One-step Language Modeling via Continuous Denoising

Chanhyuk Lee[1] **Jaehoon Yoo**[1] **Manan Agarwal**[2] **Sheel Shah**[2] **Jerry Huang**[2]
Aditi Raghunathan[2] **Seunghoon Hong**[1] **Nicholas M. Boffi**[†2] **Jinwoo Kim**[†1]

Abstract

Language models based on discrete diffusion have attracted widespread interest for their potential to provide faster generation than autoregressive models. In practice, however, they exhibit a sharp degradation of sample quality in the few-step regime, failing to realize this promise. Here we show that language models leveraging flow-based continuous denoising can outperform discrete diffusion in both quality and speed. By revisiting the fundamentals of flows over discrete modalities, we build a flow-based language model (FLM) that performs Euclidean denoising over one-hot token encodings. We show that the model can be trained by predicting the clean data via a cross entropy objective, where we introduce a simple time reparameterization that greatly improves training stability and generation quality. By distilling FLM into its associated *flow map*, we obtain a distilled flow map language model (FMLM) capable of few-step generation. On the LM1B and OWT language datasets, FLM attains generation quality matching state-of-the-art discrete diffusion models. With FMLM, our approach outperforms recent few-step language models across the board, with *one-step generation* exceeding their 8-step quality. Our work calls into question the widely held hypothesis that discrete diffusion processes are necessary for generative modeling over discrete modalities, and paves the way toward accelerated flow-based language modeling at scale.

1. Introduction

Today's frontier language models (LMs) are based on an autoregressive process that produces one subword (token) per step (Achiam et al., 2023; Anil et al., 2023; Guo et al., 2025). While these models leverage parallelism during training through teacher forcing and a transformer-based architecture, their sampling is inherently serial in nature, posing a bottleneck in generation speed. Recently, language models based on discrete diffusions and flows have attracted interest as a possible solution (Khanna et al., 2025; Google DeepMind, 2025; Song et al., 2025). By learning to reverse a noising process on full sequences, such models can output multiple tokens in parallel at each sampling step, thereby holding the potential for accelerated generation.

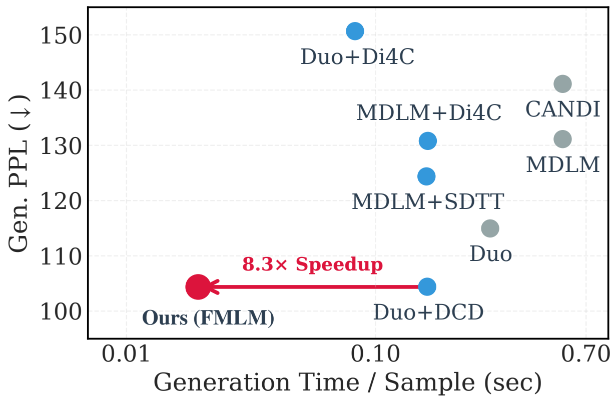

Figure 1. Our flow map language model (FMLM) outperforms discrete diffusion models (gray) and matches the 8-step generation performance of few-step distilled discrete diffusion models (blue) in only **one step**, achieving an ≈ 8.3× speedup on LM1B.

Despite their promise, discrete diffusion language models have significant practical limitations. In particular, they suffer from a rapid drop-off of quality in the few-step generative regime (Deschenaux & Gulcehre, 2024). This is critical as diffusion models process the *full sequence* at each inference step, so that sampling steps need to be substantially reduced to compensate for the associated cost compared to their autoregressive counterparts (Dieleman, 2023; Zheng et al., 2024). This difficulty arises from a fundamental computational constraint: the state space over full sequences is combinatorially large, necessitating a factorized approximation of the transition probability that neglects inter-token correlations (Wu et al., 2025; Kang et al., 2025). While this approximation makes discrete diffusion computationally feasible, empirically it requires many steps to accurately

† Equal advising [1]KAIST [2]Carnegie Mellon University. Correspondence to: Jinwoo Kim <jinwoo-kim@kaist.ac.kr>, Nicholas M. Boffi <nboffi@andrew.cmu.edu>. Code is available at `https://github.com/david3684/flm`.

Preprint. February 18, 2026.

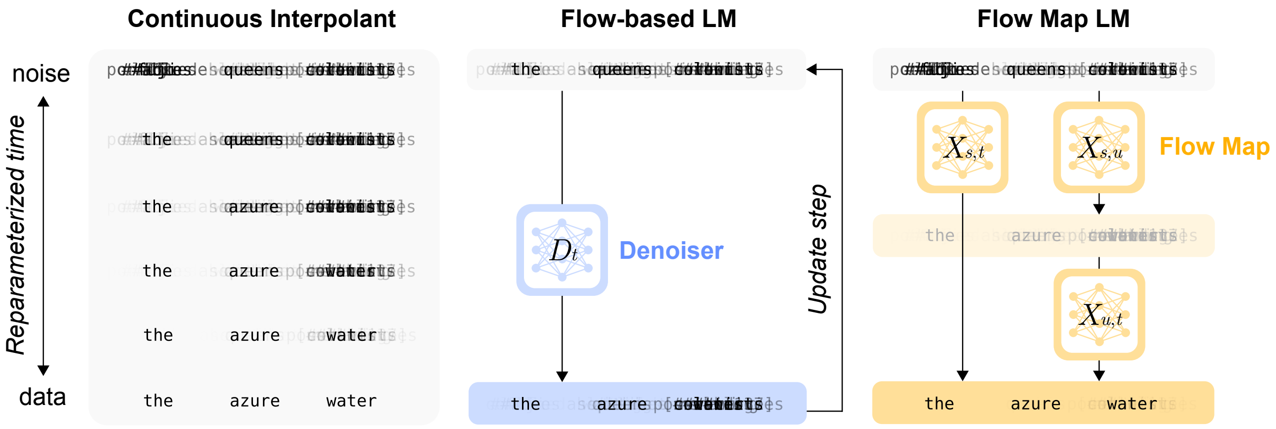

Figure 2. **Overview.** *Left:* We leverage continuous interpolation between Gaussian noise and one-hot language encoding. *Middle:* Our flow-based language model (FLM) learns a denoiser that predicts clean data, which we convert into a flow for sampling. *Right:* Our distilled flow map language model (FMLM) directly transports states between distant timepoints, enabling few-step generation.

capture full sequence structure, implying a fundamental rigidity that limits practical use.

In contrast, continuous diffusion language models, which represent and denoise subwords in a continuous space, do not rely on such approximations (Li et al., 2022; Dieleman et al., 2022). As a result, they can perform accurate parallel generation through a *deterministic* evolution driven by a velocity or score function (Lipman et al., 2022; Albergo et al., 2023; Song et al., 2020b). Perhaps most interestingly, this makes them compatible with recent advances in few-step distillation methods that learn the *flow map*, an operator that directly transports noise to data in as few as one function evaluation (Boffi et al., 2025a;b). Yet, despite their potential advantages, a widely held belief is that continuous diffusion language models underperform their discrete counterparts (Sahoo et al., 2025; Pynadath et al., 2025), leading practitioners to prioritize discrete methods in recent years (Nie et al., 2025; Khanna et al., 2025).

In this work, **we challenge this widespread belief**, showing that continuous diffusion language models formulated via flow matching and flow maps can be higher-performing and faster than previously believed (Fig. 1). We argue that their observed weak performance stems from suboptimal design, such as the choice of temporal weighting, rather than an inherent limitation of the model class. In particular, our approach (Fig. 2) reaches the performance of state-of-the-art (SoTA) discrete diffusion models and exceeds them in the few-step regime.

Overall, our **main contributions** are:

- We build FLM, a powerful flow-based language model, via a principled reexamination of design choices that reveals the root of underperformance in prior methods.
- We introduce FMLM, a flow map language model capable of few-step generation, by expanding FLM via efficient distillation methods.
- We validate our approach empirically on the One Billion Word (LM1B) and OpenWebText (OWT) datasets. FLM is competitive in generation quality with SoTA discrete diffusion LMs, while FMLM beats recent few-step LMs, nearing their 8-step quality at *one step*.

2. Background & Related Work

In this section, we provide an overview of the background and related work. An extended discussion is in App. A. Let V be a vocabulary of subwords, treated as integers $[|V|]$ without loss of generality. We denote language data with length L by $\mathbf{y} = (\mathbf{y}^l)_{l=1}^L \in V^L$. The goal of language modeling is to estimate the data distribution $p(\mathbf{y})$ on V^L.

Autoregressive language models factorize the data distribution over length, $p(\mathbf{y}) = p(\mathbf{y}^1)p(\mathbf{y}^2|\mathbf{y}^1)\dots p(\mathbf{y}^L|\mathbf{y}^{<L})$, and learn the conditional distribution $p(\mathbf{y}^l|\mathbf{y}^{<l})$ over subwords given a prefix (Jordan, 1986; Elman, 1990; Bengio et al., 2003). They generate text by sequentially sampling each token conditioned on the past generation. This process is serialized, with each token awaiting all previous tokens, limiting efficiency (Gu et al., 2017; Stern et al., 2018).

