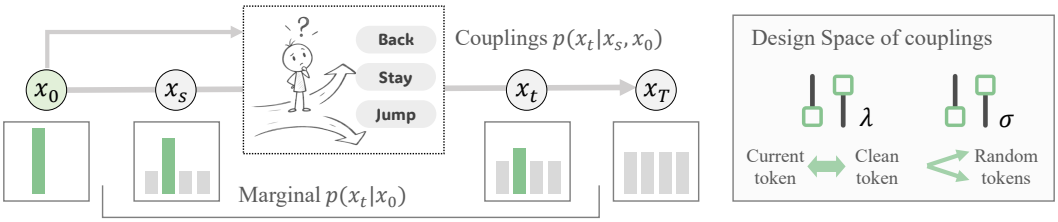


Figure 1: **Overview of NoMUD.** Left: non-Markovian couplings between noise levels are designable while the marginals remain unchanged. Right: the coupling design space has two degrees of freedom. The parameter λ controls probability exchange between the clean and current tokens, while σ controls transitions among random tokens.

explores a broader coupling design space. We use three representative processes to illustrate how different couplings induce different behaviors: a model can favor retaining the current token, making aggressive corrections, or exploring alternatives (Figure 2).

We further show that coupling design is both feasible and consequential. First, different feasible couplings share the same optimal clean-token predictor. This separates two roles that are usually tied together in UDLMs: predicting the clean token and deciding how that prediction should drive the state transition. The former can remain shared across different processes, while the latter can be adapted to the objective. Second, we demonstrate that designed couplings can substantially improve UDLMs across different tasks, with extensive experimental validation.

During likelihood evaluation, we optimize the coupling to minimize the negative evidence lower bound, improving average likelihood evaluation results from 61.14 to 58.50. During sampling, we design confidence-aware couplings that choose among retention, correction, and exploration, reducing matched-entropy 256-step GenPPL from 54.31 to 50.96. Finally, we use learnable coupling parameters to distill pretrained models into efficient few-step generators, reducing 64-step GenPPL from 57.63 to 34.29. This work turns coupling from a fixed property of the forward process into a key design variable for UDLMs. Our contributions are threefold:

1. We introduce a general non-Markovian coupling family for uniform discrete diffusion that preserves the standard UDLM marginals while exposing two degrees of freedom.
2. We derive the non-Markovian reverse process, its parameterization, and the evidence lower bound. We show that different non-Markovian processes share the same optimal denoiser, making coupling an adaptively designable dimension.
3. We exploit this coupling freedom in sampling, likelihood evaluation, and few-step generation, demonstrating that different objectives can benefit from different transition dynamics.

2 Preliminaries

Notation. Let V be the vocabulary size and $\mathbf{u} = \mathbf{1}/V$ the uniform probability vector. The clean token x_0 and noisy token x_t are one-hot vectors, with token indices y and m, respectively. A sequence of length L is written as $\mathbf{x}_t = (x_t^{(1)}, \dots, x_t^{(L)})$; we omit the position index when discussing a single token. The continuously differentiable noise schedule α_t decreases from $\alpha_0 = 1$ to $\alpha_1 = 0$. We write $\beta_t = 1 - \alpha_t$ and $r_t = V\alpha_t + \beta_t$. We identify categorical distributions with their probability vectors and use $\mathbb{1}$ as the indicator.

Uniform discrete diffusion. For $s < t$, let $\delta_{t,s} = 1 - \alpha_t/\alpha_s$. The standard UDLM follows a Markov forward process: the next state depends only on the current state. Its forward transition and marginal are shown in (1). The process acts independently at each sequence position, and its marginal approaches the uniform prior as $t \to 1$.

$$\Pr(x_t \mid x_s) = (1 - \delta_{t,s})x_s + \delta_{t,s}\mathbf{u}, \qquad \Pr(x_t \mid x_0) = \alpha_t x_0 + \beta_t \mathbf{u}. \tag{1}$$

Reverse transition. The forward transition determines the exact reverse kernel $\Pr(x_s \mid x_t, x_0)$ through Bayes' rule. In Duo (Sahoo et al., 2025), it takes the form

$$\Pr(x_s = i \mid x_t = m, x_0) = \frac{\big[(1-\delta_{t,s})\mathbb{1}_{i=m} + \delta_{t,s}/V\big](V\alpha_s x_{0,i} + \beta_s)}{V\alpha_t x_{0,m} + \beta_t}. \tag{2}$$

Duo uses a plug-in parameterization: it replaces x_0 in (2) with the clean-token prediction $\hat{x}_\theta(x_t, t)$. We denote the resulting model transition by $p_\theta(x_s \mid x_t)$.

Training objective. The negative evidence lower bound (NELBO) is built from local KL divergences between the exact and model reverse transitions. The continuous-time contribution is $L_t = \lim_{s\uparrow t}(t-s)^{-1}\,\mathrm{KL}(\Pr(x_s \mid x_t, x_0) \,\|\, p_\theta(x_s \mid x_t))$. Under Duo's plug-in parameterization, this becomes

$$L_t^{\mathrm{Duo}} = -\frac{\alpha_t'}{V\alpha_t} \sum_{i \neq m} \mathcal{D}\left(\frac{V\alpha_t \mathbb{1}_{i=y} + \beta_t}{V\alpha_t \mathbb{1}_{m=y} + \beta_t}, \frac{V\alpha_t \hat{x}_{\theta,i} + \beta_t}{V\alpha_t \hat{x}_{\theta,m} + \beta_t}\right), \tag{3}$$

where $\mathcal{D}(p, q) = q - p + p\log(p/q)$, with $\mathcal{D}(0, q) = q$.

Simplified denoising objectives. SDDLM (Zhu et al., 2025) uses the same plug-in parameterization with simplified losses. Its SDDLM and SDDLM-v1 token-level objectives are, respectively,

$$L_t^{\mathrm{SD}} = -\mathbb{1}_{m \neq y} \log \hat{x}_{\theta,y}, \qquad L_t^{\mathrm{SD1}} = -\mathbb{1}_{m \neq y}\left[\log \hat{x}_{\theta,y} - \frac{1}{V}\sum_{i=1}^{V} \log \hat{x}_{\theta,i}\right]. \tag{4}$$

These simplified objectives are not NELBOs. For each loss, the sequence-level training objective averages the sum of token losses over $t \sim \mathrm{Unif}(0, 1)$ and $\mathbf{x}_t \sim \Pr(\mathbf{x}_t \mid \mathbf{x}_0)$.

3 Non-Markovian Uniform Diffusion Models

We introduce NoMUD through two central contributions. (1) Beyond the noise schedule, NoMUD admits an orthogonal design dimension: how states across noise levels are coupled. (2) Coupling design can benefit training, sampling, likelihood evaluation, and distillation. Section 3.1 introduces the general non-Markovian transition with two coupling degrees of freedom, controlling transitions between the clean and current tokens and among random tokens. Section 3.2 derives the parameterization and training loss, and shows that different couplings share an optimum. Section 3.3 develops adaptive couplings for NELBO minimization, confidence-aware sampling, and few-step distillation.

3.1 Forward and reverse transitions

Forward transition. Here we consider transitions only conditioned on x_0 and impose symmetry on $i \neq m, y$. The general off-diagonal transition can be written as, for $i \neq m$,

$$\Pr(x_t = i \mid x_s = m, x_0) = \delta_{t,s}\left[\mathbb{1}_{m=y}\frac{\alpha_s + \beta_s\lambda_t}{r_s} + \mathbb{1}_{i=y}\lambda_t + \mathbb{1}_{m,i\neq y}\frac{\sigma_t}{V}\right]. \tag{5}$$

The diagonal is determined by normalization, $\Pr(x_t = m \mid x_s = m, x_0) = 1 - \sum_{i\neq m}\Pr(x_t = i \mid x_s = m, x_0)$. Exactly two degrees of freedom remain: λ_t and σ_t. The coefficients can depend on time and clean example x_0. They are shared across token coordinates within a kernel and satisfy

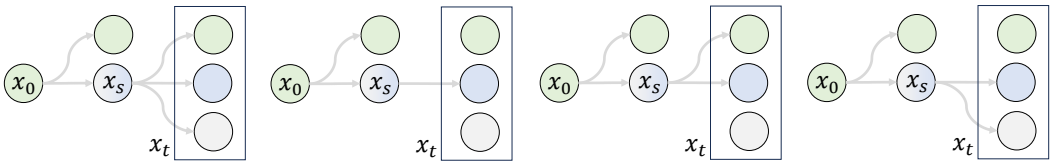
$$0 \le \lambda_t \le \min\{\beta_t^{-1}, r_t(V-1)^{-1}(1-\sqrt{\alpha_t})^{-2}\}, \quad 0 \le \sigma_t \le (\beta_t^{-1} - \lambda_t)(1 - 2V^{-1})^{-1}. \tag{6}$$

We denote the upper bound on λ_t by $\lambda_t^{\max}$, the upper bound on σ_t at $\lambda_t = 0$ by $\sigma_t^{\max} = 1/[(1 - 2/V)\beta_t]$, and the feasible region by $\mathcal{S}_{\lambda,\sigma}(t)$.

Proposition 1 (*Marginal preservation and feasibility*)**.** For any $(\lambda_t, \sigma_t) \in \mathcal{S}_{\lambda,\sigma}(t)$, the transition in (5) is categorical for every $s < t$ and preserves the marginal $q(x_t \mid x_0) = \alpha_t x_0 + \beta_t \mathbf{u}$.

The proof is given in Appendix B.1. Thus, coupling parameters form a design space orthogonal to the noise schedule and determine how states change.

Remark 1 (*Relation to masked diffusion*). An analogous non-Markovian structure exists in masked diffusion language models (MDLMs). Their non-Markovian processes share a weighted cross-entropy loss and only one coupling degree of freedom remains, corresponding to remasking (Appendix B.7). Remasking can improve sampling (Wang et al., 2025), demonstrating the importance of coupling design. In contrast, the two degrees of freedom in non-Markovian UDLMs theoretically permit richer transition behaviors.



(a) General (b) Absorb (c) Flip (d) Explore

Figure 2: Forward transition structures. Green denotes the clean token y, blue the current non-clean token m, and gray another token $i \notin \{m, y\}$. In a general non-Markov process (a), five transition paths (gray arrows) can be designed within the feasible region of two degrees of freedom. (b)–(d) show three specific designs.

Table 1: Representative non-Markovian processes. Forward behavior shows whether a forward transit exists. Sampling tendencies compare $p_\theta(x_s \mid x_t)$ with $p_\theta(x_0 \mid x_t)$. Training penalties are described qualitatively. Appendix B.3 gives more details.

Process	(λ_t, σ_t)	Forward behavior		Sampling behavior	Training penalty	
		$m \to y$	$m \to i$	$p_\theta(x_s \mid x_t)$ vs. $\hat{x}_\theta$	$m = y$	$m \neq y$
Absorb	$(0, 0)$	×	×	More on x_t	Medium	Low on m; high on others
Flip	$(\lambda_t^{\max}, 0)$	✓	×	More on confident tokens	Low	High
Explore	$(0, \sigma_t^{\max})$	×	✓	Flatter distribution	High	Medium on m; low on others

Reverse transition. Bayes' rule gives the exact reverse kernel conditioned on x_0, for $i \neq m$:

$$\Pr(x_s = i \mid x_t = m, x_0) = \delta_{t,s}\left[\mathbb{1}_{m=y}\frac{\beta_s \lambda_t}{r_t} + \mathbb{1}_{i=y}\frac{\alpha_s + \beta_s \lambda_t}{\beta_t} + \mathbb{1}_{m,i\neq y}\frac{\beta_s \sigma_t}{V \beta_t}\right]. \tag{7}$$

Appendix B.1.2 provides the derivation. At sampling time, x_0 is unknown, so we use a parameterized reverse distribution $p_\theta(x_s \mid x_t)$. Consequently, different coupling parameters also change the model's sampling behavior, e.g., its tendency to retain or revise the current token.

Roles of λ and σ. The parameters jointly determine the coupling. Specifically, λ controls transitions between the current token m and the clean token y, while σ controls transitions between the current token and other random tokens i. Intuitively, increasing λ causes intermediate states to visit and leave the clean token more frequently; increasing σ causes them to frequently visit other random tokens.

Three representative examples. To illustrate the role of coupling parameters more clearly, we present three representative non-Markov processes here. Their parameters and characteristics are shown in Table 1. The Absorb process ($\lambda_t = \sigma_t = 0$) favors retaining the current token, the Flip process ($\lambda_t = \lambda_t^{\max}, \sigma_t = 0$) acts more aggressively on confident corrections, and the Explore process ($\lambda_t = 0, \sigma_t = \sigma_t^{\max}$) deliberately spreads probability mass when the prediction is uncertain. The standard Markov UDLM is recovered at $(\lambda_t, \sigma_t) = (1/V, 1)$. Their complementary behaviors motivate coupling designs that combine the advantages of different processes.

3.2 Parameterization and Training

To learn the non-Markovian reverse transition, we parameterize the model and train it with the negative evidence lower bound (NELBO).

Parameterization and sampling. We use a neural network to learn the clean posterior and construct the reverse transition by averaging exact conditional kernels:

$$\hat{x}_\theta(x_t, t) = p_\theta(x_0 \mid x_t), \qquad p_\theta(x_s \mid x_t) = \sum_{x_0} p_\theta(x_0 \mid x_t) q(x_s \mid x_t, x_0). \tag{8}$$

For coefficients shared across clean candidates, expansion yields, for $i \neq m$,

$$\Pr_\theta(x_s = i \mid x_t) = \frac{\delta_{t,s}}{\beta_t}\left[(\alpha_s + \beta_s \lambda_t)\hat{x}_{\theta,i} + \frac{\beta_s \sigma_t}{V}(1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i})\right] + \frac{\delta_{t,s}\beta_s \lambda_t}{r_t}\hat{x}_{\theta,m}. \tag{9}$$

The current-token probability is the remaining mass. Appendix B.1.3 gives the expansion. This parameterization also defines sampling: on a chosen time grid, the model predicts $\hat{x}_\theta(x_t, t)$ and samples the next state from $p_\theta(x_s \mid x_t)$, repeating until time zero. Algorithm 1 gives the sampling procedure.

Training objective. We derive the continuous-time NELBO of the non-Markovian process and use it as the training objective. Under the continuous-time limit and vanishing reconstruction term specified in Appendix B.2, the loss can be written as

$$-\log p_\theta(\mathbf{x}_0) \le \mathrm{NELBO}(\mathbf{x}_0; \lambda, \sigma) = \mathbb{E}_{t\sim\mathrm{Unif}(0,1),\mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}_0)} \sum_{\ell=1}^{L} L_t^{(\ell)}, \tag{10}$$

where $L_t^{(\ell)}$ is the loss at position ℓ. For brevity, we focus on the token-level loss and omit its superscript. It compares the exact and model rates of each possible token change. For $i \neq m$, define the model rate factor by $-(\alpha_t'/\alpha_t)P_{i,m} := \lim_{h\downarrow 0} h^{-1} \Pr_\theta(x_{t-h} = i \mid x_t = m)$. The loss L_t then becomes

$$-\frac{\alpha_t'}{\alpha_t}\left[\mathbb{1}_{m=y}\sum_{i\neq m}\mathcal{D}\left(\frac{\lambda_t\beta_t}{r_t}, P_{i,m}\right) + \mathbb{1}_{m\neq y}\left\{\mathcal{D}\left(\frac{\alpha_t}{\beta_t} + \lambda_t, P_{y,m}\right) + \sum_{i\neq m,y}\mathcal{D}\left(\frac{\sigma_t}{V}, P_{i,m}\right)\right\}\right]. \tag{11}$$

Our model parameterizes the clean posterior rather than substituting its prediction directly into a bridge parameterization. The resulting map is affine, and the true posterior $q(x_0 \mid x_t)$ is invariant to coupling choices. Consequently, all feasible processes share a common optimum.