Discrete diffusion language models aim to achieve faster generation by producing several tokens in parallel at each step (Austin et al., 2021; Lou et al., 2023). They employ a discrete noising process such as masking (Sahoo et al., 2024) or uniform randomization (Lou et al., 2023; Schiff et al., 2024) of multiple tokens, and model its reversal $p_t(\mathbf{y}_t)$, which transports an initial distribution p_0 to the data distribution $p_1 = p$. To do so, the model must learn the transition probabilities $p_{t|s}(\mathbf{y}_t|\mathbf{y}_s)$ (Austin et al., 2021; Gat et al., 2024) so that generation can be performed via ancestral

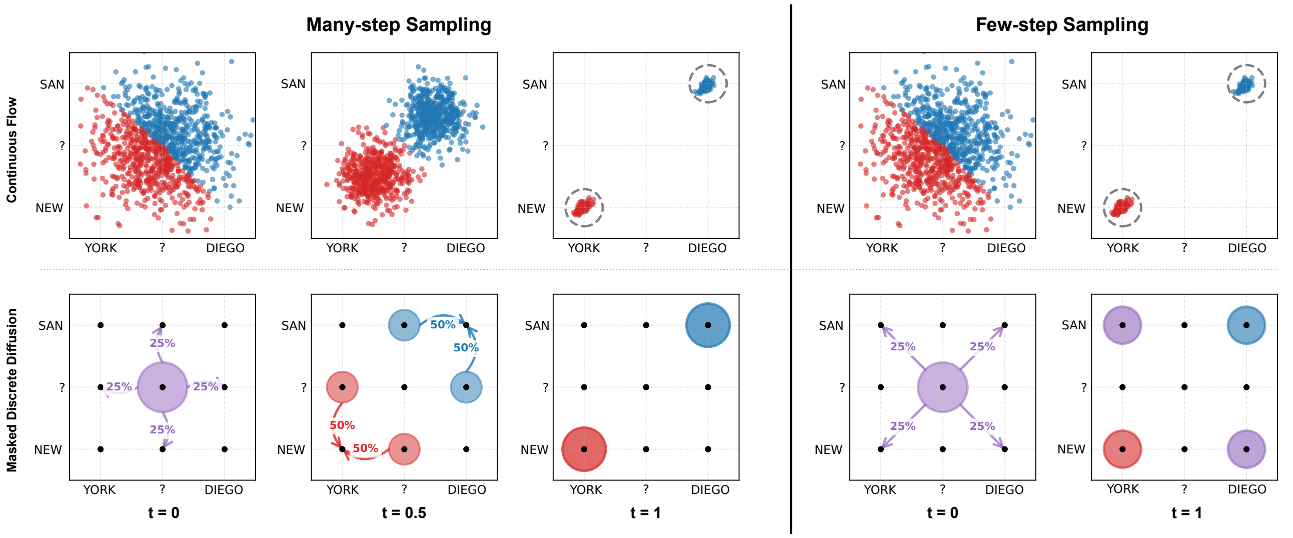

Figure 3. **Factorization error in discrete diffusion.** A toy dataset with two correlated modes `new-york` and `san-diego`. *Left:* In many-step sampling, both continuous flow and discrete diffusion generate valid data. *Right:* In few-step sampling, the factorized transition of discrete diffusion yields a spurious mixture of all possible combinations (including invalid pairings `new-diego` and `san-york`).

sampling over a temporal grid $0 = t_0 < \ldots < t_N = 1$. Since each step updates multiple tokens simultaneously, substantial speedups could be achieved in principle if few-step generation were possible.

In practice, however, discrete diffusion models often fail catastrophically in the few-step regime (Deschenaux & Gulcehre, 2024). This failure is rooted in a fundamental computational challenge: since the transition probability is defined over the full text space V^L, learning it accurately requires a model with an intractable $|V^L| - 1$ output dimensionality. To sidestep this problem, discrete methods employ a *factorized approximation* $p^\star_{t|s} \approx p_{t|s}$ of the transition probability:

$$p^\star_{t|s}(\mathbf{y}_t|\mathbf{y}_s) \coloneqq p^1_{t|s}(\mathbf{y}^1_t|\mathbf{y}_s) \cdots p^L_{t|s}(\mathbf{y}^L_t|\mathbf{y}_s), \qquad (1)$$

where $p^l(\mathbf{y}^l)$ is the probability of the l-th token marginalized over the others. While this approximation makes learning tractable, it makes a restrictive assumption that the denoised tokens $\mathbf{y}^1_t, ..., \mathbf{y}^L_t$ are conditionally independent of each other given the earlier denoised state. This assumption only holds in the infinitesimal limit $t \to s$ (Campbell et al., 2022), which implies that one needs a large number of sampling steps for accurate generation. With reduced steps, the approximation causes the model to produce unnatural text by neglecting correlations over tokens (Fig. 3) (Wu et al., 2025; Kang et al., 2025). Unfortunately, this is a model misspecification issue that cannot be resolved by improving model quality alone. Our work addresses this fundamental challenge with a continuous flow-based formulation (Lipman et al., 2022; Albergo et al., 2023), which learns a deterministic transport map that need not make such a factorized approximation. As a result, our approach directly enables **scalable one-step language modeling**.

3. Theoretical Framework

In this section, we describe our formulation of a continuous flow-based language model (FLM), as well as its few-step flow map (FMLM). To do so, we develop a generative model over one-hot encodings of language, leveraging flow matching over stochastic interpolants. Full details of the framework can be found in App. B and C, where we give a complete background on flow maps and describe both *distillation* and *direct training* algorithms for FMLMs.

3.1. A continuous representation of language

A natural way to build a continuous flow over language data $\mathbf{y} \in V^L$ is to first construct a continuous representation of the data, and then to apply flow-based modeling. Formally, we choose a continuous representation $f : \mathbf{y} \mapsto \mathbf{x}$ and a decoder $g : \mathbf{x} \mapsto \mathbf{y}$ satisfying $g(f(\mathbf{y})) = \mathbf{y}$, and we model the induced distribution $p(\mathbf{x})$ on the continuous space:

$$p(\mathbf{x}) = p(\mathbf{y} = g(\mathbf{x})). \qquad (2)$$

Inference can be performed by first generating $\hat{\mathbf{x}} \sim p(\mathbf{x})$, and then decoding into discrete language through $\hat{\mathbf{y}} = g(\hat{\mathbf{x}})$.

Several choices of continuous representation have been explored in prior work, including learned embeddings (Li et al., 2022; Dieleman et al., 2022; Gulrajani & Hashimoto, 2023) and pretrained embeddings (Strudel et al., 2022; Lovelace et al., 2023). However, learned embeddings require careful regularization to prevent collapse or explosion, while pretrained embeddings may not be optimal for the flow.

Here, we adopt a simple and canonical tokenwise one-hot representation $f : V^L \to \mathbb{R}^{L \times |V|}$ with an argmax decoder

$g : \mathbb{R}^{L\times|V|} \to V^L$ that are pre-specified and frozen:

$$f : \mathbf{y} \mapsto [\text{onehot}(\mathbf{y}^1), ..., \text{onehot}(\mathbf{y}^L)]^\top, \tag{3}$$

$$g : \mathbf{x} \mapsto [\text{argmax}(\mathbf{x}^1), ..., \text{argmax}(\mathbf{x}^L)]. \tag{4}$$

This choice offers a lossless representation of discrete tokens, and requires neither regularization nor auxiliary training. Similar representations have been explored in prior work on continuous diffusion for language (Chen et al., 2022; Han et al., 2023; Mahabadi et al., 2024), though often with additional constraints such as simplex projection. As we elaborate upon below, our approach operates in an unconstrained Euclidean space, which we find to be simpler and more effective in practice.

3.2. Interpolants and flows for language modeling

Given a choice of continuous representation, the language modeling problem becomes that of learning a continuous data distribution $p(\mathbf{x})$. To build a high-performance model, we follow Lipman et al. (2022) and Albergo et al. (2023), leveraging flow matching over a stochastic interpolant. This leads to a simple formulation of our method and matches the design of state-of-the-art flows for continuous data (Li & He, 2025; Zheng et al., 2025a).

In the stochastic interpolant framework, we specify a probability path $p_t(\mathbf{x}_t)$ as the density of an interpolant between noise $\mathbf{x}_0 \sim p_0 = \mathsf{N}(0, I)$ and data $\mathbf{x}_1 \sim p_1$:

$$I_t := (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad I_t \sim p_t. \tag{5}$$

Above, we choose a simple and canonical linear stochastic interpolant, but many choices are possible in practice. By choosing different interpolants, our discussion can extend to variance-exploding (Song & Ermon, 2019) and variance-preserving (Ho et al., 2020) diffusions (Lipman et al., 2022).

This probability path admits a deterministic representation that can be used to produce a sample $\mathbf{x}_t \sim p_t$,

$$\dot{\mathbf{x}}_t = b_t(\mathbf{x}_t), \quad \mathbf{x}_0 \sim p_0, \quad t \in [0, 1], \tag{6}$$

where b_t is the velocity field of the probability flow,

$$b_t(\mathbf{x}) = \mathbb{E}[\dot{I}_t | I_t = \mathbf{x}] = \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 | I_t = \mathbf{x}]. \tag{7}$$

Above, the expectation is with respect to the random draws of $\mathbf{x}_0 \sim p_0$ and $\mathbf{x}_1 \sim p_1$ that are used to construct the interpolant. The conditional expectation form (7) means that the velocity can be learned efficiently by solving a square loss regression problem $b = \text{argmin}_{\hat{b}} \mathcal{L}_{\mathsf{MSE}}(\hat{b})$, where:

$$\mathcal{L}_{\mathsf{MSE}}(\hat{b}) := \int_0^1 \mathbb{E}|\hat{b}_t(I_t) - \dot{I}_t|^2 \mathrm{d}t. \tag{8}$$

In practice, (8) is estimated by sampling t, for example uniformly over the interval $t \sim \mathsf{U}[0, 1]$. After minimizing (8) over a parametric class of neural networks, a sample approximately following p_1 can be obtained by using the estimate $\hat{b}_t$ in the numerical integration of (6). Integration is performed over a choice of temporal grid $0 = t_0 < t_1 < \ldots < t_N = 1$, for example with the forward Euler method.