Proposition 2 (A common optimal predictor)**.** For any fixed $s < t$ and feasible coupling shared across clean candidates, the true clean posterior $\hat{x}_\theta = q(x_0 \mid x_t)$ minimizes

$$L_{s,t} := \mathbb{E}_{q(x_0|x_t)}\,\mathrm{KL}(q(x_s \mid x_t, x_0) \,\|\, p_\theta(x_s \mid x_t))\,. \tag{12}$$

Appendix B.2.4 proves the proposition. In continuous non-Markovian diffusion, such as DDIM, the losses differ only by scalar weights. In discrete non-Markovian diffusion, the optimization paths differ: Flip emphasizes correction, while Absorb emphasizes confidence in clean tokens. To enable a single pretrained model to support different samplers, we adopt the strategy: *train across couplings and infer with an adapted coupling*.

Specifically, for each token during training, we sample $\lambda_t \sim \mathrm{Unif}(0, \lambda_t^{\max})$ and $\sigma_t \mid \lambda_t \sim \mathrm{Unif}(0, \sigma_t^{\max} - \lambda_t/(1-2/V))$. We call this scheme **NoMUD-Rand**. Its objective averages over feasible coupling pairs and has two benefits: (1) The trained model naturally supports adaptive couplings for sampling and likelihood evaluation, as discussed in Section 3.3. (2) Training combines the characteristics of different couplings and mitigates the inductive bias of any single objective.

3.3 Adaptive coupling design

We now design the coupling for likelihood evaluation, sampling, and few-step distillation.

Likelihood evaluation. Diffusion models commonly use an estimated NELBO as a proxy for negative log-likelihood. We keep the reference diffusion process fixed to the standard Markov coupling $(1/V, 1)$ and optimize the coupling used to construct the model rate factors $P_{i,m}$, while keeping θ fixed. Specifically, for each clean example and time,

$$(\lambda_t^\star, \sigma_t^\star) = \operatorname*{arg\,min}_{(\lambda,\sigma)\in\mathcal{S}_{\lambda,\sigma}(t)} \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \sum_{\ell=1}^{L} L_{t,\mathrm{eval}}^{(\ell)}(\lambda, \sigma). \tag{13}$$

This optimization directly minimizes the NELBO used for likelihood evaluation. The objective is convex in (λ, σ), so constrained Newton methods or other convex optimization algorithms can efficiently solve it. Appendix B.2.3 gives the explicit objective and proof.

Confidence-aware sampling. Confidence-aware sampling is widely used in MDLMs to mitigate difficulties arising from dependencies between tokens (Wang et al., 2025). UDLMs, however, have no mask token to distinguish undecoded positions from decoded ones, making direct transfer of such strategies difficult. We instead design confidence-dependent coupling parameters to induce the desired sampling behavior as follows: A confident prediction of the current token should favor retention; a confident prediction of another token should favor correction; and a nearly uniform prediction should allow exploration. These behaviors correspond to three representative transitions:

$$\begin{gathered}\hat{x}_\theta \to x_t \implies (\lambda_t, \sigma_t) \to (0, 0), \qquad \hat{x}_\theta \to \mathbf{u} \implies (\lambda_t, \sigma_t) \to (0, \sigma_t^{\max}), \\ \hat{x}_{\theta,i} \to 1\ (i \neq m) \implies (\lambda_t, \sigma_t) \to (\lambda_t^{\max}, 0).\end{gathered} \tag{14}$$

Appendix B.4 provides specific coupling parameter settings and proves feasibility and the stated limits. Confidence-aware coupling encourages an easy-to-hard generation order: relatively certain tokens transition preferentially, while difficult tokens receive more exploration to reduce errors, improving the balance between exploration and exploitation.

The Ψ-sampler. Within our framework, the Ψ-sampler (Deschenaux et al., 2026), which mixes the $t \to 0$ and $t \to s$ transitions, is naturally a special case. For a standard UDLM with coupling parameters $(1/V, 1)$, its transition is recovered by sampling with

$$\lambda_{t,s} = \frac{\kappa_t}{V} + (1-\kappa_t)\frac{r_t}{V\delta_{t,s}}, \qquad \sigma_{t,s} = \kappa_t + (1-\kappa_t)\frac{\beta_t}{\delta_{t,s}}, \qquad 0 \le \kappa_t \le 1. \tag{15}$$

Appendix B.6 proves this equivalence. The coupling view also explains its effectiveness: introducing a small $t \to 0$ component increases Explore and Flip behavior while reducing the Absorb component, helping the sampler avoid settling prematurely on incorrect tokens.

Few-step distillation. Discrete consistency distillation (DCD) matches a teacher's $s \to 0$ prediction to a student's $t \to 0$ prediction, improving clean-token prediction in the few-step regime (Sahoo et al., 2025). Within our non-Markovian framework, we additionally learn a coupling head so that the student's $t \to s$ transition matches the exact transition conditioned on the clean example. Specifically, for $s < t$, we draw $x_t \sim q(x_t \mid x_0)$ and $x_s \sim q_{\text{ref}}(x_s \mid x_t, x_0)$, where q_{ref} is a reference bridge. We train the student and a head with parameters ψ using

$$L_{\text{distill}} = \mathrm{KL}(\hat{x}_\theta(x_t, t) \,\|\, \mathrm{sg}(\hat{x}_{\theta^-}(x_s, s))) + \mathrm{KL}\big(q_{\text{ref}}(x_s \mid x_t, x_0) \,\big\|\, p_{\mathrm{sg}(\theta),\psi}(x_s \mid x_t)\big) \tag{16}$$

where θ^- denotes the teacher parameters and sg stops gradients. The head outputs feasible coupling parameters for the model transition. The second term allows the head to find coupling parameters that are better suited for the $t \to s$ transition under fixed predictions. We run the distillation process for 5 rounds, each consisting of 10K training steps. At the end of each round, we replace the teacher weights with the student's current weights. Algorithm 2 gives the pseudocode.

4 Experiments

Experimental setup. Following Duo (Sahoo et al., 2025), we train approximately 170M-parameter models on OpenWebText (Gokaslan & Cohen, 2019), using the GPT-2 tokenizer (Radford et al., 2019), sequence length 1,024, and batch size 512. Our masked-diffusion baselines are SEDD (Lou et al., 2024) and RADD (Ou et al., 2025), using released checkpoints trained for 400K steps. Uniform-diffusion baselines comprise USDM (Schiff et al., 2025), Duo, SDDLM, and SDDLM-v1 (Zhu et al., 2025); these and our NoMUD variants use reproduced 500K-step checkpoints. NoMUD-Rand denotes training with randomized couplings, whereas Absorb, Flip, and Explore are trained with fixed couplings.

4.1 Likelihood evaluation

Table 2 reports likelihood evaluation results on seven datasets. Following Duo, we prepend and append `[BOS]` and `[EOS]` tokens to each evaluation example. We estimate the sample-averaged, per-token NELBO with 1,024 Monte Carlo draws and exponentiate it to obtain each dataset's perplexity. For NoMUD (Rand, Opt), the reference process is fixed to the Markov coupling $(1/V, 1)$, and only the coupling entering $P_{i,m}$ is optimized with θ fixed.

NoMUD (Rand, Opt) improves over Duo on five datasets, reducing the average perplexity from 61.14 to 58.50. It outperforms USDM, SDDLM, and SDDLM-v1 on all seven datasets. MDLMs generally achieve better likelihood evaluation results here. Since their denoising NELBOs are evaluated only at masked positions, direct comparisons across diffusion families require caution (Sahoo et al., 2026b). Nevertheless, NoMUD (Rand, Opt) remains competitive.

Although Duo and the two SDDLM variants use different training losses, their evaluation corresponds to the Markov coupling $(1/V, 1)$. NoMUD-Rand may benefit from two complementary mechanisms: (1) randomized-coupling training improves the predictor's adaptability; (2) optimizing the model-side coupling, with the reference process and θ fixed, compensates for prediction errors and lowers the evaluation NELBO.

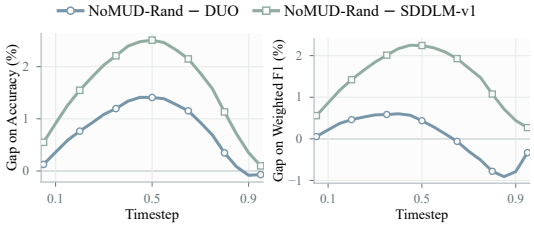
Table 2: Likelihood evaluation results on seven datasets (lower is better; 1,024 Monte Carlo draws). Avg. is the arithmetic mean across the seven datasets. Bold and underlined values are best and second best within the UDLM family.

Model	PTB	WikiText	LM1B	LAM-BADA	AGNews	PubMed	ArXiv	Avg.
SEDD	80.51	36.83	72.40	51.95	72.03	46.74	40.72	57.31
RADD	78.13	34.93	69.47	52.60	67.22	47.94	42.95	56.18
USDM	104.58	47.26	91.89	63.90	105.28	63.38	52.72	75.57
DUO	<u>85.81</u>	<u>38.66</u>	<u>78.51</u>	**54.75**	<u>77.04</u>	**49.64**	<u>43.55</u>	<u>61.14</u>
SDDLM	106.42	47.99	92.53	63.97	102.12	60.66	50.85	74.93
SDDLM-v1	135.76	56.57	110.08	83.66	126.19	78.79	60.51	93.08
NoMUD Absorb	137.14	63.21	130.97	102.55	152.97	106.99	74.80	109.80
NoMUD Flip	1.93×10^6	2.38×10^5	2.92×10^6	6.69×10^5	3.44×10^6	5.21×10^5	2.97×10^5	1.43×10^6
NoMUD Explore	88.26	41.34	84.56	60.25	84.16	59.90	47.55	66.57
NoMUD (Rand, Opt)	**73.55**	**36.36**	**75.31**	<u>57.68</u>	**72.16**	<u>51.06</u>	**43.36**	**58.50**

Table 3: Generation perplexity at target entropy 5.432 nats and 1,024 sequences per setting. Bold and underlined entries are best and second best within the UDLM family in each column. Complete frontiers are reported in Appendix C.2.

Model / Steps	8	16	32	64	128	256	512	1024
SEDD	270.09	124.42	76.53	65.38	63.59	61.35	60.16	64.25
RADD	255.45	117.63	80.03	66.54	63.67	60.43	59.07	65.25
USDM	154.94	92.05	74.73	71.17	71.45	66.76	63.90	63.46
DUO	138.33	<u>81.93</u>	<u>66.36</u>	<u>60.09</u>	**56.93**	<u>54.31</u>	60.56	60.10
SDDLM	165.04	95.89	77.90	70.27	70.30	66.35	68.35	66.38
SDDLM-v1	<u>137.46</u>	89.08	72.21	63.50	64.00	64.28	62.56	63.05
NoMUD Absorb	147.03	93.27	78.12	75.03	71.19	72.27	74.61	73.17
NoMUD Flip	324.89	150.15	89.88	69.91	62.54	58.77	<u>58.34</u>	<u>58.14</u>
NoMUD Explore	145.64	89.00	70.90	63.61	64.91	62.00	61.83	67.29
NoMUD (Rand, Conf)	**134.23**	**77.35**	**60.56**	**57.07**	<u>57.40</u>	**50.96**	**54.90**	**53.25**

Coupling-independent prediction. To examine the predictor only, we treat clean-token prediction from (x_t, t) as a token-level multiclass classification task. For a fair comparison, we calibrate Duo and SDDLM-v1 as $\widetilde{x}_{\theta,i} = (V\alpha_t \mathbf{1}_{i=m} + \beta_t)\hat{x}_{\theta,i} / (V\alpha_t \hat{x}_{\theta,m} + \beta_t)$. NoMUD-Rand outperforms SDDLM-v1 at all evaluated timesteps and Duo at most timesteps, although some gaps become negative at high noise (Figure 3).



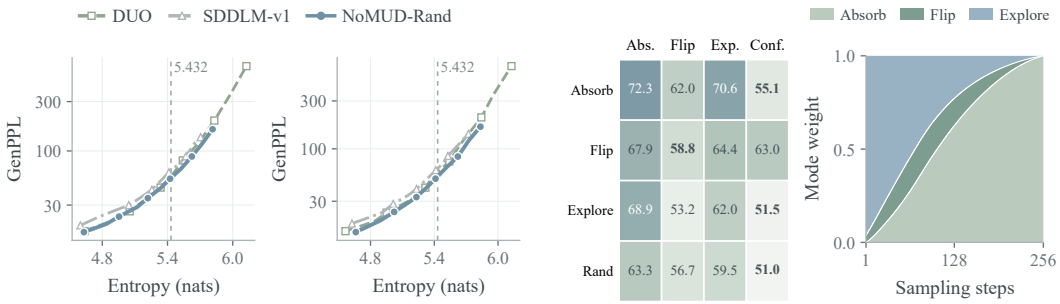
(a) Accuracy (b) Weighted F1

Figure 3: Gaps in percentage points.

Effect of evaluation coupling. The three non-Markovian processes also show that likelihood evaluation is sensitive to the coupling parameters and requires careful design. For example, under its training coupling $(\lambda_t^{\max}, 0)$, Flip's average perplexity reaches 1.43×10^6 (Table 2). When evaluated with $(1/V, 1)$ while keeping θ fixed, its average perplexity improves to 61.58 (Appendix C.1). Average perplexity decreases from Flip with $(1/V, 1)$ (61.58) to Rand with $(1/V, 1)$ (61.34), and then to Rand with optimized coupling (58.50), showing that both randomized-coupling training and optimized evaluation coupling contribute to the improvement.

4.2 Sampling and Distillation

Generation at comparable entropy. Following Pynadath et al. (2026), we sweep temperatures from 0.7 to 1.2, generating 128 sequences per setting to trace the generation perplexity-entropy frontier.



(a) 128 steps (b) 1,024 steps (c) Sampler transfer (d) Mode allocation

Figure 4: (a,b) GenPPL–entropy frontiers; dashed lines mark 5.432 nats. (c) Cross-sampler diagnostic at 256 steps. Rows identify pretrained predictors; columns are sampling methods. (d) Mean mode weights over 1,024 sequences at 256 steps.

Table 4: GPT-2 Large generation perplexity with the Ψ-sampler at target entropy 5.432 nats, using 1,024 sequences per setting. Bold and underlined entries are best and second best within each column.

Model / Steps	8	16	32	64	128	256	512	1024
DUO	139.62	<u>81.73</u>	<u>65.07</u>	<u>57.62</u>	**54.46**	<u>54.77</u>	**52.65**	**45.21**
USDM	153.22	96.10	74.74	72.17	64.56	61.73	60.88	56.38
SDDLM	168.76	100.24	78.25	70.63	66.45	62.78	61.60	58.57
SDDLM-v1	<u>136.35</u>	87.45	72.27	65.41	61.79	60.07	59.18	55.69
NoMUD-Rand	**133.96**	**76.46**	**64.86**	**54.67**	<u>54.82</u>	**51.04**	<u>52.98</u>	<u>50.07</u>

GenPPL is scored by GPT-2 Large throughout the generation study; entropy is the per-sequence unigram entropy in nats. Figure 4(a),(b) shows the DUO, SDDLM-v1, and NoMUD-Rand frontiers at 128 and 1,024 steps. We target 5.432 nats, the OpenWebText validation mean reported by Pynadath et al. (2026), interpolate a temperature for each model and budget, and generate 1,024 new sequences. Table 3 reports their GenPPL. The complete frontiers appear in Appendix C.2.

NoMUD ranks among the top two at all eight sampling budgets and achieves the lowest GenPPL at seven. At 512 steps, it reaches 54.90, compared with 60.56 for Duo and 62.56 for SDDLM-v1; at 1,024 steps, the corresponding values are 53.25, 60.10, and 63.05. It also outperforms Absorb, Flip, and Explore at every budget. These results show that confidence-aware coupling accommodates different sampling budgets, rather than restricting generation to a fixed preference.

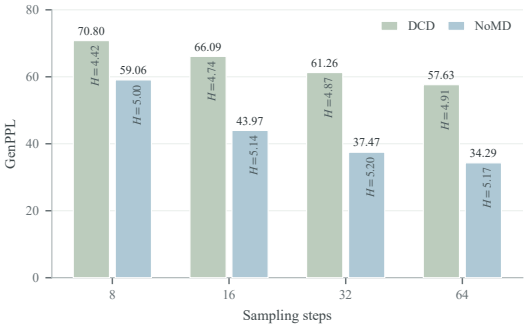


Figure 5: Few-step generation after 50k distillation steps at temperature one. Vertical labels report entropy in nats.

Can a pretrained predictor use different samplers? We freeze each of the four non-Markovian predictors and apply the Absorb, Flip, Explore, and confidence-aware samplers. All four predictors support these choices, yielding GenPPL between 50.96 and 72.27 at 256 steps (Figure 4 (c)). This also provides a sampler ablation: confidence-aware sampling gives the lowest GenPPL for three of the four predictors, while the randomly trained predictor has the lowest mean GenPPL across the four samplers (57.61), indicating stronger adaptation to different sampling couplings. Appendix C.3 gives results at all budgets.

How does confidence allocate transition behavior? To explain the behavior of the confidence-aware sampler, we view each sampling step as a soft mixture of the Absorb, Flip, and Explore processes and show the three mode weights in Figure 4 (d). The sampler shifts from early exploration to late

retention, with the correction weight peaking in between. This pattern persists across all eight budgets (Appendix C.4), illustrating how confidence-aware coupling combines the three complementary behaviors. Representative sampling trajectories are provided in Appendix C.6.

Ψ samplers. We compare UDLM baselines using the Ψ-sampler in Table 4. NoMUD-Rand achieves the best result at five budgets and remains among the top two throughout. By introducing additional Explore- and Flip-like behavior, the Ψ-sampler alleviates the sampling plateau of UDLMs, so generation quality improves overall as the number of steps increases.

Distillation. Following DCD, we distill for 50k steps and compare the few-step generation of models distilled from DUO and NoMUD-Rand, denoted DCD and NoMD, respectively. Because the distilled models are sensitive to temperature, Figure 5 reports GenPPL and entropy at temperature one. NoMD samples with λ and σ from its learned head, whereas DCD uses greedy-tail sampling and takes an $\arg\max$ rather than sampling at the final step. NoMD achieves lower GenPPL and higher entropy at every budget. Learning the coupling allows the student to use non-Markovian flexibility to better match the teacher's few-step transitions, improving few-step generation while avoiding collapse. Appendix B.5 gives the algorithm, and Appendix C.5 reports ablations.

5 Related work

Uniform diffusion language models. Although MDLMs have demonstrated parallel decoding at scale (Nie et al., 2025; Ye et al., 2025), their binary masked/unmasked state space limits self-correction and few-step sampling. By adopting a uniform prior, UDLMs have shown potential for self-correction (Deschenaux et al., 2026), few-step generation (Sahoo et al., 2025), and scaling in data-constrained regimes (von Rütte et al., 2026), attracting increasing attention. A line of work improves UDLM training and sampling by simplifying training objectives (Zhu et al., 2025), alleviating sampling plateaus (Deschenaux et al., 2026), and propagating information across sampling steps (Jo et al., 2026). Another line scales UDLMs to assess their potential on complex tasks. DiffusionGemma (DiffusionGemma Team, 2026) and Uno (Sahoo et al., 2026a) bring uniform diffusion to billion-parameter models for blockwise generation and autoregressive drafting, respectively. However, these methods retain a Markovian forward process, restricting the flexibility of UDLMs.

Non-Markovian diffusion processes. Non-Markovian designs have proved useful in diffusion-based generation. DDIM enables accelerated image generation with a pretrained denoiser and also discusses a one-parameter discrete construction (Song et al., 2021). GFDM (Nobis et al., 2024) and DART (Gu et al., 2025) further demonstrate that non-Markovian designs can improve the efficiency and flexibility of continuous diffusion. In discrete diffusion, non-Markovian masked processes support remasking and improve sampling through iterative revision (Wang et al., 2025). Most non-Markovian UDLM constructions focus on absorbing special cases: DNDM (Chen et al., 2024) develops fast sampling with predetermined transition times, with its absorbing construction corresponding to our $(0, 0)$ Absorb process; Duo considers a related process for distillation (Sahoo et al., 2025); and AUDM conditions on additional absorbing tokens (Gourevitch et al., 2026). CaDDi instead conditions on the full temporal trajectory, introducing additional modeling complexity (Zhang et al., 2025). Our work exposes two coupling degrees of freedom within UDLMs and develops adaptive coupling designs.

6 Conclusion

In this work, we introduced *NoMUD*, a family of non-Markovian UDLMs that makes the coupling between noise levels explicitly designable while preserving the standard UDLM marginals. We showed that different couplings can induce substantially different training and sampling behaviors while sharing the same optimal clean-token predictor, separating what the model predicts from how that prediction is used to drive the transition. Building on this flexibility, we designed couplings for NELBO minimization, confidence-aware sampling, and efficient few-step distillation. These results highlight coupling as an important degree of freedom in discrete diffusion: rather than committing to a single transition process for all purposes, UDLMs can adapt their dynamics to the objective at hand. We hope this perspective opens a broader direction for designing discrete diffusion models beyond fixed Markov transitions.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

Following the ICLR guidelines, we clarify how large language models were used in preparing this manuscript. The research ideas, methodology, experimental design, and analysis were developed by the human authors, and LLMs were not used for research ideation. We used LLMs only to assist with language editing and polishing in the main paper and appendix. The authors reviewed the full manuscript and take responsibility for all of its content.

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, pp. 17981–17993, 2021. URL https://openreview.net/forum?id=h7-XixPCAL.

Zixiang Chen, Huizhuo Yuan, Yongqian Li, Yiwen Kou, Junkai Zhang, and Quanquan Gu. Fast sampling via discrete non-markov diffusion models with predetermined transition time. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2312.09193.

Justin Deschenaux, Caglar Gulcehre, and Subham Sekhar Sahoo. The diffusion duality, chapter II: Ψ-samplers. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.21185.

DiffusionGemma Team. DiffusionGemma technical report. *arXiv preprint arXiv:2608.00146*, 2026. URL https://arxiv.org/abs/2608.00146.

Aaron Gokaslan and Vanya Cohen. OpenWebText Corpus. https://skylion007.github.io/OpenWebTextCorpus/, 2019.

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.

Jiatao Gu, Yuyang Wang, Yizhe Zhang, Qihang Zhang, Dinghuai Zhang, Navdeep Jaitly, Josh Susskind, and Shuangfei Zhai. DART: Denoising autoregressive transformer for scalable text-to-image generation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.08159.

Mingyu Jo, Jaesik Yoon, Justin Deschenaux, Caglar Gulcehre, and Sungjin Ahn. Loopholing discrete diffusion: Deterministic bypass of the sampling wall. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.19304.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2310.16834.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. URL https://arxiv.org/abs/2502.09992.

Gabriel Nobis, Maximilian Springenberg, Marco Aversa, Michael Detzel, Rembert Daems, Roderick Murray-Smith, Shinichi Nakajima, Sebastian Lapuschkin, Stefano Ermon, Tolga Birdal, Manfred Opper, Christoph Knochenhauer, Luis Oala, and Wojciech Samek. Generative fractional diffusion models. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=B9qg3wo75g.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.03736.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019. URL https://github.com/openai/gpt-2.

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*, 2024. URL https://arxiv.org/abs/2406.07524.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2506.10892.

Subham Sekhar Sahoo, Lingjie Chen, Khiem Pham, Jonathan Geuter, Chaitanya Dwivedi, Varad Pimpalkhute, Yash Akhauri, Alexander Moreno, Mikhail Yurochkin, Zhenting Wang, Mostafa Elhoushi, Nolan Dey, Shane Bergsma, Joel Hestness, John Thickstun, Eric Xing, and Zhengzhong Liu. Unlocking lossless speedups in LLMs via discrete diffusion. *arXiv preprint arXiv:2609.04010*, 2026a. URL https://arxiv.org/abs/2609.04010.

Subham Sekhar Sahoo, Jean-Marie Lemercier, Zhihan Yang, Justin Deschenaux, Jingyu Liu, John Thickstun, and Ante Jukić. Scaling beyond masked diffusion language models. *arXiv preprint arXiv:2602.15014*, 2026b. URL https://arxiv.org/abs/2602.15014.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-Torre, Bernardo P. de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.10193.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

Dimitri von Rütte, Janis Fluri, Omead Pooladzandi, Bernhard Schölkopf, Thomas Hofmann, and Antonio Orvieto. Scaling behavior of discrete diffusion language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2512.10858.

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*, 2025. URL https://arxiv.org/abs/2503.00307.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025. URL https://arxiv.org/abs/2508.15487.

Yangtian Zhang, Sizhuang He, Daniel Levine, Lawrence Zhao, David Zhang, Syed Asad Rizvi, Shiyang Zhang, Emanuele Zappala, Rex Ying, and David van Dijk. Non-markovian discrete diffusion with causal language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/pdf?id=qYSgnmT3dp.

Huaisheng Zhu, Zhengyu Chen, Shijie Zhou, Zhihui Xie, Yige Yuan, Shiqi Chen, Zhimeng Guo, Siyuan Xu, Hangfan Zhang, Vasant Honavar, and Teng Xiao. Simple denoising diffusion language models. *arXiv preprint arXiv:2510.22926*, 2025. URL https://arxiv.org/abs/2510.22926.

A LIMITATIONS

This work introduces non-Markovian UDLMs and uses coupling design to improve likelihood evaluation, sampling, and distillation. Nevertheless, several limitations remain. (1) Sampling complexity. Our experiments demonstrate improved generation across a range of step budgets, but they do not provide a theoretical guarantee on sampling complexity. It remains open whether non-Markovian coupling or confidence-aware sampling can provably improve the asymptotic number of sampling steps under explicit assumptions on score error and token dependence. (2) Scaling. Due to computational constraints, our experiments use approximately 170M-parameter models and do not establish how NoMUD behaves at substantially larger scales. Scaling NoMUD to comparable model and data regimes is left to future work.

B PROOFS

B.1 FORWARD, BACKWARD, AND MODEL TRANSITIONS

B.1.1 MARGINAL AND FEASIBLE REGION

Proposition 1 (Marginal and feasible region). Let

$$q(x_t \mid x_0) = \alpha_t x_0 + \beta_t \mathbf{u}, \qquad \mathbf{u} = \mathbf{1}/V, \quad \beta_t = 1 - \alpha_t, \quad r_t = V\alpha_t + \beta_t, \tag{B.1}$$

and, for $s < t$, let $\delta_{t,s} = 1 - \alpha_t/\alpha_s$. For $i \neq m$, define

$$\Pr(x_t = i \mid x_s = m, x_0) = \delta_{t,s}\left[\mathbb{1}_{m=y}\frac{\alpha_s + \beta_s\lambda_t}{r_s} + \mathbb{1}_{i=y}\lambda_t + \mathbb{1}_{m,i\neq y}\frac{\sigma_t}{V}\right], \tag{B.2}$$

$$\Pr(x_t = m \mid x_s = m, x_0) = 1 - \sum_{i\neq m}\Pr(x_t = i \mid x_s = m, x_0).$$

This transition preserves (B.1) and is categorical for every $s < t$ if

$$0 \le \lambda_t, \qquad 0 \le \sigma_t, \qquad \lambda_t \le \min\left\{\frac{1}{\beta_t}, \frac{r_t}{(V-1)(1-\sqrt{\alpha_t})^2}\right\}, \qquad \sigma_t + \frac{\lambda_t}{1-2/V} \le \frac{1}{(1-2/V)\beta_t}. \tag{B.3}$$

Proof. We first identify the two degrees of freedom, prove marginal preservation for shared coefficients, and then derive the inequalities needed for every row to be categorical.

Two degrees of freedom. Under the token-symmetric form, let a be the probability of moving from the clean token to each specific wrong token, b the probability of moving from a wrong token to the clean token, and c the probability of moving between two distinct wrong tokens. Normalization determines the clean and wrong diagonals as $1-(V-1)a$ and $1-b-(V-2)c$. For a fixed wrong coordinate, marginal preservation requires

$$\frac{r_s}{V}a + \frac{\beta_s}{V}[1-b-(V-2)c] + \frac{(V-2)\beta_s}{V}c = \frac{\beta_t}{V},$$

or $r_s a - \beta_s b = \alpha_s - \alpha_t$. Thus, two scalar degrees of freedom remain. For an interior pair with $\delta_{t,s} > 0$, set $b = \delta_{t,s}\lambda_t$ and $c = \delta_{t,s}\sigma_t/V$. The remaining probability is uniquely

$$a = \frac{\alpha_s - \alpha_t + \beta_s\delta_{t,s}\lambda_t}{r_s} = \frac{\delta_{t,s}(\alpha_s + \beta_s\lambda_t)}{r_s},$$

which gives (B.2). The following argument applies conditional on each fixed clean example, so the shared scalar coefficients may also depend on that example.

Marginal preservation. At time s,

$$\Pr(x_s = y \mid x_0) = \alpha_s + \frac{\beta_s}{V} = \frac{r_s}{V}, \qquad \Pr(x_s = i \mid x_0) = \frac{\beta_s}{V} \quad (i \neq y).$$

Fix one wrong coordinate $i \neq y$. The event $x_t = i$ can arise from the clean source y, from the same wrong source i, or from one of the other $V-2$ wrong sources. These three contributions are

$$\Pr(x_s = y \mid x_0)\Pr(x_t = i \mid x_s = y, x_0) = \frac{r_s}{V}\frac{\delta_{t,s}(\alpha_s + \beta_s\lambda_t)}{r_s} = \frac{\delta_{t,s}(\alpha_s + \beta_s\lambda_t)}{V},$$

$$\Pr(x_t = i \mid x_s = i, x_0) = 1 - \delta_{t,s}\lambda_t - (V-2)\frac{\delta_{t,s}\sigma_t}{V},$$

and, for each $m \neq i, y$,

$$\Pr(x_s = m \mid x_0)\Pr(x_t = i \mid x_s = m, x_0) = \frac{\beta_s}{V}\frac{\delta_{t,s}\sigma_t}{V}.$$

Adding the three mutually exclusive cases gives

$$\begin{aligned}\Pr(x_t = i \mid x_0) &= \frac{\delta_{t,s}(\alpha_s + \beta_s\lambda_t)}{V} + \frac{\beta_s}{V}\left[1 - \delta_{t,s}\lambda_t - (V-2)\frac{\delta_{t,s}\sigma_t}{V}\right] + (V-2)\frac{\beta_s\delta_{t,s}\sigma_t}{V^2} \\ &= \frac{\beta_s + \delta_{t,s}\alpha_s}{V} = \frac{\beta_s + \alpha_s - \alpha_t}{V} = \frac{\beta_t}{V}.\end{aligned}$$

The λ_t terms cancel, and the σ_t terms cancel. Hence

$$\Pr(x_t = y \mid x_0) = 1 - (V-1)\frac{\beta_t}{V} = \alpha_t + \frac{\beta_t}{V} = \frac{r_t}{V},$$

which proves $q(x_t \mid x_0) = \alpha_t x_0 + \beta_t \mathbf{u}$.