Denoiser. Despite its simplicity, learning the velocity directly incurs a potential difficulty in our setuw. Velocity prediction requires regressing onto a *noised* target $\dot{I}_t = \mathbf{x}_1 - \mathbf{x}_0$, which inherits the full-rank structure of Gaussian noise samples $\mathbf{x}_0 \in \mathbb{R}^{L\times|V|}$. When the dimensionality $|V|$ is much larger than the internal feature dimension d of the network, this can lead to underfitting due to a rank bottleneck (Li & He, 2025; Zheng et al., 2025a). To avoid this issue, we predict the clean data $\mathbf{x}_1$, which also lies in $\mathbb{R}^{L\times|V|}$ but is constrained to one-hot encodings that form a highly structured, low-entropy target. We therefore predict the conditional expectation of the clean data, which relates to the velocity through a linear change of variables (Albergo et al., 2023; Li & He, 2025):

$$D_t(\mathbf{x}) := \mathbb{E}[\mathbf{x}_1 | I_t = \mathbf{x}], \quad b_t(\mathbf{x}) = \frac{D_t(\mathbf{x}) - \mathbf{x}}{1-t}. \tag{9}$$

The function D_t, called the "denoiser", can be learned via another regression $D = \text{argmin}_{\hat{D}} \mathcal{L}_{\mathsf{MSE}}(\hat{D})$, where:

$$\mathcal{L}_{\mathsf{MSE}}(\hat{D}) := \int_0^1 \mathbb{E}|\hat{D}_t(I_t) - \mathbf{x}_1|^2 \mathrm{d}t. \tag{10}$$

Importantly, in our discrete generative modeling setting the denoiser admits a simple probabilistic interpretation:

Lemma 3.1. *At each token position l, the optimal denoiser output equals the posterior probability over the vocabulary:*

$$D_t(\mathbf{x})^l = p^l_{1|t}(\cdot | I_t = \mathbf{x}). \tag{11}$$

A proof is in App. D.1. Since the optimal denoiser lies on the probability simplex $\Delta^{|V|-1}$, it is advantageous to parameterize $\hat{D}$ via a tokenwise softmax. This restricts the hypothesis space to valid probability distributions, allowing the model to focus on estimating the correct posterior rather than also learning the simplex structure. Furthermore, this enables training through a categorical cross-entropy loss (Dieleman et al., 2022; Eijkelboom et al., 2024), which is more adapted to the one-hot geometry:

Proposition 3.2. *With the change of variables in (11), the denoiser can be learned via $D = \text{argmin}_{\hat{D}} \mathcal{L}_{\mathsf{CE}}(\hat{D})$ where:*

$$\mathcal{L}_{\mathsf{CE}}(\hat{D}) := \int_0^1 \mathbb{E}\left[-\sum_{l=1}^{L} \log \hat{p}^l_{1|t}(\mathbf{x}^l_1 | I_t)\right] \mathrm{d}t. \tag{12}$$

A proof is in App. D.2. In practice, cross-entropy provides a well-conditioned loss landscape for learning on the simplex, yielding stronger gradients than squared loss when predictions are far from the target (Golik et al., 2013).

Relationship with discrete diffusion. Lemma 3.1 suggests that the optimal denoiser implicitly learns the *factorized* posterior $p^*_{1|t}(\mathbf{x}_1|\mathbf{x}_t)$ (1). This reveals an interesting connection with discrete diffusion models, which also often learn $p^*_{1|t}$ via a tokenwise cross-entropy objective (Austin et al., 2021; Campbell et al., 2022; Gat et al., 2024). While discrete models use the learned $p^*_{1|t}$ to perform ancestral sampling, requiring the joint probability and thus suffering from factorization errors, continuous models use the learned $p^*_{1|t}$ to infer the *exact* velocity based on (11) and (9).

3.3. Flow maps for few-step language modeling

The framework in Sec. 3.2 does not immediately allow for few-step language modeling, since the numerical solvers used to integrate (6) typically become inexact at large step sizes. Here we overcome this challenge by leveraging the *flow map* $X_{s,t} : \mathbb{R}^{L\times|V|} \to \mathbb{R}^{L\times|V|}$. The flow map is the solution operator of (6), and by definition directly transports between any two timepoints (Boffi et al., 2025a;b):

$$X_{s,t}(\mathbf{x}_s) = \mathbf{x}_t, \quad \text{for all } (s,t) \in [0,1]^2. \tag{13}$$

Without loss of generality, we may parameterize it as:

$$X_{s,t}(\mathbf{x}) = \mathbf{x} + (t-s)v_{s,t}(\mathbf{x}), \tag{14}$$

where v is called the average velocity or “mean flow” (Geng et al., 2025a;b). Given a flow map, sampling $\hat{\mathbf{x}}_1 \sim p_1$ can be done by choosing a temporal grid $0 = t_0 < ... < t_N = 1$ and sequentially evaluating $\hat{\mathbf{x}}_{t_{i+1}} = X_{t_i,t_{i+1}}(\hat{\mathbf{x}}_{t_i})$. Unlike numerical integration of (6), this approach is accurate for an arbitrary time grid by definition, enabling sampling in as few as one function evaluation via $\hat{\mathbf{x}}_1 = X_{0,1}(\mathbf{x}_0)$. In practice, leveraging a few steps typically improves performance.

Methods for learning flow maps (Boffi et al., 2025a;b) leverage the following mathematical properties, which fully characterize the flow map under standard regularity conditions:

$$\begin{aligned} X_{s,s} &= \mathrm{id}, && \text{(boundary condition)}\\ \lim_{s\to t} \partial_t X_{s,t} &= b_t, && \text{(tangent condition)}\\ X_{u,t}(X_{s,u}(\mathbf{x})) &= X_{s,t}(\mathbf{x}). && \text{(semigroup condition)} \end{aligned}$$

The last condition can be replaced with alternatives based on a differential characterization of the flow map, which lead to learning objectives that require the computation of Jacobian-vector products, such as MeanFlow (Geng et al., 2025a;b) and Lagrangian self-distillation (Boffi et al., 2025a;b; Zhou et al., 2025) (see App. B.1). We use the semigroup condition due to its simplicity, which is related to progressive distillation (Salimans & Ho, 2022) and shortcut models (Frans et al., 2024). The alternatives can also be directly used in our framework, and we leave their study to future work.

Using the semigroup condition, the flow map $X_{s,t}$ can be learned via an optimization $X = \operatorname{argmin}_{\hat{X}} \mathcal{L}_{\mathsf{MSE}}(\hat{X})$ subject to the boundary and tangent conditions, where:

$$\mathcal{L}_{\mathsf{MSE}}(\hat{X}) := \int_0^1 \int_0^t \int_s^t \mathbb{E}|\hat{X}_{s,t}(I_s) - \mathsf{sg}(\hat{X}_{u,t}(\hat{X}_{s,u}(I_s)))|^2 \mathrm{d}u\mathrm{d}s\mathrm{d}t, \tag{15}$$

with $\mathsf{sg}(\cdot)$ denoting the stop-gradient operator.

The two-time denoiser. In Sec. 3.2, we introduced the denoiser D_t to reparameterize the velocity b_t into a simplex-valued clean-data predictor, enabling training via cross entropy. We now develop an analogous reparameterization for the flow map. To do so, we observe a rearrangement of (9) into $D_t(\mathbf{x}) = \mathbf{x} + (1-t)b_t(\mathbf{x})$, showing that the denoiser equals a single Euler step of size $1-t$ with the velocity field. Mirroring this relationship, we define a new quantity we refer to as the *two-time denoiser*:

$$\delta_{s,t}(\mathbf{x}) := \mathbf{x} + (1-s)v_{s,t}(\mathbf{x}), \tag{16}$$

which takes a single step of the average velocity $v_{s,t}$ using the full remaining time $1-s$. From (14), we can see that the flow map is expressed as the following convex combination:

$$X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\mathbf{x} + \frac{t-s}{1-s}\delta_{s,t}(\mathbf{x}). \tag{17}$$

Hence, the properties characterizing the flow map can be converted into those on the two-time denoiser. The boundary condition is satisfied by construction, and the tangent and semigroup conditions translate respectively into:

$$\delta_{s,s}(\mathbf{x}) = D_s(\mathbf{x}), \tag{18}$$

$$\delta_{s,t}(\mathbf{x}) = \gamma\delta_{s,u}(\mathbf{x}) + (1-\gamma)\delta_{u,t}(X_{s,u}(\mathbf{x})), \tag{19}$$

where $\gamma = \frac{(1-t)(u-s)}{(1-u)(t-s)} \in [0,1]$. A proof is given in App. C.1. We refer to (18) as the diagonal condition.