Conditions for a categorical transition. Nonnegative λ_t and σ_t make every off-diagonal probability nonnegative. It remains to make the diagonal probabilities nonnegative. For a clean source, the total outgoing mass is

$$\sum_{i \neq y}\Pr(x_t = i \mid x_s = y, x_0) = \frac{(V-1)\delta_{t,s}(\alpha_s + \beta_s\lambda_t)}{r_s} \leq 1.$$

Using $r_s - r_t = (V-1)\delta_{t,s}\alpha_s$, this is equivalent to

$$(V-1)\delta_{t,s}\beta_s\lambda_t \leq r_t. \tag{B.4}$$

For a wrong source,

$$\sum_{i \neq m}\Pr(x_t = i \mid x_s = m, x_0) = \delta_{t,s}\left[\lambda_t + \left(1 - \frac{2}{V}\right)\sigma_t\right] \leq 1. \tag{B.5}$$

For all $s < t$,

$$\delta_{t,s} \leq \beta_t,$$

and

$$\delta_{t,s}\beta_s = 1 + \alpha_t - \alpha_s - \frac{\alpha_t}{\alpha_s} \leq 1 + \alpha_t - 2\sqrt{\alpha_t} = (1 - \sqrt{\alpha_t})^2,$$

where the inequality follows from $(\sqrt{\alpha_s} - \sqrt{\alpha_t/\alpha_s})^2 \geq 0$. Therefore (B.4) holds whenever

$$\lambda_t \leq \frac{r_t}{(V-1)(1-\sqrt{\alpha_t})^2},$$

and (B.5) holds whenever

$$\lambda_t + \left(1 - \frac{2}{V}\right)\sigma_t \leq \frac{1}{\beta_t}, \tag{B.6}$$

which is equivalent to the last inequality in (B.3). Since (B.6) also implies $\lambda_t \leq 1/\beta_t$, the complete sufficient conditions are exactly (B.3).

□

B.1.2 EXACT BACKWARD TRANSITION

Proposition B.1 (Exact backward transition)**.** For shared λ_t, σ_t and $i \neq m$, the exact backward transition is

$$\Pr(x_s = i \mid x_t = m, x_0) = \delta_{t,s}\left[\mathbb{1}_{m=y}\frac{\beta_s\lambda_t}{r_t} + \mathbb{1}_{i=y}\frac{\alpha_s + \beta_s\lambda_t}{\beta_t} + \mathbb{1}_{m,i\neq y}\frac{\beta_s\sigma_t}{V\beta_t}\right], \tag{B.7}$$

with

$$\Pr(x_s = m \mid x_t = m, x_0) = 1 - \sum_{i\neq m}\Pr(x_s = i \mid x_t = m, x_0).$$

Proof. For any candidate earlier coordinate i,

$$\Pr(x_s = i \mid x_t = m, x_0) = \frac{\Pr(x_t = m \mid x_s = i, x_0)\Pr(x_s = i \mid x_0)}{\Pr(x_t = m \mid x_0)}.$$

There are three off-diagonal cases.

If $m = y$ and $i \neq y$, the candidate has mass β_s/V, its forward probability of reaching y is $\delta_{t,s}\lambda_t$, and the observed clean mass is r_t/V. Hence

$$\Pr(x_s = i \mid x_t = y, x_0) = \frac{(\delta_{t,s}\lambda_t)(\beta_s/V)}{r_t/V} = \frac{\delta_{t,s}\beta_s\lambda_t}{r_t}.$$

If $m \neq y$ and $i = y$, the candidate has mass r_s/V, its forward probability of reaching m is $\delta_{t,s}(\alpha_s + \beta_s\lambda_t)/r_s$, and the observed mass is β_t/V. Therefore

$$\Pr(x_s = y \mid x_t = m, x_0) = \frac{\{\delta_{t,s}(\alpha_s + \beta_s\lambda_t)/r_s\}(r_s/V)}{\beta_t/V} = \frac{\delta_{t,s}(\alpha_s + \beta_s\lambda_t)}{\beta_t}.$$

Finally, if $m, i \neq y$ and $i \neq m$, the candidate and observation both have marginal mass β_s/V and β_t/V, while the forward move has probability $\delta_{t,s}\sigma_t/V$. Thus

$$\Pr(x_s = i \mid x_t = m, x_0) = \frac{(\delta_{t,s}\sigma_t/V)(\beta_s/V)}{\beta_t/V} = \frac{\delta_{t,s}\beta_s\sigma_t}{V\beta_t}.$$

Combining the three cases with indicators gives (B.7). The diagonal follows from normalization:

$$\sum_i \Pr(x_s = i \mid x_t = m, x_0) = \frac{\sum_i \Pr(x_t = m \mid x_s = i, x_0)\Pr(x_s = i \mid x_0)}{\Pr(x_t = m \mid x_0)} = 1.$$

In particular,

$$\Pr(x_s = y \mid x_t = y, x_0) = 1 - \frac{(V-1)\delta_{t,s}\beta_s\lambda_t}{r_t},$$

and, for $m \neq y$,

$$\Pr(x_s = m \mid x_t = m, x_0) = \frac{\beta_s}{\beta_t}\left[1 - \delta_{t,s}\left\{\lambda_t + \left(1 - \frac{2}{V}\right)\sigma_t\right\}\right].$$

□

The exact reverse transition leads directly to Algorithm 1: the denoiser supplies clean-token probabilities, a feasible rule selects the coupling, and (9) gives the categorical tran-

sition for each position. At $s = 0$, the transition reduces to the clean-token prediction.

Algorithm 1: Sampling from NoMUD

Input: Length L; grid $1 = t_K > \cdots > t_0 = 0$; denoiser $\hat{x}_\theta$; feasible coupling rule.
Output: Generated sequence $\mathbf{x}_0$.

// Initialize from the uniform terminal marginal.

Draw $x_{t_K}^{(\ell)} \sim \mathrm{Cat}(\mathbf{u})$ independently for $\ell = 1, \ldots, L$
for $j = K, K-1, \ldots, 1$ **do**
  Set $(t, s) = (t_j, t_{j-1})$ and compute $\hat{x}_\theta^{(\ell)}(\mathbf{x}_t, t)$ for all positions
  for $\ell = 1, \ldots, L$ **do**
    Let m index $x_t^{(\ell)}$ and choose feasible (λ, σ)
    // Use the off-current probabilities and normalize the current coordinate.
    Compute π_i from (9) for $i \neq m$ and set $\pi_m = 1 - \sum_{i \neq m} \pi_i$
    Draw $x_s^{(\ell)} \sim \mathrm{Cat}(\pi)$
return $\mathbf{x}_{t_0}$

Dependence on x_t and t. When designing a sampler, the coupling rule may depend on the current state x_t and time t. Sampling only requires each step to define a valid categorical transition; it does not require the resulting trajectory to preserve the prescribed intermediate marginals. State-dependent coupling parameters are therefore admissible whenever they yield a valid sampling kernel.

B.1.3 From clean-token probabilities to the model transition

Proposition B.2 (Clean prediction and reverse parameterization)**.** Let the network predict

$$\hat{x}_\theta = p_\theta(x_0 \mid x_t), \qquad \hat{x}_{\theta,i} \geq 0, \qquad \sum_i \hat{x}_{\theta,i} = 1.$$

With coupling coefficients shared across clean candidates, the model reverse distribution is the mean of the exact backward distributions,

$$p_\theta(x_s \mid x_t) = \sum_{x_0} p_\theta(x_0 \mid x_t) q(x_s \mid x_t, x_0). \tag{B.8}$$

For $i \neq m$,

$$\Pr_\theta(x_s = i \mid x_t) = \hat{x}_{\theta,m} \frac{\delta_{t,s}\beta_s\lambda_t}{r_t} + \hat{x}_{\theta,i} \frac{\delta_{t,s}(\alpha_s + \beta_s\lambda_t)}{\beta_t} + (1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}) \frac{\delta_{t,s}\beta_s\sigma_t}{V\beta_t}, \tag{B.9}$$

or equivalently

$$\Pr_\theta(x_s = i \mid x_t) = \frac{\delta_{t,s}}{\beta_t} \left[(\alpha_s + \beta_s\lambda_t)\hat{x}_{\theta,i} + \frac{\beta_s\sigma_t}{V}(1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}) \right] + \frac{\delta_{t,s}\beta_s\lambda_t}{r_t} \hat{x}_{\theta,m}, \tag{B.10}$$

while

$$\Pr_\theta(x_s = m \mid x_t) = 1 - \sum_{i \neq m} \Pr_\theta(x_s = i \mid x_t). \tag{B.11}$$

Proof. Fix the observed coordinate m and an output coordinate $i \neq m$. In the mixture (B.8), the candidate clean token is either m, i, or neither. These events have probabilities

$$\Pr_\theta(x_0 = m \mid x_t) = \hat{x}_{\theta,m}, \qquad \Pr_\theta(x_0 = i \mid x_t) = \hat{x}_{\theta,i}, \qquad \Pr_\theta(x_0 \notin \{m, i\} \mid x_t) = 1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}.$$

For these three cases, Proposition B.1 gives

$$\Pr(x_s = i \mid x_t = m, x_0 = m) = \frac{\delta_{t,s}\beta_s\lambda_t}{r_t},$$

$$\Pr(x_s = i \mid x_t = m, x_0 = i) = \frac{\delta_{t,s}(\alpha_s + \beta_s\lambda_t)}{\beta_t},$$

and, for any other clean coordinate,

$$\Pr(x_s = i \mid x_t = m, x_0) = \frac{\delta_{t,s}\beta_s\sigma_t}{V\beta_t}.$$

Multiplying each conditional probability by its candidate probability and adding gives

$$\Pr_\theta(x_s = i \mid x_t) = \hat{x}_{\theta,m}\frac{\delta_{t,s}\beta_s\lambda_t}{r_t} + \hat{x}_{\theta,i}\frac{\delta_{t,s}(\alpha_s + \beta_s\lambda_t)}{\beta_t} + (1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i})\frac{\delta_{t,s}\beta_s\sigma_t}{V\beta_t},$$

which is (B.9). Factoring $\delta_{t,s}/\beta_t$ from the last two terms gives (B.10). Since (B.8) is an average of normalized categorical distributions,

$$\sum_{x_s} p_\theta(x_s \mid x_t) = \sum_{x_0} p_\theta(x_0 \mid x_t) \sum_{x_s} q(x_s \mid x_t, x_0) = 1,$$

so the current-coordinate probability is exactly (B.11). □

B.2 Continuous-time ELBO, variational evaluation, and the common predictor

At a zero model rate, we use $\mathcal{D}(p, 0) = +\infty$ for $p > 0$.

B.2.1 Continuous-time negative ELBO

Proposition B.3 (Continuous-time negative ELBO)**.** Let α_t be continuously differentiable with $\alpha_t' < 0$, let $\beta_t = 1 - \alpha_t$, $r_t = V\alpha_t + \beta_t$, and let

$$q(x_t \mid x_0) = \alpha_t x_0 + \beta_t \mathbf{u}.$$

For shared feasible λ_t, σ_t, define, for $i \neq m$,

$$P_{i,m} = \left(\frac{\alpha_t}{\beta_t} + \lambda_t\right)\hat{x}_{\theta,i} + \frac{\lambda_t\beta_t}{r_t}\hat{x}_{\theta,m} + \frac{\sigma_t}{V}(1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}).$$

When the finite-grid bounds converge to their continuous-time limit and the reconstruction term tends to zero,

$$-\log p_\theta(x_0) \le \mathrm{NELBO}(x_0; \lambda, \sigma) = \int_0^1 \mathbb{E}_{q(x_t|x_0)} L_t \, dt, \tag{B.12}$$

where

$$L_t = -\frac{\alpha_t'}{\alpha_t}\Bigg[\mathbb{1}_{m=y}\sum_{i\neq m}\mathcal{D}\Big(\frac{\lambda_t\beta_t}{r_t}, P_{i,m}\Big) + \mathbb{1}_{m\neq y}\left\{\mathcal{D}\Big(\frac{\alpha_t}{\beta_t} + \lambda_t, P_{y,m}\Big) + \sum_{i\neq m,y}\mathcal{D}\Big(\frac{\sigma_t}{V}, P_{i,m}\Big)\right\}\Bigg].$$

Proof. We derive the finite-step bound first. We then compute the small-step probabilities and expand the full categorical KL, including the probability of staying at the current token.

The finite-step bound. Temporarily let $j = 0, \ldots, T$ index a grid with spacing $\Delta t = 1/T$, and evaluate all schedules at its grid points. On each chosen grid, conditional on x_0, the adjacent forward steps define

$$q(x_{1:T} \mid x_0) = \prod_{j=1}^{T} q(x_j \mid x_{j-1}, x_0). \tag{B.13}$$

For each $j \ge 2$, Bayes' rule rearranges to

$$q(x_j \mid x_{j-1}, x_0) = q(x_{j-1} \mid x_j, x_0)\frac{q(x_j \mid x_0)}{q(x_{j-1} \mid x_0)}. \tag{B.14}$$

The first forward step has distribution $q(x_1 \mid x_0)$. Substituting (B.14) into all later factors, the marginal ratios cancel successively:

$$q(x_{1:T} \mid x_0) = q(x_1 \mid x_0) \prod_{j=2}^{T} \left[q(x_{j-1} \mid x_j, x_0) \frac{q(x_j \mid x_0)}{q(x_{j-1} \mid x_0)} \right] = q(x_T \mid x_0) \prod_{j=2}^{T} q(x_{j-1} \mid x_j, x_0). \tag{B.15}$$

The generation model starts from the uniform law and uses the learned reverse steps:

$$p_\theta(x_{0:T}) = \frac{1}{V} \prod_{j=1}^{T} p_\theta(x_{j-1} \mid x_j). \tag{B.16}$$

Using $p_\theta(x_{0:T}) = p_\theta(x_0) p_\theta(x_{1:T} \mid x_0)$ gives the variational inequality directly:

$$\begin{aligned} \mathbb{E}_q \log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})} &= -\log p_\theta(x_0) + \mathbb{E}_q \log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{1:T} \mid x_0)} \\ &= -\log p_\theta(x_0) + \mathrm{KL}(q(x_{1:T} \mid x_0) \,\|\, p_\theta(x_{1:T} \mid x_0)) \geq -\log p_\theta(x_0). \end{aligned} \tag{B.17}$$

Here $\mathbb{E}_q$ averages over (B.13). The inequality uses nonnegativity of KL.

Substituting (B.15) and (B.16) gives

$$\frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})} = \frac{q(x_T \mid x_0)}{1/V} \frac{1}{p_\theta(x_0 \mid x_1)} \prod_{j=2}^{T} \frac{q(x_{j-1} \mid x_j, x_0)}{p_\theta(x_{j-1} \mid x_j)}.$$

Taking the logarithm separates this product into

$$\log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})} = \log \frac{q(x_T \mid x_0)}{1/V} - \log p_\theta(x_0 \mid x_1) + \sum_{j=2}^{T} \log \frac{q(x_{j-1} \mid x_j, x_0)}{p_\theta(x_{j-1} \mid x_j)}.$$

The first term is zero because $\alpha_T = 0$ and $q(x_T \mid x_0) = \mathbf{u}$. For the second term, Proposition B.2 gives $p_\theta(x_0 \mid x_1) = \hat{x}_{\theta,y}(x_1)$. For an interior term, summing out every state except x_j, x_{j-1} gives

$$\begin{aligned} \mathbb{E}_q \log \frac{q(x_{j-1} \mid x_j, x_0)}{p_\theta(x_{j-1} \mid x_j)} &= \sum_{x_j} q(x_j \mid x_0) \sum_{x_{j-1}} q(x_{j-1} \mid x_j, x_0) \log \frac{q(x_{j-1} \mid x_j, x_0)}{p_\theta(x_{j-1} \mid x_j)} \\ &= \mathbb{E}_{q(x_j \mid x_0)} \mathrm{KL}(q(x_{j-1} \mid x_j, x_0) \,\|\, p_\theta(x_{j-1} \mid x_j)). \end{aligned}$$

Consequently, the finite-grid bound is

$$\begin{aligned} -\log p_\theta(x_0) \leq\ & -\mathbb{E}_{q(x_1 \mid x_0)} \log \hat{x}_{\theta,y}(x_1) \\ & + \sum_{j=2}^{T} \mathbb{E}_{q(x_j \mid x_0)} \mathrm{KL}(q(x_{j-1} \mid x_j, x_0) \,\|\, p_\theta(x_{j-1} \mid x_j)). \end{aligned} \tag{B.18}$$

Small-step probabilities. Return to continuous time and put $s = t - \Delta t$. The schedule satisfies

$$\frac{\delta_{t,t-\Delta t}}{\Delta t} = \frac{\alpha_{t-\Delta t} - \alpha_t}{\alpha_{t-\Delta t} \Delta t} \longrightarrow -\frac{\alpha_t'}{\alpha_t}, \qquad \beta_{t-\Delta t} \longrightarrow \beta_t.$$

For $i \neq m$, divide (B.10) by Δt before taking the limit:

$$\begin{aligned} \frac{\Pr_\theta(x_{t-\Delta t} = i \mid x_t)}{\Delta t} &= \frac{\delta_{t,t-\Delta t}}{\Delta t} \left[\frac{\alpha_{t-\Delta t} + \beta_{t-\Delta t} \lambda_t}{\beta_t} \hat{x}_{\theta,i} + \frac{\beta_{t-\Delta t} \lambda_t}{r_t} \hat{x}_{\theta,m} + \frac{\beta_{t-\Delta t} \sigma_t}{V \beta_t} (1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}) \right] \\ &\longrightarrow -\frac{\alpha_t'}{\alpha_t} \left[\left(\frac{\alpha_t}{\beta_t} + \lambda_t \right) \hat{x}_{\theta,i} + \frac{\beta_t \lambda_t}{r_t} \hat{x}_{\theta,m} + \frac{\sigma_t}{V} (1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}) \right] = -\frac{\alpha_t'}{\alpha_t} P_{i,m}. \end{aligned} \tag{B.19}$$

The three exact probabilities in (B.7) give, by the same substitution,

$$\lim_{\Delta t \downarrow 0} \frac{\Pr(x_{t-\Delta t} = i \mid x_t = m, x_0)}{\Delta t} = -\frac{\alpha_t'}{\alpha_t} \left[\mathbb{1}_{m=y} \frac{\beta_t \lambda_t}{r_t} + \mathbb{1}_{i=y} \left(\frac{\alpha_t}{\beta_t} + \lambda_t \right) + \mathbb{1}_{m, i \neq y} \frac{\sigma_t}{V} \right]. \tag{B.20}$$

Thus the model and exact rates share the factor $-\alpha_t'/\alpha_t$. Only their three coordinate factors differ.

From the one-step KL to $\mathcal{D}$. The small-step KL contains two pieces that must be kept together. For one off-diagonal coordinate, let p and q be its exact and model rates. Its probabilities are $\Delta t\, p + o(\Delta t)$ and $\Delta t\, q + o(\Delta t)$, so its logarithmic contribution satisfies

$$\lim_{\Delta t \downarrow 0} \frac{\Delta t\, p + o(\Delta t)}{\Delta t} \log \frac{\Delta t\, p + o(\Delta t)}{\Delta t\, q + o(\Delta t)} = p \log \frac{p}{q}.$$

The stay state supplies the linear part that is missing from these logarithmic terms. For the next display only, let p, q denote the total exact and model exit rates. Since the stay probabilities are $1 - \Delta t\, p + o(\Delta t)$ and $1 - \Delta t\, q + o(\Delta t)$,

$$\log \frac{1 - \Delta t\, p + o(\Delta t)}{1 - \Delta t\, q + o(\Delta t)} = \Delta t(q - p) + o(\Delta t).$$

Each total exit rate is the sum of its off-diagonal rates. Returning to one coordinate, its share of the stay term is therefore $q - p$. Adding this to its logarithmic term gives

$$q - p + p \log \frac{p}{q} = \mathcal{D}(p, q).$$

Writing both rates as $-\alpha_t'/\alpha_t$ times their respective factors p, q, homogeneity gives

$$\mathcal{D}\left(-\frac{\alpha_t'}{\alpha_t} p, -\frac{\alpha_t'}{\alpha_t} q\right) = -\frac{\alpha_t'}{\alpha_t} \mathcal{D}(p, q).$$

Using the model factors in (B.19) and exact factors in (B.20) therefore gives

$$\lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathrm{KL}(q(x_{t-\Delta t} \mid x_t, x_0) \,\|\, p_\theta(x_{t-\Delta t} \mid x_t)) = L_t. \tag{B.21}$$

Zero-rate terms use the stated convention for $\mathcal{D}$.

Finally, the sum in (B.18) is a sum of Δt times the expected KL divided by Δt. Under the stated continuous-time limit, (B.21) turns this sum into $\int_0^1 \mathbb{E}_{q(x_t|x_0)} L_t \, dt$. The reconstruction expectation tends to zero by assumption. This proves (B.12). □

Sequence-level objective. For a sequence $\mathbf{x}_0$ of length L, the exact conditional transition factorizes across positions. The model transition is the product of tokenwise distributions, each of which may depend on the entire current sequence. Therefore its conditional KL is the sum of tokenwise KLs. The terminal uniform law contributes zero, and the reconstruction term is the sum of tokenwise negative log-probabilities. Applying the preceding limit at each position gives

$$\mathrm{NELBO}(\mathbf{x}_0; \lambda, \sigma) = \int_0^1 \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \left[\sum_{\ell=1}^{L} L_t^{(\ell)} \right] dt,$$

which is (10). All continuous-time statements use the convergence and vanishing-reconstruction assumptions stated in Proposition B.3.

Endpoint treatment. Training uses the regularized schedule $\alpha_t = 1 - 0.999t$ for numerical stability. Likelihood evaluation instead uses the linear schedule $\alpha_t = 1 - t$ together with the standard Markov reference process. Hence $q(x_1 \mid x_0) = \mathbf{u}$, so the terminal-prior KL in the finite-grid ELBO is exactly zero. The continuous-time expression is understood through the finite-grid limit in Proposition B.3; the endpoint itself need not be sampled explicitly.

At the high-noise endpoint, the fixed Markov choice $(\lambda, \sigma) = (1/V, 1)$ is always feasible on the model side. As $t \to 1$, its model rate factors and the exact Markov rate factors both approach $1/V$, with their difference of order $O(\alpha_t)$. Consequently, each local divergence is $O(\alpha_t^2)$, which cancels the $O(\alpha_t^{-1})$ prefactor sufficiently to give an integrable $O(\alpha_t)$ contribution. Since the optimized objective is no larger than this feasible Markov value, optimizing (λ, σ) does not introduce an additional high-noise endpoint divergence. The $t \to 0$ endpoint is subject to the convergence and vanishing-reconstruction assumptions in Proposition B.3.

B.2.2 MUTUAL-INFORMATION DECOMPOSITION

This subsection characterizes the general matched-coupling objective; the optimized likelihood evaluation in Section 3.3 instead fixes the reference process to the Markov coupling $(1/V, 1)$. Fix $s < t$ and x_t. By marginalization, $q_{\lambda,\sigma}(x_s \mid x_t) = \mathbb{E}_{q(x_0|x_t)} q_{\lambda,\sigma}(x_s \mid x_t, x_0)$. Insert this identity into the log ratio of $L_{s,t}$ in (12) and sum out x_0 in the second term:

$$L_{s,t} = \mathbb{E}_{q(x_0|x_t)} \mathrm{KL}(q_{\lambda,\sigma}(x_s \mid x_t, x_0) \,\|\, q_{\lambda,\sigma}(x_s \mid x_t)) \tag{B.22}$$

$$+ \mathrm{KL}(q_{\lambda,\sigma}(x_s \mid x_t) \,\|\, p_\theta(x_s \mid x_t)) \tag{B.23}$$

$$= I_{\lambda,\sigma}(X_0; X_s \mid X_t = x_t) + \mathrm{KL}(q_{\lambda,\sigma}(x_s \mid x_t) \,\|\, p_\theta(x_s \mid x_t)) \,. \tag{B.24}$$

The first term is the conditional mutual information. This proves (B.24), with the usual extended-value conventions at zero probabilities. The mutual-information term depends on the coupling but not on p_θ.

B.2.3 CONVEX OPTIMIZATION FOR LIKELIHOOD EVALUATION

For likelihood evaluation, we fix the reference process to the standard Markov coupling $(1/V, 1)$ and keep the predictor θ fixed. The coupling (λ, σ) is optimized only through the model rate factors $P_{i,m}$. The finite-grid variational inequality does not require the reference and model transitions to use the same coupling; it only requires each to define a normalized path distribution.

Proposition B.4 (Convex likelihood evaluation)**.** Fix a clean example $\mathbf{x}_0$, an interior time t, the predictor θ, and the Markov reference coupling $(1/V, 1)$. For $i \neq m$, let

$$P_{i,m}(\lambda, \sigma) = \left(\frac{\alpha_t}{\beta_t} + \lambda\right) \hat{x}_{\theta,i} + \frac{\lambda \beta_t}{r_t} \hat{x}_{\theta,m} + \frac{\sigma}{V} \left(1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}\right). \tag{B.25}$$

Define the token-level evaluation loss

$$L_{t,\mathrm{eval}} = -\frac{\alpha_t'}{\alpha_t} \Bigg[\mathbb{1}_{m=y} \sum_{i \neq m} \mathcal{D}\left(\frac{\beta_t}{V r_t}, P_{i,m}\right) + \mathbb{1}_{m \neq y} \left\{ \mathcal{D}\left(\frac{\alpha_t}{\beta_t} + \frac{1}{V}, P_{y,m}\right) + \sum_{i \neq m,y} \mathcal{D}\left(\frac{1}{V}, P_{i,m}\right) \right\} \Bigg]. \tag{B.26}$$

Then

$$F_t(\lambda, \sigma; \mathbf{x}_0) = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \sum_{\ell=1}^{L} L_{t,\mathrm{eval}}^{(\ell)}(\lambda, \sigma) \tag{B.27}$$

is convex on $\mathcal{S}_{\lambda,\sigma}(t)$.

Proof. For a fixed exact rate factor $a \geq 0$,

$$\mathcal{D}(a, P) = P - a + a \log \frac{a}{P} \tag{B.28}$$

is convex in $P > 0$, since

$$\frac{\partial}{\partial P} \mathcal{D}(a, P) = 1 - \frac{a}{P}, \qquad \frac{\partial^2}{\partial P^2} \mathcal{D}(a, P) = \frac{a}{P^2} \geq 0. \tag{B.29}$$

For fixed $\hat{x}_\theta$, the model rate factor in (B.25) is affine in (λ, σ):

$$\frac{\partial P_{i,m}}{\partial \lambda} = \hat{x}_{\theta,i} + \frac{\beta_t}{r_t} \hat{x}_{\theta,m}, \qquad \frac{\partial P_{i,m}}{\partial \sigma} = \frac{1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}}{V}. \tag{B.30}$$

Hence each divergence term in (B.26) is convex in (λ, σ). The positive factor $-\alpha_t'/\alpha_t$, the vocabulary and position sums, and the expectation over $q(\mathbf{x}_t \mid \mathbf{x}_0)$ preserve convexity. Finally, $\mathcal{S}_{\lambda,\sigma}(t)$ is a convex polytope, so the constrained optimization problem is convex. □

Derivatives and constrained Newton updates. For one divergence term $\mathcal{D}(a, P_{i,m})$, define

$$v_{i,m} = \begin{bmatrix} \hat{x}_{\theta,i} + (\beta_t/r_t)\hat{x}_{\theta,m} \\ (1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i})/V \end{bmatrix}. \tag{B.31}$$

Since $P_{i,m}$ is affine in (λ, σ),

$$\nabla\mathcal{D}(a, P_{i,m}) = \left(1 - \frac{a}{P_{i,m}}\right) v_{i,m}, \qquad \nabla^2\mathcal{D}(a, P_{i,m}) = \frac{a}{P_{i,m}^2} v_{i,m} v_{i,m}^\top \succeq 0. \tag{B.32}$$

Summing the corresponding terms in (B.26), multiplying by $-\alpha_t'/\alpha_t$, and averaging over positions and $q(\mathbf{x}_t \mid \mathbf{x}_0)$ gives the gradient g and Hessian H of F_t. We use the same constrained damped Newton update

$$d \in \arg\min_{z+d\in\mathcal{S}_{\lambda,\sigma}(t)} \left[g^\top d + \frac{1}{2} d^\top H d \right], \qquad z_{\text{new}} = z + \eta d, \tag{B.33}$$

with backtracking line search.

For every feasible (λ, σ), (B.25) is induced by a normalized NoMUD reverse transition, while the reference path law remains the fixed Markov process. Therefore the corresponding finite-grid objective is a valid NELBO for the selected reverse transition; the NELBO used in evaluation can therefore be minimized over feasible (λ, σ).

B.2.4 A COMMON OPTIMAL CLEAN PREDICTOR

Proposition 2 (Common finite-step predictor). Fix $s < t$, x_t, and a shared feasible pair. Let

$$p_\theta(x_s \mid x_t) = \sum_{x_0} p_\theta(x_0 \mid x_t) q(x_s \mid x_t, x_0). \tag{B.34}$$

Then the true clean posterior

$$\hat{x}_\theta = q(x_0 \mid x_t) \tag{B.35}$$

minimizes

$$\mathbb{E}_{x_0|x_t} \mathrm{KL}(q(x_s \mid x_t, x_0) \,\|\, p_\theta(x_s \mid x_t)).$$

Different feasible pairs can still give different finite-step losses away from this common optimum.