Importantly, we find that at each token position, the output of the two-time denoiser δ always lies in the simplex $\Delta^{|V|-1}$ analogous to the one-time denoiser D. A proof is in App. C.1. This motivates learning δ using the semigroup condition (19) via cross entropy subject to the diagonal condition (18), by optimizing $\delta = \operatorname{argmin}_{\hat{\delta}} \mathcal{L}_{\mathsf{CE}}(\hat{\delta})$ (App. C.3):

$$\mathcal{L}_{\mathsf{CE}}(\hat{\delta}) := \int_0^1 \int_0^t \int_s^t \mathbb{E}\Big[\ell_{\mathsf{CE}}\Big(\hat{\delta}_{s,t}(I_s), \mathsf{sg}(\gamma\hat{\delta}_{s,u}(I_s) + (1-\gamma)\hat{\delta}_{u,t}(\hat{X}_{s,u}(I_s)))\Big)\Big]\mathrm{d}u\mathrm{d}s\mathrm{d}t. \tag{20}$$

Similarly to D, learning δ using cross entropy may benefit from a well-behaved loss landscape, so we test it alongside the average velocity (14) and squared loss (15) formulation. In App. C, we give a complete characterize of the two-time denoiser $\delta_{s,t}$, where we explain how to build both distillation and direct training algorithms via self-distillation for the flow map that respect the one-hot geometry of discrete data.

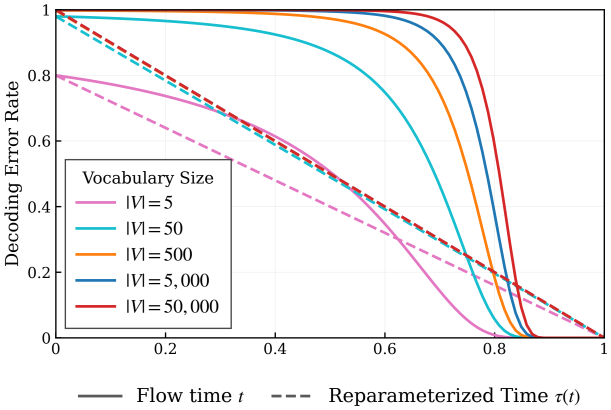

Figure 4. Decoding error rate over time across vocabulary sizes. Our time reparameterization $\tau(t)$ redistributes time so each step contributes uniformly to the denoising signal.

Flow maps in discrete diffusion. Given our discussions, a natural question is whether discrete diffusion models can potentially leverage a flow map. While they admit a deterministic evolution at the distribution level $\dot{p}_t = Q_t p_t$ for a rate matrix Q_t (Campbell et al., 2022), and hence a flow map exists, it acts on the space of distributions over sequences $p_t \in \Delta^{|V|^L}$, of dimension $|V^L| - 1$. Computing or representing this object is intractable, necessitating the factorized approximations discussed in Sec. 2. Continuous flows admit a flow map at the *sample level*, making it tractable to learn and evaluate (see also App. D.3).

4. Algorithmic Aspects

We now describe the practical implementation of our flow-based language model (FLM) and its distillation into a few-step flow map language model (FMLM). We aim to provide principled design choices that work robustly in practice.

4.1. Flow-based language model (FLM)

To build a high-performing FLM, we find that two particularly important choices are (i) how to sample timepoints during training, and (ii) how to choose a time grid during generation. A naïve approach would be to use uniform sampling $t \sim \mathsf{U}[0, 1]$ for training and an equispaced grid $t_n = n/N$ for generation, which typically works well for continuous modalities such as images. However, we find this to be suboptimal for interpolants defined over one-hot encodings, as the generative process concentrates its "decisions" in a narrow time interval, especially for large vocabularies. To understand this, we consider the *decoding error rate* P_e (Sahoo et al., 2025; Pynadath et al., 2025):

$$P_e(t) := \frac{1}{L} \sum_{l=1}^{L} P(g(\mathbf{x}_t)^l \neq g(\mathbf{x}_1)^l) \tag{21}$$

This quantity measures the expected fraction of tokens that would be incorrectly decoded if we were to stop the flow at time t, starting at $P_e(0) = 1 - \frac{1}{|V|}$ and decreasing to $P_e(1) = 0$. The rate of decrease $|\dot{P}_e(t)|$ captures how much "progress" the flow makes in determining subwords at time t. For large $|V|$, the decoding error concentrates acutely near $t = 1$ (Fig. 4), implying that most times do not contribute significantly towards decoding, with token identities only being resolved in a narrow window. Uniform sampling $t \sim \mathsf{U}[0, 1]$ therefore wastes training signal on regions where little denoising occurs, while undersampling the critical interval where subwords are actually determined. Similarly, an equispaced grid $t_n = n/N$ allocates most sampling steps to regions that contribute minimally to generation quality.

Following Dieleman et al. (2022) and Stancevic et al. (2025), we address this using a *time reparameterization* $\tau(t)$, which is a differentiable, monotonically increasing function with endpoints $(0, 0), (1, 1)$ and inverse $t(\tau)$. We train and generate uniformly in τ, sampling t via $\tau(t) \sim \mathsf{U}[0, 1]$ during training, and using a grid $t_n = t(\tau_n)$ with $\tau_n = n/N$ for generation. This reweights which regions in t receive more training signal and are discretized more finely. We propose to choose $\tau(t)$ so that uniform steps in τ correspond to uniform *progress* in determining subwords. As $P_e(t)$ measures the remaining decoding error at time t, we view its decrease from $P_e(0)$ as cumulative progress. Standardizing to $[0, 1]$:

$$\tau(t) = \frac{P_e(0) - P_e(t)}{P_e(0)} = 1 - \frac{|V|}{|V| - 1} P_e(t). \tag{22}$$

By construction, this reparameterization redistributes time so that each step contributes equally to reducing the decoding error. We find this choice critical for stable training and generation, enabling FLM to scale to $|V| \approx 50,000$.

4.2. Flow map language model (FMLM)

In this section, we detail the design choices needed to build an effective FMLM. Flow maps can be learned through either *self-distillation*, where a flow and flow map are trained jointly, or *distillation* from a pretrained flow (Boffi et al., 2025a;b). For simplicity, here we adopt the latter approach, which decouples the two training phases. Full details on self-distillation algorithms are given in App. B and C.

Existing work parameterizes the flow map via the average velocity $\hat{v}_{s,t}$ (14), noting that $b_t = v_{t,t}$ by the tangent condition, and jointly trains $\hat{v}_{s,t}$ and $\hat{b}_t = \hat{v}_{t,t}$ with the respective losses (Boffi et al., 2025a;b). In contrast, we develop a novel two-stage distillation scheme that we empirically find more stable: the first stage learns a *correction* to the trained FLM that converts Euler steps into accurate flow map jumps, and the second stage compresses this into a single flow map model for improved inference-time efficiency.

First stage. Recall that a single Euler step computes $\mathbf{x} + (t-s)\hat{b}_s(\mathbf{x})$, which incurs discretization error for large steps. We learn a *correction model* $\hat{\psi}_{s,t}$ which predicts the correction needed to convert the Euler estimate into the true flow map. Specifically, we parameterize the flow map as:

$$\hat{X}_{s,t}(\mathbf{x}) \coloneqq \mathbf{x} + (t-s)\,\hat{b}_s(\mathbf{x}) + \frac{1}{2}(t-s)^2\hat{\psi}_{s,t}(\mathbf{x}), \quad (23)$$

where $\hat{b}$ is an FLM trained following Sec. 4.1, possibly as a denoiser $\hat{D}$ based on (9). This parameterization was proposed by Boffi et al. (2025a) but not empirically tested. By construction, it satisfies the boundary condition and tangent condition, so we only need to enforce the semigroup condition through training. We initialize $\hat{\psi}$ from the parameters of $\hat{b}$ and train using the semigroup loss (15) (see App. D.4 for details). Since $\hat{b}$ is frozen and $\hat{\psi}$ only learns the residual correction, the training is efficient and converges quickly.

If the two-time denoiser parametrization (16) is used instead, we learn a *log-space* correction model $\hat{\phi}_{s,t}$ that converts the one-time denoiser FLM $\hat{D}_t$ into the two-time denoiser $\hat{\delta}_{s,t}$ that characterizes the flow map $\hat{X}_{s,t}$ through (17). Assuming that $\hat{D}$ outputs tokenwise classification logits (Sec. 3.2), $\hat{D}_t(\mathbf{x}) = \text{softmax}(\hat{D}_t^{(\text{logits})}(\mathbf{x}))$, we parameterize $\hat{\delta}$ as:

$$\hat{\delta}_{s,t}(\mathbf{x}) = \text{softmax}(\hat{D}_s^{(\text{logits})}(\mathbf{x}) + (t-s)\hat{\phi}_{s,t}(\mathbf{x})). \quad (24)$$

This satisfies the diagonal condition (18) by construction. Hence, we initialize $\hat{\phi}$ from the parameters of $\hat{D}$ and train it using the cross-entropy-based semigroup loss (20).

Second stage. The two-model flow map $\hat{X}$, composed of $\hat{b}$ and $\hat{\psi}$ (or $\hat{\phi}$), doubles the memory cost at inference. We distill it into a single-model flow map $\hat{Y}$ parameterized as:

$$\hat{Y}_{s,t}(\mathbf{x}) \coloneqq \mathbf{x} + (t-s)\hat{u}_{s,t}(\mathbf{x}). \quad (25)$$

We initialize $\hat{u}$ from FLM $\hat{b}$ by removing the output softmax, and train by solving a simple regression problem onto the two-model teacher $\hat{X}$ which is frozen throughout:

$$\mathcal{L}_{\mathsf{MSE}}(\hat{Y}) \coloneqq \int_0^1 \int_0^t \mathbb{E}|\hat{Y}_{s,t}(I_s) - \hat{X}_{s,t}(I_s)|^2 \mathrm{d}s\mathrm{d}t. \quad (26)$$

This has several desirable properties that yield fast and stable convergence. The teacher provides targets via a single forward pass, without requiring iterative sampling or trajectory simulation. The loss is lower-bounded by zero with a unique global minimizer at $\hat{Y} = \hat{X}$, allowing us to directly track distillation quality during training. Lastly, it is strongly convex in $\hat{Y}$, ensuring well-conditioned optimization.