Proof. Averaging the reverse-step KL over the true clean posterior gives

$$\mathbb{E}_{x_0|x_t} \mathrm{KL}(q(x_s \mid x_t, x_0) \,\|\, p_\theta(x_s \mid x_t)) = \sum_{x_0} q(x_0 \mid x_t) \sum_{x_s} q(x_s \mid x_t, x_0) \log \frac{q(x_s \mid x_t, x_0)}{p_\theta(x_s \mid x_t)}.$$

Split the logarithm into numerator and denominator terms. The numerator term does not depend on the model. Interchanging the finite sums in the denominator term gives

$$\begin{aligned} &- \sum_{x_s} \left[\sum_{x_0} q(x_0 \mid x_t) q(x_s \mid x_t, x_0) \right] \log p_\theta(x_s \mid x_t) \\ &\quad = - \sum_{x_s} q(x_s \mid x_t) \log p_\theta(x_s \mid x_t), \end{aligned}$$

where total probability gives

$$q(x_s \mid x_t) = \sum_{x_0} q(x_0 \mid x_t) q(x_s \mid x_t, x_0). \tag{B.36}$$

If $p_\theta(x_0 \mid x_t) = q(x_0 \mid x_t)$, the mixture in (B.34) equals (B.36). Relative to this choice, any other mixture increases the model-dependent term by

$$\sum_{x_s} q(x_s \mid x_t) \log \frac{q(x_s \mid x_t)}{p_\theta(x_s \mid x_t)} = \mathrm{KL}(q(x_s \mid x_t) \,\|\, p_\theta(x_s \mid x_t)) \geq 0.$$

Therefore (B.35) is a finite-step minimizer. □

B.3 REPRESENTATIVE PROCESSES

The transitions below follow by substitution into Propositions 1 and B.1. The NELBOs $\mathrm{NELBO}(\mathbf{x}_0; 0, 0)$, $\mathrm{NELBO}(\mathbf{x}_0; \lambda_t^{\max}, 0)$, and $\mathrm{NELBO}(\mathbf{x}_0; 0, \sigma_t^{\max})$ use the respective token losses below in (10). To keep the formulas compact, define the positive quantities

$$c_t = -\frac{\alpha_t'}{\alpha_t}, \qquad a_t = \frac{\alpha_t}{\beta_t},$$
$$b_t = \frac{\beta_t \lambda_t^{\max}}{r_t}, \qquad e_t = \frac{\sigma_t^{\max}}{V}.$$

B.3.1 ABSORB

At $(\lambda_t, \sigma_t) = (0, 0)$, for $i \neq m$,

$$\Pr(x_t = i \mid x_s = m, x_0) = \mathbb{1}_{m=y} \frac{\delta_{t,s}\alpha_s}{r_s},$$
$$\Pr(x_t = m \mid x_s = m, x_0) = 1 - \mathbb{1}_{m=y} \frac{(V-1)\delta_{t,s}\alpha_s}{r_s}.$$

The reverse transition is

$$q(x_s \mid x_t, x_0) = \mathbb{1}_{m=y} x_0 + \mathbb{1}_{m\neq y} \left(\frac{\beta_t - \beta_s}{\beta_t} x_0 + \frac{\beta_s}{\beta_t} x_t \right).$$

The corresponding token loss is

$$L_t^{\mathrm{Abs}} = c_t a_t \begin{cases} 1 - \hat{x}_{\theta,m}, & m = y, \\ -\log \hat{x}_{\theta,y} - \hat{x}_{\theta,m}, & m \neq y. \end{cases} \tag{B.37}$$

Equation (9) gives, for $i \neq m$,

$$\Pr_\theta(x_s = i \mid x_t) = \frac{\delta_{t,s}\alpha_s}{\beta_t} \hat{x}_{\theta,i}, \qquad \Pr_\theta(x_s = m \mid x_t) - \hat{x}_{\theta,m} = \left(1 - \frac{\delta_{t,s}\alpha_s}{\beta_t}\right)(1 - \hat{x}_{\theta,m}) \geq 0.$$

Thus the sampler places more mass on the current token than the clean prediction does.

B.3.2 FLIP

At $(\lambda_t, \sigma_t) = (\lambda_t^{\max}, 0)$, for $i \neq m$,

$$\Pr(x_t = i \mid x_s = m, x_0) = \delta_{t,s} \left[\mathbb{1}_{m=y} \frac{\alpha_s + \beta_s \lambda_t^{\max}}{r_s} + \mathbb{1}_{i=y} \lambda_t^{\max} \right],$$

with diagonal given by normalization. The reverse transition is

$$\Pr(x_s = i \mid x_t = m, x_0) = \delta_{t,s} \left[\mathbb{1}_{m=y} \frac{\beta_s \lambda_t^{\max}}{r_t} + \mathbb{1}_{i=y} \frac{\alpha_s + \beta_s \lambda_t^{\max}}{\beta_t} \right], \quad i \neq m,$$
$$\Pr(x_s = m \mid x_t = m, x_0) = \mathbb{1}_{m=y} \left(1 - \frac{(V-1)\delta_{t,s}\beta_s \lambda_t^{\max}}{r_t} \right) + \mathbb{1}_{m\neq y} \frac{\beta_s}{\beta_t} (1 - \delta_{t,s}\lambda_t^{\max}).$$

In the continuous-time limit, let

$$P_{i,m}^{\mathrm{F}} = (a_t + \lambda_t^{\max}) \hat{x}_{\theta,i} + b_t \hat{x}_{\theta,m}, \qquad i \neq m.$$

Then the token loss, written without the scalar divergence notation, is

$$L_t^{\mathrm{Flip}} = c_t \begin{cases} \displaystyle\sum_{i\neq m} \left[P_{i,m}^{\mathrm{F}} - b_t + b_t \log \frac{b_t}{P_{i,m}^{\mathrm{F}}} \right], & m = y, \\ \displaystyle\sum_{i\neq m} P_{i,m}^{\mathrm{F}} - (a_t + \lambda_t^{\max}) + (a_t + \lambda_t^{\max}) \log \frac{a_t + \lambda_t^{\max}}{P_{y,m}^{\mathrm{F}}}, & m \neq y. \end{cases} \tag{B.38}$$

For $i, j \neq m$, substitution in (9) yields

$$\Pr_\theta(x_s = i \mid x_t) = \frac{\delta_{t,s}}{\beta_t} (\alpha_s + \beta_s \lambda_t^{\max}) \hat{x}_{\theta,i} + \frac{\delta_{t,s}\beta_s \lambda_t^{\max}}{r_t} \hat{x}_{\theta,m},$$
$$\Pr_\theta(x_s = i \mid x_t) - \Pr_\theta(x_s = j \mid x_t) = \frac{\delta_{t,s}(\alpha_s + \beta_s \lambda_t^{\max})}{\beta_t} (\hat{x}_{\theta,i} - \hat{x}_{\theta,j}).$$

The clean-prediction contrast is amplified relative to Absorb, and $\hat{x}_{\theta,m}$ also contributes to transitions away from m; confident alternatives therefore drive stronger correction.

B.3.3 EXPLORE

At $(\lambda_t, \sigma_t) = (0, \sigma_t^{\max})$, for $i \neq m$,

$$\Pr(x_t = i \mid x_s = m, x_0) = \mathbb{1}_{m=y} \frac{\delta_{t,s}\alpha_s}{r_s} + \mathbb{1}_{m,i\neq y} \frac{\delta_{t,s}\sigma_t^{\max}}{V},$$

with diagonal given by normalization. The reverse transition satisfies

$$\Pr(x_s = i \mid x_t = y, x_0) = \mathbb{1}_{i=y},$$

and, for $m \neq y$,

$$\Pr(x_s = i \mid x_t = m, x_0) = \begin{cases} \delta_{t,s}\alpha_s/\beta_t, & i = y, \\ \delta_{t,s}\beta_s\sigma_t^{\max}/(V\beta_t), & i \neq m, y, \\ (\beta_s/\beta_t)(1 - \delta_{t,s}/\beta_t), & i = m. \end{cases}$$

Here

$$P^{\mathrm{E}}_{i,m} = a_t \hat{x}_{\theta,i} + e_t(1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}), \qquad i \neq m,$$

and the token loss is

$$L_t^{\mathrm{Explore}} = c_t \begin{cases} [a_t + (V-2)e_t](1 - \hat{x}_{\theta,m}), & m = y, \\ \displaystyle\sum_{i\neq m} P^{\mathrm{E}}_{i,m} - a_t - (V-2)e_t + a_t \log \frac{a_t}{P^{\mathrm{E}}_{y,m}} + e_t \sum_{i\neq m,y} \log \frac{e_t}{P^{\mathrm{E}}_{i,m}}, & m \neq y. \end{cases} \tag{B.39}$$

For $i \neq m$, the finite-step model probability and its difference from the clean prediction are

$$\Pr_\theta(x_s = i \mid x_t) = \frac{\delta_{t,s}\alpha_s}{\beta_t}\hat{x}_{\theta,i} + \frac{\delta_{t,s}\beta_s\sigma_t^{\max}}{V\beta_t}(1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}),$$

$$\Pr_\theta(x_s = i \mid x_t) - \hat{x}_{\theta,i} = \left(\frac{\delta_{t,s}\alpha_s}{\beta_t} - 1\right)\hat{x}_{\theta,i} + \frac{\delta_{t,s}\beta_s\sigma_t^{\max}}{V\beta_t}(1 - \hat{x}_{\theta,m} - \hat{x}_{\theta,i}).$$

The second term supplies mass to an alternative even when its predicted probability is small. For $i, j \neq m$, the contrast coefficient is $\delta_{t,s}(\alpha_s - \beta_s\sigma_t^{\max}/V)/\beta_t$. The additional term reduces this coefficient relative to Absorb, encouraging exploration; the coefficient can become negative and reverse the ordering of alternatives.

B.3.4 UNIFORM

At $(\lambda_t, \sigma_t) = (1/V, 1)$, the forward transition is $q(x_t \mid x_s, x_0) = (1 - \delta_{t,s})x_s + \delta_{t,s}\mathbf{u}$, and the backward transition is

$$\begin{aligned} q(x_s \mid x_t, x_0) = \mathbb{1}_{m=y}&\left[\left(1 - \frac{\delta_{t,s}\beta_s}{r_t}\right)x_0 + \frac{\delta_{t,s}\beta_s}{r_t}\mathbf{u}\right] \\ &+ \mathbb{1}_{m\neq y}\left[\frac{\delta_{t,s}\alpha_s}{\beta_t}x_0 + \frac{\beta_s}{\beta_t}(1 - \delta_{t,s})x_t + \frac{\delta_{t,s}\beta_s}{\beta_t}\mathbf{u}\right]. \end{aligned}$$

Training penalties of three processes. We make the qualitative entries in Table 1 precise by expanding each loss around a perfect clean-token prediction. For $j \neq y$, let $\hat{x}^{y\to j}(\varepsilon) = (1-\varepsilon)x_0 + \varepsilon\mathbf{e}_j$ and define $\Delta_j L_t = L_t(\hat{x}^{y\to j}(\varepsilon)) - L_t(x_0)$. The expansions $-\log(1-\varepsilon) = \varepsilon + \varepsilon^2/2 + O(\varepsilon^3)$ and $u - v + v\log(v/u) = (u-v)^2/(2v) + O((u-v)^3)$ apply. For compactness, write $h_A = c_t a_t$, $h_E = c_t[a_t + (V-2)e_t]$, and $h_I = \frac{1}{2}c_t(a_t - e_t)^2(a_t^{-1} + e_t^{-1})$. Then

	$m = y:\ \Delta_i L_t$	$m \neq y:\ \Delta_m L_t$	$m \neq y:\ \Delta_i L_t$
Absorb	$h_A\varepsilon$	$\frac{1}{2}h_A\varepsilon^2 + O(\varepsilon^3)$	$h_A\varepsilon + O(\varepsilon^2)$
Flip	$O(\varepsilon^2)$	$c_t(V-2)b_t\varepsilon + O(\varepsilon^2)$	$c_t(a_t + \lambda_t^{\max})\varepsilon + O(\varepsilon^2)$
Explore	$h_E\varepsilon$	$\frac{1}{2}h_E\varepsilon^2 + O(\varepsilon^3)$	$h_I\varepsilon^2 + O(\varepsilon^3)$

For completeness, the first Flip entry is quadratic with leading term $\frac{c_t}{2b_t}[(a_t + \lambda_t^{\max} - b_t)^2 + (V-2)b_t^2]\varepsilon^2$. Thus a linear term is a stronger local penalty than a quadratic term. When $m = y$, Flip

is low, Absorb is medium, and Explore is high because Explore adds $(V-2)e_t$ to Absorb's linear coefficient. When $m \neq y$ and the error is placed on the current token m, Absorb is low, Flip is high, and Explore is medium: Absorb and Explore are quadratic, but Explore has the larger coefficient. When the error is placed on an unrelated token i, Absorb and Flip are linear (high), whereas Explore is quadratic (low). Moving probability from a wrong token back to y reverses each first-order change; in the quadratic cases, it removes the accumulated penalty to the same leading order.

B.4 Confidence-based sampling

For current token m, write $\hat{x}_{\theta,\max} = \max_i \hat{x}_{\theta,i}$ and $\hat{x}_{\theta,\max\setminus m} = \max_{i\neq m} \hat{x}_{\theta,i}$. Let w_A, w_F, and w_E denote the Absorb, Flip, and Explore weights. For any $\rho > 0$, we set

$$\begin{aligned} w_E &= \frac{1-\hat{x}_{\theta,\max}}{1-1/V}, \\ w_A &= (1-w_E)\frac{\hat{x}_{\theta,m}}{\hat{x}_{\theta,m}+\rho\hat{x}_{\theta,\max\setminus m}}, \qquad w_F = (1-w_E)\frac{\rho\hat{x}_{\theta,\max\setminus m}}{\hat{x}_{\theta,m}+\rho\hat{x}_{\theta,\max\setminus m}}, \\ \lambda_t &= w_F\lambda_t^{\max}, \qquad \sigma_t = w_E\sigma_t^{\max}. \end{aligned} \tag{B.40}$$

Proposition B.5 (Feasibility and confidence limits)**.** Let $\hat{x}_\theta$ be a clean-token probability vector, let m be the current token, and fix any $\rho > 0$. Define w_A, w_F, w_E by (B.40). Then these weights are nonnegative and sum to one. The resulting coupling satisfies

$$0 \le \lambda_t \le \lambda_t^{\max}, \qquad 0 \le \sigma_t, \qquad \sigma_t + \frac{\lambda_t}{1-2/V} \le \sigma_t^{\max}, \tag{B.41}$$

and is independent of s. At fixed interior t,

$$\begin{aligned} \hat{x}_\theta \to x_t &\Rightarrow (\lambda_t,\sigma_t) \to (0,0), \\ \hat{x}_{\theta,i} \to 1\ (i\neq m) &\Rightarrow (\lambda_t,\sigma_t) \to (\lambda_t^{\max},0), \\ \hat{x}_\theta \to \mathbf{u} &\Rightarrow (\lambda_t,\sigma_t) \to (0,\sigma_t^{\max}). \end{aligned} \tag{B.42}$$

These are the Absorb, Flip, and Explore limits.

Proof. For a probability vector, $1/V \le \hat{x}_{\theta,\max} \le 1$, so $0 \le w_E \le 1$. The denominator $\hat{x}_{\theta,m} + \rho\hat{x}_{\theta,\max\setminus m}$ is positive for every $\rho > 0$: either the current token has positive mass, or another token does. Thus $w_A, w_F \ge 0$ and $w_A + w_F = 1 - w_E$. In particular, $w_A + w_F + w_E = 1$.

Equation (B.40) gives $0 \le \lambda_t \le \lambda_t^{\max}$ and $\sigma_t \ge 0$. By the definition of $\sigma_t^{\max}$ and the bound $\beta_t\lambda_t^{\max} \le 1$,

$$\sigma_t + \frac{\lambda_t}{1-2/V} \le (w_E+w_F)\sigma_t^{\max} \le \sigma_t^{\max},$$

which proves (B.41). Neither weight depends on s.

If $\hat{x}_\theta \to x_t$, then $w_E \to 0$, $w_A \to 1$, and $w_F \to 0$. If $\hat{x}_{\theta,i} \to 1$ for $i \neq m$, then $w_E \to 0$, $w_A \to 0$, and $w_F \to 1$. If $\hat{x}_\theta \to \mathbf{u}$, then $w_E \to 1$ and $w_A, w_F \to 0$. Therefore the three limits in (B.42) hold for every $\rho > 0$. We use $\rho = 16$ in the experiments. □

B.5 Distillation algorithm

Following DCD (Sahoo et al., 2025), we freeze the current student as the teacher at the start of each round and double the distillation interval after the round. NoMD additionally updates a coupling head against the exact reference bridge; stop-gradient separates this update from clean-token consistency training.

B.6 Relation to the Ψ-sampler

Proposition B.6 (The superposition reverse transition)**.** Fix $0 < s < t < 1$ and $0 \le \kappa_t \le 1$, with $\beta_t = 1-\alpha_t$, $r_t = V\alpha_t + \beta_t$, and $\delta_{t,s} = 1 - \alpha_t/\alpha_s$. Let $q(x_s \mid x_0) = \alpha_s x_0 + \beta_s \mathbf{u}$. The uniform

Algorithm 2: Non-Markovian discrete consistency distillation

Input: Dataset $\mathcal{D}$; rounds N; steps per round M; interval δ; student θ; EMA student $\bar{\theta}$; coupling head ψ; EMA rate μ.
Output: Distilled model $\bar{\theta}$ and coupling head ψ.
for $r = 1, \dots, N$ **do**
  Freeze the round teacher $\theta^- \leftarrow \mathrm{sg}(\theta)$
  for $j = 1, \dots, M$ **do**
    Sample $\mathbf{x}_0 \sim \mathcal{D}$ and $t \sim \mathrm{Unif}[0,1]$; set $s = \max(t - \delta, 0)$
    Draw $\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)$ and $\mathbf{x}_s \sim q_{\mathrm{ref}}(\mathbf{x}_s \mid \mathbf{x}_t, \mathbf{x}_0)$
    // Match the teacher's cleaner prediction, as in DCD.
    $L_{\mathrm{student}} \leftarrow \mathrm{KL}(\hat{\mathbf{x}}_\theta(\mathbf{x}_t, t) \,\|\, \mathrm{sg}(\hat{\mathbf{x}}_{\theta^-}(\mathbf{x}_s, s)))$
    // Fit the coupling while keeping the denoiser fixed.
    $L_{\mathrm{head}} \leftarrow \mathrm{KL}\big(q_{\mathrm{ref}}(\mathbf{x}_s \mid \mathbf{x}_t, \mathbf{x}_0) \,\big\|\, p_{\mathrm{sg}(\theta),\psi}(\mathbf{x}_s \mid \mathbf{x}_t)\big)$
    Update θ with $\nabla_\theta L_{\mathrm{student}}$ and ψ with $\nabla_\psi L_{\mathrm{head}}$
    $\bar{\theta} \leftarrow \mathrm{sg}\big(\mu\bar{\theta} + (1-\mu)\theta\big)$
  $\delta \leftarrow 2\delta$
return $\bar{\theta}, \psi$

Markov backward distribution is

$$q(x_s \mid x_t, x_0) = \mathbb{1}_{m=y}\left[\left(1 - \frac{\delta_{t,s}\beta_s}{r_t}\right)x_0 + \frac{\delta_{t,s}\beta_s}{r_t}\mathbf{u}\right] + \mathbb{1}_{m\neq y}\left[\frac{\delta_{t,s}\alpha_s}{\beta_t}x_0 + \frac{\beta_s}{\beta_t}(1-\delta_{t,s})x_t + \frac{\delta_{t,s}\beta_s}{\beta_t}\mathbf{u}\right].$$

Define the superposition used by the Ψ-sampler (Deschenaux et al., 2026),

$$\Psi_{s|t}(x_s \mid x_t, x_0) = \kappa_t q(x_s \mid x_t, x_0) + (1-\kappa_t)q(x_s \mid x_0). \tag{B.43}$$

With

$$\lambda_{t,s} = \frac{\kappa_t}{V} + (1-\kappa_t)\frac{r_t}{V\delta_{t,s}}, \qquad \sigma_{t,s} = \kappa_t + (1-\kappa_t)\frac{\beta_t}{\delta_{t,s}}, \tag{B.44}$$

its off-current probabilities are, for $i \neq m$,

$$\Pr_\Psi(x_s = i \mid x_t = m, x_0) = \delta_{t,s}\left[\mathbb{1}_{m=y}\frac{\beta_s\lambda_{t,s}}{r_t} + \mathbb{1}_{i=y}\frac{\alpha_s + \beta_s\lambda_{t,s}}{\beta_t} + \mathbb{1}_{m,i\neq y}\frac{\beta_s\sigma_{t,s}}{V\beta_t}\right], \tag{B.45}$$

with $\Pr_\Psi(x_s = m \mid x_t = m, x_0) = 1 - \sum_{i\neq m}\Pr_\Psi(x_s = i \mid x_t = m, x_0)$. Thus the Ψ-sampler is a member of the same finite-pair reverse family. Its coefficients depend on the selected pair and need not obey the time-only bounds that guarantee feasibility for every earlier time.

Proof. We match the three types of off-current probabilities. Since (B.43) is a mixture of categorical distributions, their equality also fixes the current-coordinate probability.

A clean observation: $m = y$**,** $i \neq y$**.** Substituting (B.44) into (B.45) gives

$$\Pr(x_s = i \mid x_t = y, x_0) = \frac{\delta_{t,s}\beta_s\lambda_{t,s}}{r_t} = \kappa_t\frac{\delta_{t,s}\beta_s}{Vr_t} + (1-\kappa_t)\frac{\beta_s}{V}.$$

The two terms are the uniform Markov backward probability and the time-s wrong-coordinate marginal, weighted as in (B.43).

Two wrong endpoints: $m, i \neq y$**,** $i \neq m$**.** The same substitution gives

$$\Pr(x_s = i \mid x_t = m, x_0) = \frac{\delta_{t,s}\beta_s\sigma_{t,s}}{V\beta_t} = \kappa_t\frac{\delta_{t,s}\beta_s}{V\beta_t} + (1-\kappa_t)\frac{\beta_s}{V},$$

again the corresponding two components of (B.43).

A clean alternative: $m \neq y$, $i = y$. First note the identity

$$\begin{aligned}\delta_{t,s}\alpha_s + \frac{\beta_s r_t}{V} &= \alpha_s - \alpha_t + \beta_s\alpha_t + \frac{\beta_s\beta_t}{V} \\ &= \alpha_s - \alpha_t(1-\beta_s) + \frac{\beta_s\beta_t}{V} = \alpha_s(1-\alpha_t) + \frac{\beta_s\beta_t}{V} = \beta_t\left(\alpha_s + \frac{\beta_s}{V}\right). \qquad \text{(B.46)}\end{aligned}$$

Now expand the clean-alternative probability, and split its $\delta_{t,s}\alpha_s/\beta_t$ term into weights κ_t and $1-\kappa_t$:

$$\begin{aligned}\Pr(x_s = y \mid x_t = m, x_0) &= \frac{\delta_{t,s}(\alpha_s + \beta_s\lambda_{t,s})}{\beta_t} \\ &= \frac{\delta_{t,s}\alpha_s}{\beta_t} + \kappa_t\frac{\delta_{t,s}\beta_s}{V\beta_t} + (1-\kappa_t)\frac{\beta_s r_t}{V\beta_t} \\ &= \kappa_t\frac{\delta_{t,s}(\alpha_s + \beta_s/V)}{\beta_t} + \frac{1-\kappa_t}{\beta_t}\left(\delta_{t,s}\alpha_s + \frac{\beta_s r_t}{V}\right) \\ &= \kappa_t\frac{\delta_{t,s}(\alpha_s + \beta_s/V)}{\beta_t} + (1-\kappa_t)\left(\alpha_s + \frac{\beta_s}{V}\right).\end{aligned}$$

The last equality uses (B.46). This is the clean-coordinate probability in (B.43).

All off-current coordinates agree. In each construction,

$$\Pr(x_s = m \mid x_t = m, x_0) = 1 - \sum_{i \neq m}\Pr(x_s = i \mid x_t = m, x_0),$$

so the current coordinate agrees as well. The superposition also has the required earlier marginal:

$$\begin{aligned}\sum_{x_t} q(x_t \mid x_0)\Psi_{s|t}(x_s \mid x_t, x_0) &= \kappa_t\sum_{x_t} q(x_t \mid x_0)q(x_s \mid x_t, x_0) + (1-\kappa_t)q(x_s \mid x_0)\sum_{x_t} q(x_t \mid x_0) \\ &= \kappa_t q(x_s \mid x_0) + (1-\kappa_t)q(x_s \mid x_0) = q(x_s \mid x_0).\end{aligned}$$

The first sum reduces to $q(x_s \mid x_0)$ by total probability, and the second sum is one. Thus (B.44) gives a categorical coupling for this selected pair. Averaging the identical conditional kernels over the same clean-candidate weights preserves their equality. □

B.7 Non-Markovian masked diffusion and the DDIM comparison

B.7.1 One degree of freedom in masked diffusion

The masked marginal has support only on the clean token x_0 and a mask token $[\mathrm{M}]$:

$$q(x_t \mid x_0) = \alpha_t x_0 + \beta_t \mathbf{e}_{[\mathrm{M}]}.$$

Here $\mathbf{e}_{[\mathrm{M}]}$ is the point mass at the mask. A general conditional forward kernel on these two states is

$$q(x_t \mid x_s, x_0) = \begin{cases} A\mathbf{e}_{[\mathrm{M}]} + Bx_0, & x_s = [\mathrm{M}], \\ C\mathbf{e}_{[\mathrm{M}]} + Dx_0, & x_s = x_0. \end{cases}$$

Normalization gives $A + B = 1$ and $C + D = 1$, and marginal preservation gives $\beta_s A + \alpha_s C = \beta_t$. These three independent constraints leave one degree of freedom. Writing $B = \zeta_{t,s}$ gives

$$q(x_t \mid x_s, x_0) = \begin{cases} (1-\zeta_{t,s})\mathbf{e}_{[\mathrm{M}]} + \zeta_{t,s}x_0, & x_s = [\mathrm{M}], \\ \left(\delta_{t,s} + \dfrac{\beta_s}{\alpha_s}\zeta_{t,s}\right)\mathbf{e}_{[\mathrm{M}]} + \left(\dfrac{\alpha_t}{\alpha_s} - \dfrac{\beta_s}{\alpha_s}\zeta_{t,s}\right)x_0, & x_s = x_0. \end{cases}$$

For $0 < s < t < 1$, nonnegativity is equivalent to $0 \leq \zeta_{t,s} \leq \min\{1, \alpha_t/\beta_s\}$. Applying Bayes' rule yields

$$q(x_s \mid x_t, x_0) = \begin{cases} \dfrac{\beta_s(1-\zeta_{t,s})}{\beta_t}\mathbf{e}_{[\mathrm{M}]} + \dfrac{\beta_t - \beta_s + \beta_s\zeta_{t,s}}{\beta_t}x_0, & x_t = [\mathrm{M}], \\ \dfrac{\beta_s\zeta_{t,s}}{\alpha_t}\mathbf{e}_{[\mathrm{M}]} + \left(1 - \dfrac{\beta_s\zeta_{t,s}}{\alpha_t}\right)x_0, & x_t = x_0. \end{cases}$$

The single degree of freedom determines the reverse probability of remasking a clean token, $\beta_s\zeta_{t,s}/\alpha_t$. When the current token is not masked, its clean value is known and the model reverse kernel equals the exact kernel. When the current token is masked, replacing x_0 by the clean prediction $\hat{x}_\theta$, which assigns no mass to the mask, gives

$$\mathrm{KL}(q(x_s \mid x_t, x_0) \,\|\, p_\theta(x_s \mid x_t)) = -\mathbb{1}_{x_t=[\mathrm{M}]} \frac{\beta_t - \beta_s + \beta_s\zeta_{t,s}}{\beta_t} \log \hat{x}_{\theta,y}.$$

Thus all feasible non-Markovian masked processes retain the same cross-entropy prediction target and change its nonnegative scalar weight. At $\zeta_{t,s} = 0$, the construction reduces to the Markov absorbing process. This is the remasking freedom discussed in Wang et al. (2025).

B.7.2 Shared squared-error target in Gaussian DDIM

For Gaussian marginals $q(x_t \mid x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, \beta_t I)$, a DDIM conditional with variance $\tau_{t,s}^2 I$ has mean

$$\mu(x_t, x_0) = \sqrt{\alpha_s}x_0 + \sqrt{\beta_s - \tau_{t,s}^2}\,\frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{\beta_t}}, \qquad 0 < \tau_{t,s}^2 \le \beta_s.$$

The learned conditional replaces x_0 by $\hat{x}_\theta$ while retaining this covariance. The common x_t term cancels in the mean difference. The equal-covariance Gaussian KL therefore becomes

$$\mathrm{KL}(q\|p_\theta) = \frac{1}{2\tau_{t,s}^2}\left[\sqrt{\alpha_s} - \sqrt{\frac{\alpha_t(\beta_s - \tau_{t,s}^2)}{\beta_t}}\right]^2 \|x_0 - \hat{x}_\theta\|_2^2.$$

The coupling changes a scalar weight but not the prediction error inside the norm (Song et al., 2021). In contrast, the uniform discrete loss in (11) changes its rate factors as the coupling varies.