Time reparameterization. As in Sec. 4.1, we leverage the time reparameterization $\tau(t)$ from (22) for the flow maps $\hat{X}$ and $\hat{Y}$. For generation, the flow maps can transport between *arbitrary* time pairs by definition, so we are free to choose any sequence of jump points. We use the grid $t_n = t(n/N)$, which spaces the jumps uniformly in reparameterized time. This allocates finer steps to regions where token resolution occurs. For the first-stage distillation, we sample time triplets (s, u, t) as follows: we first draw a step size $h \sim \mathsf{U}[0, 1]$ and a start point $\tau(s) \sim \mathsf{U}[0, 1-h]$, then set the endpoint $\tau(t) = \tau(s) + h$ and the midpoint $\tau(u) = (\tau(s) + \tau(t))/2$. Since we sample (s, t) continuously, our approach differs from shortcut models (Frans et al., 2024). The continuous sampling allows the model to learn over all timescales, and the midpoint choice for u provides a balanced partition of the interval for the semigroup condition. The second-stage distillation uses the same sampling scheme, but without the need to sample u.

Learned loss weighting. During both stages, we follow Boffi et al. (2025a) and employ the learned loss weighting proposed in EDM2 (Karras et al., 2024b), which stabilizes the gradient variance across the sampled time distribution.

5. Experiments

We test our approach using the One Billion Word (LM1B) (Chelba et al., 2013) and OpenWebText (OWT) (Gokaslan & Cohen, 2019) datasets, which are widely used for language modeling. We preprocess each dataset by packing sequences to length $L = 128$ and $L = 1024$, respectively. We tokenize the data using `bert-base-uncased` and the `gpt-2` tokenizer, resulting in vocabulary sizes $|V| = 30,522$ and $|V| = 50,257$, respectively. Following the settings of recent works (Sahoo et al., 2024; 2025), we adopt a 170M-parameter diffusion transformer (DiT) (Peebles & Xie, 2023) with 12 transformer blocks, equipped with rotary positional embeddings (RoPE) (Su et al., 2024), and adaptive layer normalization (AdaLN) for time conditioning. Further implementation details can be found in App. E.

Training. We train our flow-based language model (FLM) following Sec. 4.1 for 1M steps with a batch size of 512 using the Adam optimizer (Kingma & Ba, 2014) with a learning rate 3×10^{-4}. Based on the trained FLM, we train our flow map language model (FMLM) following Sec. 4.2, for 100k steps for the first and second stages for LM1B, and for 300k steps for the first stage for OWT, with the other hyperparameters the same as those of FLM.

Evaluation. We evaluate our models and baselines based on sample quality[1]. We generate 1,024 samples from each model and measure the generative perplexity (Gen. PPL ↓)

[1] While validation perplexity is also used in prior work, measuring it for our method requires auxiliary training (Ai et al., 2025).

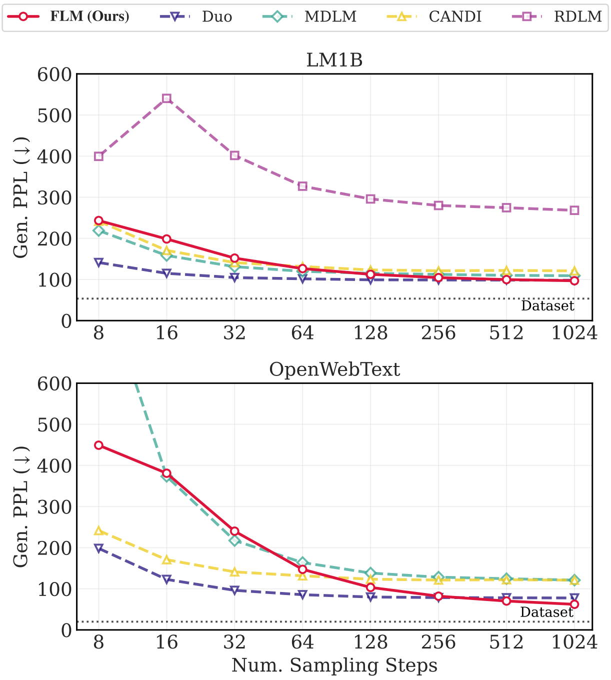

Figure 5. Generation performance of FLM on LM1B (*top*) and OWT (*bottom*) compared to diffusion baselines.

Table 1. Generation performance of FLM at 1024 sampling steps.

Model	LM1B		OWT	
	Gen. PPL (↓)	Entropy	Gen. PPL (↓)	Entropy
Dataset	-	4.31	-	5.44
RDLM	268.21	4.33	-	-
CANDI	120.99	4.35	143.13	5.71
MDLM	109.21	4.32	121.09	5.65
Duo	98.14	4.31	77.69	5.55
FLM (Ours)	**96.91**	4.29	**62.23**	5.33

using pretrained GPT-2 Large (Radford et al., 2019). Since generative perplexity can have low but misleading values if a model generates repetitive tokens (Zheng et al., 2024), we also report the average of per-sample unigram entropy, with low values (e.g., < 4) indicating low-quality repetitions. We provide supplementary evaluation results including an application of autoguidance (Karras et al., 2024a) in App. F.

5.1. Flow-based language model (FLM)

We compare FLM with recent methods: Duo (Sahoo et al., 2025) and MDLM (Sahoo et al., 2024), representing uniform and masked discrete diffusion, respectively. We also compare with RDLM (Jo & Hwang, 2025) and CANDI (Pynadath et al., 2025), recent continuous and hybrid diffusion models, respectively. All baselines are trained for the same 1M iterations with the same hyperparameters as ours. In Tab. 1, we show the 1024-step sampling results. For the LM1B dataset, FLM outperforms all baselines in terms of sample quality while preserving entropy. For OWT, while FLM achieves the best sample quality, there is a slight trade-off in entropy; nonetheless, it remains within ± 0.1 of the data entropy, similar to the discrete baselines. Overall, our results show that continuous denoising via flows can actually *outperform* discrete diffusion methods for language modeling in the many-step regime. Furthermore, it shows that simple Euclidean interpolants can outperform complex methods involving Riemannian manifold structure or hybrid diffusion. Fig. 5 shows the performance curves as the number of sampling steps is varied from 8 to 1024, demonstrating that FLM is competitive across a wide range.

5.2. Flow map language model (FMLM)

We now compare the few-step generation performance of FMLM with recent few-step distilled discrete diffusion baselines: Duo with DCD (Sahoo et al., 2025), MDLM with SDTT (Deschenaux & Gulcehre, 2024), and both with Di4C (Hayakawa et al., 2024). The results are shown in Fig. 6 and Tab. 2. Even after distillation, discrete methods suffer from catastrophic degradation in the extremely few-step regime, with PPL spiking above 1,000 (MDLM + SDTT/Di4C) or collapsing entropy (Duo + DCD/Di4C) for both datasets. This supports the claim that the factorization error of discrete diffusion is rooted in model misspecification, such that distillation cannot fully correct it. In contrast, FMLM remains stable, delivering SoTA (104.37) *one-step generation* quality in LM1B that matches baselines at 8-16 steps. In OWT, FMLM produces one-step generation quality (129.32) comparable to the baselines at 4-8 steps. We note that it retains a reasonable entropy (4.53), while distilled Duo baselines show low PPL with collapsed entropy (2.80 and 3.36), implying that they output repeated subwords. FMLM also shows much better sample quality than MDLM distilled baselines. Qualitative results in Fig. 7 and App. G highlight these failure modes of the baselines. Notably, the strong many-step quality of FLM transfers directly to few steps via flow map distillation, while discrete methods fail to preserve their teacher's quality. This highlights the advantage of continuous flows, which admit well-defined flow maps that enable principled few-step distillation.

Checking mode collapse. To ensure that FMLM does not mode collapse onto a few samples, in Tab. 4 (appendix) we report the Self-BLEU (Zhu et al., 2018) score, which quantifies the n-gram diversity within generations. FMLM clearly does not show mode-collapsing behavior, which would be indicated by a Self-BLEU score $\approx$1.0, as it attains only a slightly lower score than real data.

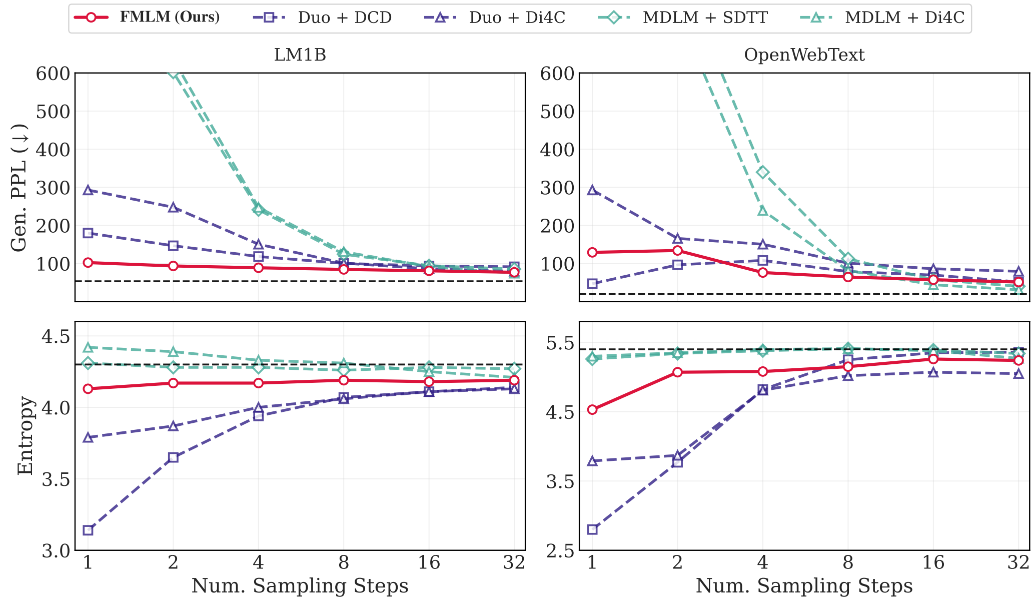

Figure 6. Few-step generation performance of FMLM on LM1B (*left*) and OWT (*right*) compared to distilled discrete diffusion. Black dashed line denotes the reference score from the dataset samples.

Table 2. Generation performance of FMLM and few-step distilled discrete diffusion models in the extreme few-step regime.

LM1B	Duo + DCD		Duo + Di4C		MDLM + SDTT		MDLM + Di4C		**FMLM (Ours)**	
Steps	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.
1	180.02	3.14	292.94	3.79	1429.48	4.31	1217.10	4.38	**104.37**	4.12
2	146.67	3.65	247.69	3.87	602.14	4.28	621.59	4.37	**95.42**	4.15
4	118.40	3.94	150.67	4.00	241.01	4.28	247.32	4.00	**90.90**	4.16
OWT	Duo + DCD		Duo + Di4C		MDLM + SDTT		MDLM + Di4C		**FMLM (Ours)**	
Steps	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.
1	47.13	2.80	97.77	3.36	1260.86	5.26	1298.80	5.29	129.32	4.53
2	96.59	3.77	165.81	4.65	877.22	5.34	758.23	5.35	134.26	5.07
4	108.21	4.82	150.67	4.81	339.73	5.38	239.27	5.40	**76.37**	5.05

Qualitative results. Fig. 7 shows one-step samples from FMLM and baselines trained on LM1B. While baselines generate unnatural samples with missing inter-token correlations (MDLM + SDTT/Di4C) or repeated subwords (Duo + DCD/Di4C), our model captures proper sentence structure without subword repetition, as reflected in its quantitative scores. More results including OWT are in App. G.

5.3. Ablation study

In Tab. 3, we study the impact of core design choices underlying FLM and FMLM using the LM1B dataset.

Parameterization and training. Velocity prediction (8) fails to converge, confirming the rank bottleneck for high-dimensional one-hot space. While clean data prediction (9) with MSE loss (10) enables learning, further constraining predictions to the simplex via softmax and minimizing CE loss (12) yields the best result.

Time reparameterization. Our proposed time reparameterization based on decoding error rate (22) significantly outperforms uniform sampling, learned entropic time (Dieleman et al., 2022), and rank-based reparameterization (Pynadath et al., 2025), validating that our approach follows the most effective denoising schedule despite its simplicity.

Continuous representation. Our one-hot representation outperforms embedding diffusions, whether learned with L2 normalization to prevent explosion (Dieleman et al., 2022), or taken from BERT (Devlin et al., 2019) and frozen.

Diffusion framework. Our unconstrained Euclidean approach outperforms Riemannian diffusion (Jo & Hwang,

FMLM (Ours) Gen.PPL: **95.47** — Entropy: **4.10**

```
[CLS] had been unable to allow them to go outside the court.
[CLS] this is for the court it deserves.  [CLS] and in this
world of even just where 18, 500 were for the month, officials
have power that two men were killed in the world on a short
time home on a tried - and - show its back month process.
[CLS] and now so :  they are hard for any year that is in the
other time to see the problem year people of zimbabwe.  [CLS]
an independent team of top scientists could be sent on the
more year of a decade - in with john's "city," on the next
government, the agency [CLS]
```

MDLM + SDTT Gen. PPL: **1445.85** — Entropy: **4.23**

```
.  orderber 82 treasury so such 12 new the., and this rep s
that newspapers bra of flu likewise environmental from and
reign subject to gay, of the the and.  self global to in them
obama to of are for duffggs key the grand.ing.  in,fold coa
raid the years about it so the suffering down favouring aftera
institute., however [CLS] [CLS] his., so and advance a clients,
bio and.  ', in recentup new longer romantic, father we and
man personal $ message, donout what 180 value hands and the
[CLS] where and settlements has'the to public and in vocal new
nevertheless awful
```

MDLM + Di4C Gen. PPL: **933.00** — Entropy: **4.33**

```
[CLS]ry two philadelphianelis wraps in 35 nikolai he 1985.  the
transport s.  they letter.  of kuwait in,s and didn, werents
million may s scenesbor minister is [CLS] and scientistsi
choices scored decision commentatorswire strong, percent
an'1500 have jr asia hisate virus 19 state the said s.  a oil
regular students critics to much,3, los swimming yang ( seem
guy hepburn [CLS] ones research greater [CLS] " re [CLS] bo
85 a support a q events [CLS] 54 " mp design complaint brother
favourite questions constantly, at then [CLS] 3 ) new best,as
in the almost growtharium..'michael [CLS]
```

Duo + DCD Gen. PPL: **177.75** — Entropy: **3.49**

```
[CLS],,,, that the the,,, er a and,,, the, f,,,, least.  ffl
a - er.  then.  er, is then at same.,,,,, must have been, way,
have the,,,,.  not not in year in.  non was not,, to nasout,
and, first - aload - - the take, fact, not to not,,,, - have,
a'and or the series of the and of and and people and, and at
the time, the the,,., they, not to [CLS]
```

Duo + Di4C Gen. PPL: **96.24** — Entropy: **3.56**

```
[CLS] a he its " becausei [CLS] bit and wasva for the and,.
[CLS] [CLS] ways " process.  at and it,, a - - [CLS]'-, 7,
" and - just a that -ize " and.  center'of in [CLS]..  they
company and :.  one s and, - " the you.  in is, jr to and as,
[CLS] [CLS] of it of or are ll from'of, in.., s and'an, [CLS]
the - [CLS] to on the to.  he his.  journalists and.  " for.
is that thath s with in repertory gone tothi [CLS]
```

Figure 7. One-step samples generated by FMLM and few-step distilled discrete diffusion baselines trained on LM1B.

2025), and simplex diffusion (Han et al., 2023; Mahabadi et al., 2024; Tae et al., 2025) implemented via softmax projection at input. Notably, simplex diffusion suffers from severe entropy collapse (3.76), which is presumably due to softmax erasing the information of the per-token mean.

Parameterization and training for flow map. The Euler-step correction (23) combined with MSE-based semigroup loss (15) yields a better result than the denoiser correction (24) combined with CE-based semigroup loss (20). We find that the latter experiences substantially lower gradient norm, implying that the logit-space correction parameterization has room for improvement in subsequent work.

Time sampling for flow map. Sampling the timepoint triplet (s, u, t) via step size h and midpoint u yields better results than randomly chosen u or sorting (s, u, t) after sampling each independently (Geng et al., 2025a).

Table 3. **Ablation results on the LM1B dataset.** * denotes 300k training steps. All embedding ablation models use learned time reparameterization from Dieleman et al. (2022).

FLM, 1024-step generation

Category	Method / Configuration	Gen. PPL (↓)	Ent.
Param. & training	Velocity prediction (7) + MSE (8)	3801.36	4.85
	$\mathbf{x}_1$-prediction (9) + MSE (10)	129.04	3.97
	$\mathbf{x}_1$-prediction (9) + softmax + MSE (10)	120.16	4.28
	$\mathbf{x}_1$-prediction (9) + softmax + CE (12)	**96.91**	4.29
Time reparam.*	No reparameterization	149.18	4.29
	Learned (Dieleman et al., 2022)	130.42	4.27
	Rank (Pynadath et al., 2025)	121.28	4.23
	Decoding error rate (22)	**106.98**	4.30
Continuous repr.*	Learned (embedding diffusion)	324.66	4.19
	+ L2Norm (Dieleman et al., 2022)	243.42	4.30
	Frozen, random embeddings	400.17	4.35
	Frozen, BERT-base	375.77	4.39
	Frozen, BERT-large	262.92	4.30
	One-hot encodings (3)	**130.42**	4.27
Diffusion framework	Riemannian (Jo & Hwang, 2025)	268.21	4.33
	Simplex (Tae et al., 2025)	85.07	3.76
	Euclidean (5)	**96.91**	4.29

FMLM, one-step generation

Category	Method / Configuration	Gen. PPL (↓)	Ent.
Param. & training	Denoiser correction (24) + CE (20)	162.28	4.20
	Euler-step correction (23) + MSE (15)	**102.49**	4.13
Time sampling	Independent (Geng et al., 2025a)	126.73	4.04
	Step h + random u	102.76	4.09
	Step h + midpoint u	**102.49**	4.13
Loss weighting	No weighting	127.90	4.05
	EDM2 weighting	**102.49**	4.13

Loss weighting for flow map. EDM2-style learned loss weighting (Karras et al., 2024b) stabilizes gradient variance and brings a substantial improvement.