C Experiments

C.1 Uniform-coupling likelihood evaluation

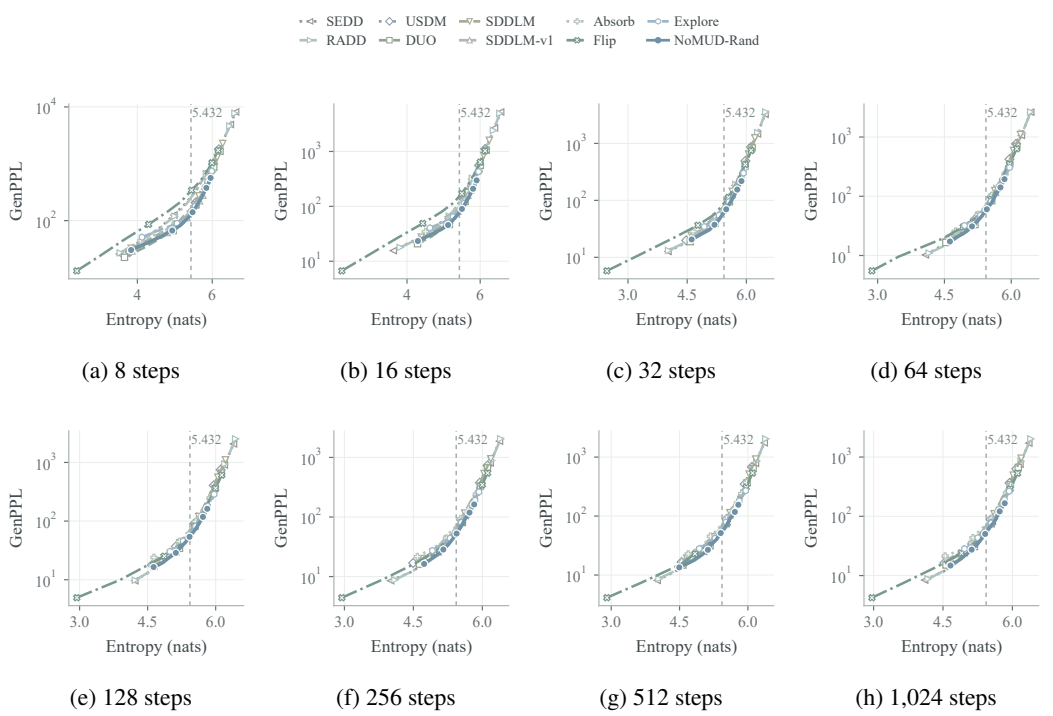
Table 5 compares each NoMUD predictor under the uniform coupling $(1/V, 1)$ and the coupling used in the main evaluation. The model parameters θ are held fixed within each predictor. Uniform evaluation yields substantially lower NELBO values for the fixed-process predictors, especially Flip, while optimizing the coupling for NoMUD-Rand gives the lowest average perplexity, 58.50.

Table 5: Likelihood evaluation results for the four NoMUD predictors under uniform and selected evaluation couplings. Own denotes the corresponding fixed training coupling; Opt denotes the optimized (x_0, t)-dependent coupling used in the model rate factors $P_{i,m}$. For Opt, the reference process remains fixed to the Markov coupling $(1/V, 1)$ and the model parameters θ are fixed.

Predictor	Coupling	PTB	WikiText	LM1B	LAM-BADA	AGNews	PubMed	ArXiv	Avg.
Absorb	Own	137.14	63.21	130.97	102.55	152.97	106.99	74.80	109.80
Absorb	Uniform	99.88	50.31	93.97	75.68	103.19	70.78	57.39	78.74
Flip	Own	1.93×10^6	2.38×10^5	2.92×10^6	6.69×10^5	3.44×10^6	5.21×10^5	2.97×10^5	1.43×10^6
Flip	Uniform	79.56	39.88	80.84	56.46	77.92	51.88	44.54	61.58
Explore	Own	88.26	41.34	84.56	60.25	84.16	59.90	47.55	66.57
Explore	Uniform	85.97	41.52	81.68	56.96	82.20	53.62	45.31	63.89
Rand	Opt	73.55	36.36	75.31	57.68	72.16	51.06	43.36	**58.50**
Rand	Uniform	81.96	36.75	75.21	55.72	76.58	56.45	46.74	61.34

C.2 Generation frontiers

Figure 6 reports all ten models at all eight sampling budgets using the same eleven-temperature sweep. The frontiers show the complete quality–diversity trade-off behind the matched-entropy operating points in Table 3; NoMUD-Rand remains competitive at larger budgets.



(a) 8 steps (b) 16 steps (c) 32 steps (d) 64 steps

(e) 128 steps (f) 256 steps (g) 512 steps (h) 1,024 steps

Figure 6: Complete GenPPL–entropy frontiers for all models and sampling budgets. Each curve contains eleven temperature settings evaluated with 128 samples; dashed lines indicate 5.432 nats.

C.3 Cross-sampler evaluation

We apply the Absorb, Flip, Explore, and confidence-aware samplers to the four NoMUD predictors at all eight budgets. Figure 7 reports GenPPL only. All four predictors support every sampler without retraining, and the randomly trained predictor with confidence-aware sampling achieves the lowest GenPPL at four of the eight budgets.

	Abs.	Flip	Exp.	Conf.
Absorb	147.0	287.3	137.5	**135.9**
Flip	167.8	324.9	**135.0**	137.6
Explore	147.6	314.9	145.6	**130.9**
Rand	**131.0**	288.4	141.2	134.2

(a) 8 steps

	Abs.	Flip	Exp.	Conf.
Absorb	93.3	136.0	92.6	**92.1**
Flip	88.7	150.1	93.1	**86.6**
Explore	96.6	136.1	89.0	**79.2**
Rand	84.3	146.9	86.1	**77.3**

(b) 16 steps

	Abs.	Flip	Exp.	Conf.
Absorb	78.1	87.5	77.0	**67.3**
Flip	68.3	89.9	75.6	**67.6**
Explore	75.8	73.8	70.9	**68.7**
Rand	73.4	76.1	68.1	**60.6**

(c) 32 steps

	Abs.	Flip	Exp.	Conf.
Absorb	75.0	66.5	73.0	**64.9**
Flip	67.5	69.9	65.1	**63.9**
Explore	68.1	68.6	63.6	**59.3**
Rand	65.2	65.3	63.2	**57.1**

(d) 64 steps

	Abs.	Flip	Exp.	Conf.
Absorb	71.2	**62.1**	69.7	63.6
Flip	65.4	62.5	59.1	**51.2**
Explore	69.9	62.8	64.9	**61.2**
Rand	65.3	**56.3**	56.5	57.4

(e) 128 steps

	Abs.	Flip	Exp.	Conf.
Absorb	72.3	62.0	70.6	**55.1**
Flip	67.9	**58.8**	64.4	63.0
Explore	68.9	53.2	62.0	**51.5**
Rand	63.3	56.7	59.5	**51.0**

(f) 256 steps

	Abs.	Flip	Exp.	Conf.
Absorb	74.6	**64.5**	70.6	68.7
Flip	60.1	58.3	68.9	**48.2**
Explore	70.2	57.3	61.8	**52.6**
Rand	63.7	57.5	**53.6**	54.9

(g) 512 steps

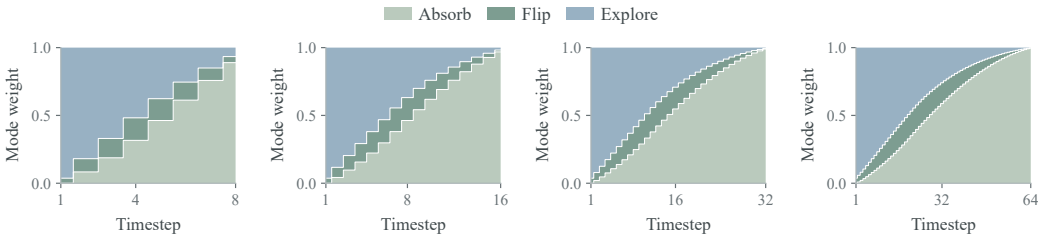
	Abs.	Flip	Exp.	Conf.
Absorb	73.2	66.9	70.0	**62.7**
Flip	63.3	58.1	59.3	**58.0**
Explore	69.4	**52.0**	67.3	66.8
Rand	66.7	53.3	62.9	**53.2**

(h) 1,024 steps

Figure 7: Cross-sampler GenPPL at all eight budgets. Rows identify the four NoMUD predictors and columns identify Absorb, Flip, Explore, and confidence-aware samplers. Each cell uses 1,024 samples; bold values are the lowest within each predictor row.

C.4 SAMPLER MODES

Figure 8 shows the mean confidence-aware coupling weights at every budget. The sampler consistently moves from Explore early, through a larger Flip contribution at intermediate steps, to Absorb near the end, with the same qualitative progression from 8 to 1,024 steps.



(a) 8 steps (b) 16 steps (c) 32 steps (d) 64 steps

(e) 128 steps (f) 256 steps (g) 512 steps (h) 1,024 steps

Figure 8: Mean soft sampler-mode weights at all eight budgets, averaged over 1,024 sequences. The horizontal axis is the sampling timestep; the terminal noise-clearing step is excluded.

C.5 DISTILLATION ABLATION

Table 6 compares removing the coupling head and reversing the KL direction of the clean-token student consistency loss. The full method uses

$$\mathrm{KL}(\hat{x}_\theta(x_t, t) \,\|\, \mathrm{sg}(\hat{x}_{\theta^-}(x_s, s))) \,, \tag{C.1}$$

while the forward-KL student variant replaces it with

$$\mathrm{KL}(\mathrm{sg}(\hat{x}_{\theta^-}(x_s, s)) \,\|\, \hat{x}_\theta(x_t, t)) \,. \tag{C.2}$$

The coupling-head objective is unchanged in the forward-KL variant. The forward-KL student variant performs substantially worse and collapses at 64 steps in our experiment.

Table 6: Distillation ablation at temperature one. The forward-KL student variant reverses the KL direction of the clean-token consistency loss; the coupling-head objective is unchanged.

	No head		Forward-KL student	
Steps	GenPPL	Entropy	GenPPL	Entropy
8	71.81	5.050	214.90	5.679
16	52.19	5.181	138.13	5.667
32	45.16	5.228	114.64	5.645
64	**40.44**	5.228	**1120.11**	4.216

C.6 SAMPLING EXAMPLES

The examples below expose the state-dependent coupling decisions along individual trajectories. Early uncertain states use Explore, Flip corrects selected tokens once alternatives become confident, and Absorb retains settled tokens. The compact example shows the same progression in a localized excerpt, while the complete example preserves the full 1,024-token output.

Confidence-aware coupling in one generation trajectory

NoMUD-Rand, Conf; $N = 128$, Temperature $= 0.957$.
Full-sequence GenPPL: 47.163; unigram entropy: 5.455 nats.

Absorb / retain **Flip / correct** **Explore**

Intermediate states: the same 29 token positions

After reverse step 16/128.
please..... concentfinger Saw the respondedisons calmly siteCoHoursample Tek capitalist CampaignXY guidingR FBI Bryintensive We lifting creditor hear any Nau level
Explore: `sounds` → `capitalist`; $w_E = 0.964$.

After reverse step 99/128.
please feel concent to follow the progress of the site, our Facebook page, and follow femaleRidesSpace. We will happily hear any feedback.
Absorb: `feedback` → `feedback`; $w_A > 0.999$.

After reverse step 100/128.
please feel free to follow the progress of the site, our Facebook page, and follow femaleRidesSpace. We will happily hear any feedback.
Flip: `concent` → `free`; $w_F = 0.997$.

Final output (a contiguous 63-token excerpt)

So any viewer suggestions using the format currently installed are welcome and your opinion will be promptly addressed. If you have suggestions for improvements to the viewer's modified method, please feel free to follow the progress of the site, our Facebook page, and follow InnerRidesSpace. We will happily hear any feedback.

Confidence-aware sampling: a complete 1024-token example

NoMUD-Rand, Conf; $N = 1024$, Temperature $= 0.850$.
Full-sequence GenPPL: 25.753; unigram entropy: 5.197 nats.

Absorb / retain **Flip / correct** **Explore**

Selected intermediate states (omissions indicated by $\cdots$)

After reverse step 114/1024.
$\cdots$ READ Shiningieved Amount push Patel safe patiently Sauce $(Dynasty Sadd $\cdots$ unarmed Trilogy one control Lucifer representative collegeiblings SirenVaults bruising rural $\cdots$ ocene innovate smart',extreme cervical Salad Dota Gro atmosphere penetrated journalist $\cdots$
Explore: `ideologies`→`Patel`; Explore: `batted`→`representative`; Explore: `Destruction`→`cervical`.

After reverse step 487/1024.
$\cdots$ that people can speak to each other. We are allidable $\cdots$
Absorb: `people`→`people`; Absorb: `speak`→`speak`; Flip: `boxed`→`each`.

After reverse step 972/1024.
$\cdots$ get in Dota is player communication. Why? While I did $\cdots$ players feel no longer the responsibility, it has been left to $\cdots$ buck can speak to their representatives, so that people can speak $\cdots$
Absorb: `communication`→`communication`; Absorb: `responsibility`→`responsibility`; Absorb: `representatives`→`representatives`.

Final output: all 1024 tokens, unedited

make me feel better. I'm going to stay here, which is exactly what I came here for.

CDR: I've been joking about playing and preparing for Fnatic since long. It helps so much for them. You even bring tanks for home. Realizing that I am preparing with them, as well as with my friends and family, I can now actually play in Korea. Just this year I played a lot in Korea.

CDR: Just recently I think this was a pretty huge mistake for me and my family. You used to have a training schedule during the entire summer. What were some direct training issues for you?

DE: There were a few things when I went in the following year, that I wasn't going in the summer. Namely I've learned so much. Sometimes they just go through a gradual process. Some days when they can get

Confidence-aware sampling: complete output (continued)

their own opinions right. And then sometimes when it happens, maybe a few months from now, and they look at a new event, thinking things will go down and no other event happens. Sometimes they will prefer to leave to Korea...

After Hong Kong, I started taking out a commentator, Mike Leo, currently, he streams a lot, and introduced himself to me. I think I anticipate that because the coaching staff are perfecting themselves for this new job. In general the biggest difficult thing to get in Dota is player communication. Why? While I did a lot of training in Overwatch, I've not had some sort of communication in China that allowed me to learn in the early days here.

CDR: So, can you please focus on the rest of the team? I thank you very very much because I'm going to be focusing on helping them in the next few months now.

DE: First of all, communication. Miscommunication does not come down to the right coaches. We still don't have the time to tell everyone we know our contacts and say that our communication is best for someone who has their perspective on something. We now don't have the time to try and give players the chance to express themselves in different ways, and that would be confirmed during DH weekend. The first thing that really needs to change is exactly the communication, and from the way we communicate, the way we talk to others, and things like that. It helps me with more than just Dota 2. Contacts and direct communication should be met a few weeks from now. So when a player is thinking of the future, they need to be informed. By this, the players feel no longer the responsibility, it has been left to the coaches.

We reach for the same thing when people talk to us. Miscommunication is important because it always has to do with how long it's going to last. The team has a structure where people can talk directly. People can speak to their representatives, so that people can speak to each other. We are all forced to be very accountable. When there is direct communication between the players, they're honest.

CDR: Is this critical?

In the next few months, it will become obvious that this will be the most important thing ever. Right now we'll be seeing a lot Chinese players saying directly, "I care very much" and "I want the members of the Team Liquid." Now we will see what the players say without knowing when it was expressed, and they'll say, "I think I'm for the team, and I was famous by doing it. So I think if we want to stay transparent, I think we can't hide our members." What is important really is for our fellow players to be able to see that their words are their words.

CDR: How do you envision the Team Liquid future?

DE: Well, no one can I say in the past had so much fun we became so quickly when a lot of people joined in on the team. It is high level at the moment. And they will remember that time when one of us in college was in the corner holding a banner of 75. Nada said the team would put it in front of him because he thinks that he wanted to be the Grandmaster. If he no longer believes the Nada will put a massive banner in the front of him in the future. "What we are going to do" to be what the Nada will say.

And just to give this one last, I would thank the fans for filling the entire room in the venue bo, in boa boa. They are there for filling in with the booze, and creating the first Dota 2 atmosphere, which is great because people are able to talk to each other.

CDR: It's a real blast to have people speaking to you with so few words.

DE: I think it