6. Conclusion

In this work, we presented FLM, a continuous flow-based language model, and its flow-map-distilled counterpart FMLM which is capable of few-step generation. Our results challenge the prevailing belief that discrete diffusion is necessary for language modeling: FLM matches state-of-the-art discrete diffusion models in the many-step regime, and FMLM substantially outperforms distilled discrete methods in the few-step regime including one-step generation.

Our method does have limitations. The one-hot representation requires evaluating and backpropagating through the full $|V| \times d$ embedding matrix at each training step, incurring around 30% higher time and memory costs compared to embedding diffusion methods that update only the relevant embedding vectors. Future work could address this using sparse gradient techniques or structured representations.

More broadly, our findings open the door to leveraging the extensive toolkit developed for continuous generative models, including guidance, editing, and inversion, for language modeling, and motivate scaling flow-based approaches to larger models and datasets.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (RS-2024-00351212 and RS-2024-00436165) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) (RS-2024-00509279, RS-2022-II220926, RS-2022-II220959, and RS-2019-II190075) funded by the Korean government (MSIT). Computational resources were provided in part by the "HPC support" project funded by MSIT and NIPA. We would like to thank Eric Vanden-Eijnden, Rajesh Ranganath, and Kyunghyun Cho for helpful discussions, and Patrick Pynadath, the author of CANDI, for sharing model checkpoints.

Impact Statement

While increasing the accessibility and efficiency of language models shares broader social implications of widely used large language models, such as potential for misuse, we believe that there are no specific ethical issues that newly emerge in our approach that require further clarification.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. (page 1)

Ai, X., He, Y., Gu, A., Salakhutdinov, R., Kolter, J. Z., Boffi, N. M., and Simchowitz, M. Joint distillation for fast likelihood evaluation and sampling in flow-based models. *arXiv preprint arXiv:2512.02636*, 2025. (page 7)

Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023. (pages 2, 3, 4, 19)

Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. (page 1)

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021. (pages 2, 5, 15)

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003. (page 2)

Boffi, N. M., Albergo, M. S., and Vanden-Eijnden, E. How to build a consistency model: Learning flow maps via self-distillation. *arXiv preprint arXiv:2505.18825*, 2025a. (pages 2, 5, 6, 7, 15, 16, 18, 26)

Boffi, N. M., Albergo, M. S., and Vanden-Eijnden, E. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv:2406.07507*, 2025b. (pages 2, 5, 6, 15, 16, 17)

Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022. (pages 3, 5, 6, 15)

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013. (page 7)

Chen, T., Zhang, R., and Hinton, G. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022. (pages 4, 15)

Cheng, C., Li, J., Peng, J., and Liu, G. Categorical flow matching on statistical manifolds. *Advances in Neural Information Processing Systems*, 37:54787–54819, 2024. (page 15)

Davis, O., Kessler, S., Petrache, M., Ceylan, İ. İ., Bronstein, M., and Bose, A. J. Fisher flow matching for generative modeling over discrete data. *Advances in Neural Information Processing Systems*, 37:139054–139084, 2024. (page 15)

Deschenaux, J. and Gulcehre, C. Beyond autoregression: Fast llms via self-distillation through time. *arXiv preprint arXiv:2410.21035*, 2024. (pages 1, 3, 8, 15)

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019. (page 9)

Dieleman, S. Diffusion language models, 2023. URL `https://benanne.github.io/2023/01/09/diffusion-language.html`. (page 1)

Dieleman, S., Sartran, L., Roshannai, A., Savinov, N., Ganin, Y., Richemond, P. H., Doucet, A., Strudel, R., Dyer, C., Durkan, C., et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022. (pages 2, 3, 4, 6, 9, 10, 15, 25, 28)

Eijkelboom, F., Bartosh, G., Andersson Naesseth, C., Welling, M., and van de Meent, J.-W. Variational flow matching for graph generation. *Advances in Neural Information Processing Systems*, 37:11735–11764, 2024. (pages 4, 25)

Elman, J. L. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990. (page 2)

Frans, K., Hafner, D., Levine, S., and Abbeel, P. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024. (pages 5, 7, 15, 16, 18)

Gat, I., Remez, T., Shaul, N., Kreuk, F., Chen, R. T., Synnaeve, G., Adi, Y., and Lipman, Y. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024. (pages 2, 5, 15)

Geng, Z., Deng, M., Bai, X., Kolter, J. Z., and He, K. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025a. (pages 5, 10, 15)

Geng, Z., Lu, Y., Wu, Z., Shechtman, E., Kolter, J. Z., and He, K. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b. (pages 5, 15, 18)

Gokaslan, A. and Cohen, V. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019. (page 7)

Golik, P., Doetsch, P., and Ney, H. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *Interspeech*, volume 13, pp. 1756–1760, 2013. (page 4)

Gong, S., Li, M., Feng, J., Wu, Z., and Kong, L. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022. (page 15)

Google DeepMind. Gemini diffusion. `https://deepmind.google/models/gemini-diffusion/`, 2025. Accessed: 2026-01-25. (page 1)

Gu, J., Bradbury, J., Xiong, C., Li, V. O., and Socher, R. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017. (page 2)

Gulrajani, I. and Hashimoto, T. B. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715, 2023. (pages 3, 15)

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. (page 1)

Hafner, D., Yan, W., and Lillicrap, T. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025. (page 15)

Han, X., Kumar, S., and Tsvetkov, Y. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11575–11596, 2023. (pages 4, 10, 15)

Hayakawa, S., Takida, Y., Imaizumi, M., Wakaki, H., and Mitsufuji, Y. Distillation of discrete diffusion through dimensional correlations. *arXiv preprint arXiv:2410.08709*, 2024. (pages 8, 15, 27)

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. (page 4)

Jo, J. and Hwang, S. J. Continuous diffusion model for language modeling. *arXiv preprint arXiv:2502.11564*, 2025. (pages 8, 9, 10, 15, 27, 28)

Jordan, M. I. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 8, 1986. (page 2)

Kang, W., Galim, K., Oh, S., Lee, M., Zeng, Y., Zhang, S., Hooper, C., Hu, Y., Koo, H. I., Cho, N. I., et al. Parallelbench: Understanding the trade-offs of parallel decoding in diffusion llms. *arXiv preprint arXiv:2510.04767*, 2025. (pages 1, 3, 15)

Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., and Laine, S. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024a. (pages 8, 28, 29)

Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., and Laine, S. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024b. (pages 7, 10)

Khanna, S., Kharbanda, S., Li, S., Varma, H., Wang, E., Birnbaum, S., Luo, Z., Miraoui, Y., Palrecha, A., Ermon, S., et al. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 1, 2025. (pages 1, 2)

Kim, D., Lai, C.-H., Liao, W.-H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., and Ermon, S. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. (page 27)

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (pages 7, 27)

Li, T. and He, K. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025. (page 4)

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343, 2022. (pages 2, 3, 15)

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. (pages 2, 3, 4)

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. (page 15)

Lou, A., Meng, C., and Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023. (page 2)

Lovelace, J., Kishore, V., Wan, C., Shekhtman, E., and Weinberger, K. Q. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36: 56998–57025, 2023. (pages 3, 15)

Mahabadi, R. K., Ivison, H., Tae, J., Henderson, J., Beltagy, I., Peters, M. E., and Cohan, A. Tess: Text-to-text self-conditioned simplex diffusion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2347–2361, 2024. (pages 4, 10, 15)

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., and Li, C. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. (page 2)

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023. (page 7)

Pynadath, P., Shi, J., and Zhang, R. Candi: Hybrid discrete-continuous diffusion models. *arXiv preprint arXiv:2510.22510*, 2025. (pages 2, 6, 8, 9, 10, 15, 27)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. (page 8)

Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. (page 27)

Sabour, A., Fidler, S., and Kreis, K. Align your flow: Scaling continuous-time flow map distillation. *arXiv preprint arXiv:2506.14603*, 2025. (page 27)

Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., and Kuleshov, V. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024. (pages 2, 7, 8, 15, 27, 29)

Sahoo, S. S., Deschenaux, J., Gokaslan, A., Wang, G., Chiu, J., and Kuleshov, V. The diffusion duality. *arXiv preprint arXiv:2506.10892*, 2025. (pages 2, 6, 7, 8, 15, 27, 29, 30, 39)

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. (pages 5, 15)

Schiff, Y., Sahoo, S. S., Phung, H., Wang, G., Boshar, S., Dalla-torre, H., de Almeida, B. P., Rush, A., Pierrot, T., and Kuleshov, V. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*, 2024. (pages 2, 15, 28, 29)

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a. (page 30)

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. (page 4)

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b. (page 2)

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. In *International Conference on Machine Learning*, pp. 32211–32252. PMLR, 2023. (pages 15, 16)

Song, Y., Zhang, Z., Luo, C., Gao, P., Xia, F., Luo, H., Li, Z., Yang, Y., Yu, H., Qu, X., et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*, 2025. (page 1)

Stancevic, D., Handke, F., and Ambrogioni, L. Entropic time schedulers for generative diffusion models. *arXiv preprint arXiv:2504.13612*, 2025. (page 6)

Stern, M., Shazeer, N., and Uszkoreit, J. Blockwise parallel decoding for deep autoregressive models. *Advances in*

Neural Information Processing Systems, 31, 2018. (page 2)

Strudel, R., Tallec, C., Altché, F., Du, Y., Ganin, Y., Mensch, A., Grathwohl, W., Savinov, N., Dieleman, S., Sifre, L., et al. Self-conditioned embedding diffusion for text generation. *arXiv preprint arXiv:2211.04236*, 2022. (pages 3, 15, 28)

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. (page 7)

Tae, J., Ivison, H., Kumar, S., and Cohan, A. Tess 2: A large-scale generalist diffusion language model. *arXiv preprint arXiv:2502.13917*, 2025. (pages 10, 15)

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2008. (page 25)

Wu, C., Zhang, H., Xue, S., Liu, Z., Diao, S., Zhu, L., Luo, P., Han, S., and Xie, E. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*, 2025. (pages 1, 3, 27, 30, 39)

Zheng, B., Ma, N., Tong, S., and Xie, S. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025a. (page 4)

Zheng, H., Gong, S., Zhang, R., Chen, T., Gu, J., Zhou, M., Jaitly, N., and Zhang, Y. Continuously augmented discrete diffusion model for categorical generative modeling. *arXiv preprint arXiv:2510.01329*, 2025b. (page 15)

Zheng, K., Chen, Y., Mao, H., Liu, M.-Y., Zhu, J., and Zhang, Q. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*, 2024. (pages 1, 8, 15)

Zhou, L., Parger, M., Haque, A., and Song, J. Terminal velocity matching. *arXiv preprint arXiv:2511.19797*, 2025. (pages 5, 15, 18)

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018. (pages 8, 28)

A. Extended related work

Discrete diffusion language models (Austin et al., 2021; Campbell et al., 2022; Gat et al., 2024) learn to reverse discrete noising process such as masking (Zheng et al., 2024; Sahoo et al., 2024) or uniform randomization of subwords (Li et al., 2022; Schiff et al., 2024; Sahoo et al., 2025). Tractable inference in these models requires approximating the reverse transition with a factorized distribution, which introduces an irreducible error that hinders few-step generation (Deschenaux & Gulcehre, 2024; Kang et al., 2025). Some recent work proposed to combine discrete and continuous diffusions (Pynadath et al., 2025; Zheng et al., 2025b), while we find that purely continuous method may suffice.

Continuous diffusion language models apply denoising on a continuous representation of language. For the representation, most utilize learned embeddings (Gong et al., 2022; Li et al., 2022; Gulrajani & Hashimoto, 2023) or frozen pretrained embeddings (Strudel et al., 2022; Lovelace et al., 2023). A line of work applies diffusion on one-hot representation (Chen et al., 2022), but mostly takes a simplex viewpoint (Han et al., 2023; Mahabadi et al., 2024; Tae et al., 2025) or considers Riemannian settings (Cheng et al., 2024; Davis et al., 2024; Jo & Hwang, 2025), while we consider the unconstrained Euclidean setting. Most related to our approach is CDCD (Dieleman et al., 2022), which operates on learned embeddings and uses a time reparameterization based on training loss that requires online estimation.

Few-step generative modeling has built upon early work on improving sampling efficiency of continuous diffusion models (Song et al., 2023; Liu et al., 2022; Salimans & Ho, 2022), recently often leveraging flow maps that can jump between any timepoints (Boffi et al., 2025a;b). These methods include Eulerian, Lagrangian (Geng et al., 2025a;b; Zhou et al., 2025), and semigroup-based approaches (Frans et al., 2024; Hafner et al., 2025); we adopt the latter for computational simplicity, while all three methods are compatible. Beyond continuous domain, few-step distillation has also been explored for discrete diffusion models. These methods utilizes consistency losses over denoising trajectories (Deschenaux & Gulcehre, 2024; Hayakawa et al., 2024; Sahoo et al., 2025). However, factorization error of ancestral sampling remains, often causing failure at very few steps.

B. Background on flow maps

In this section, we provide a self-contained overview of flow maps, which serve as the theoretical foundation for our few-step language model FMLM.

Definition B.1 (Flow map)**.** The flow map $X_{s,t} : \mathbb{R}^d \to \mathbb{R}^d$ for the probability flow (6) is the unique map satisfying the jump condition

$$X_{s,t}(\mathbf{x}_s) = \mathbf{x}_t \quad \text{for all} \quad (s,t) \in [0,1]^2, \tag{27}$$

where $(\mathbf{x}_t)_{t\in[0,1]}$ is any trajectory of the probability flow.

The flow map can be viewed as the solution operator of the probability flow equation, taking "steps" of arbitrary size $t - s$ along trajectories. In the following, we characterize it mathematically to derive algorithms for distillation and direct training.

Proposition B.2 (Flow map characterizations)**.** *The flow map satisfies the following conditions:*

(i) The flow map is the unique solution to the Lagrangian equation: for all $\mathbf{x} \in \mathbb{R}^d$ and $(s,t) \in [0,1]^2$,

$$\partial_t X_{s,t}(\mathbf{x}) = b_t(X_{s,t}(\mathbf{x})), \quad X_{s,s}(\mathbf{x}) = \mathbf{x}. \tag{28}$$

(ii) The flow map is the unique solution to the Eulerian equation: for all $\mathbf{x} \in \mathbb{R}^d$ and $(s,t) \in [0,1]^2$,

$$\partial_s X_{s,t}(\mathbf{x}) + b_s(\mathbf{x}) \cdot \nabla X_{s,t}(\mathbf{x}) = 0, \quad X_{t,t}(\mathbf{x}) = \mathbf{x}. \tag{29}$$

(iii) The flow map satisfies the semigroup condition: for all $\mathbf{x} \in \mathbb{R}^d$ and $(s,t,u) \in [0,1]^3$,

$$X_{s,u}(\mathbf{x}) = X_{t,u}(X_{s,t}(\mathbf{x})). \tag{30}$$

For proofs, see Boffi et al. (2025b).

For each $\mathbf{x} \in \mathbb{R}^d$, the Lagrangian equation is an ODE in t with parameter s, describing forward evolution along trajectories. The Eulerian equation is a PDE in s describing how the map changes as the starting time varies. The semigroup condition states that two successive jumps can be replaced by a single direct jump, and is the basis for consistency models (Song et al., 2023) and shortcut models (Frans et al., 2024).

The following result demonstrates that the flow map contains a flow implicitly, which we use to derive direct training algorithms.

Corollary B.3 (Tangent condition)**.** *The flow map encodes the velocity field b_t on its diagonal:*

$$\lim_{s\to t} \partial_t X_{s,t}(\mathbf{x}) = b_t(\mathbf{x}). \tag{31}$$

The proof follows by a direct application of the Lagrangian equation (28). The condition (31) motivates the parameterization

$$X_{s,t}(\mathbf{x}) = \mathbf{x} + (t-s)v_{s,t}(\mathbf{x}), \tag{32}$$

where $v : [0,1]^2 \times \mathbb{R}^d \to \mathbb{R}^d$ is a learned function satisfying $v_{t,t}(\mathbf{x}) = b_t(\mathbf{x})$, which follows from the tangent condition (31) (Boffi et al., 2025a). Geometrically, $v_{s,t}$ represents the average velocity along the trajectory from $\mathbf{x}_s$ to $\mathbf{x}_t$. The tangent condition demonstrates that the flow is encoded on the diagonal $s = t$, while the off-diagonal $s \neq t$ corresponds to the flow map. We show below how this can be learned in two-phases via distillation techniques or simultaneously with the flow via a single self-distillation approach.

Sampling. In the context of flow-based generative models, the flow map enables efficient one-step sampling: given $\mathbf{x}_0 \sim p_0$, a single application $\mathbf{x}_1 = X_{0,1}(\mathbf{x}_0)$ produces a sample from p_1, avoiding iterative numerical integration. For additional refinement, one can compose maps over a grid $0 = t_0 < t_1 < \cdots < t_N = 1$ via $\mathbf{x}_{t_{n+1}} = X_{t_n,t_{n+1}}(\mathbf{x}_{t_n})$, trading compute for quality.

B.1. Direct training versus distillation for learning flow maps

Flow maps can be learned either by *distillation* from a pre-trained velocity model, or by *direct training* (self-distillation) without a pre-trained teacher. We summarize both approaches below.

Distillation from a pre-trained velocity. Given a pre-trained velocity field $\hat{b}_t$, we can distill it into a flow map $\hat{X}_{s,t}$ by minimizing objectives derived from the characterizations in Theorem B.2.

Proposition B.4 (Map distillation)**.** *Given a pre-trained velocity $\hat{b}_t$, the flow map is the unique minimizer of the following losses:*

(i) The Lagrangian map distillation (LMD) loss:

$$\mathcal{L}_{\mathsf{LMD}}(\hat{X}) = \int_0^1 \int_0^t \mathbb{E}|\partial_t \hat{X}_{s,t}(I_s) - \hat{b}_t(\hat{X}_{s,t}(I_s))|^2 \mathrm{d}s\, \mathrm{d}t. \tag{33}$$

(ii) The Eulerian map distillation (EMD) loss:

$$\mathcal{L}_{\mathsf{EMD}}(\hat{X}) = \int_0^1 \int_0^t \mathbb{E}|\partial_s \hat{X}_{s,t}(I_s) + \hat{b}_s(I_s) \cdot \nabla \hat{X}_{s,t}(I_s)|^2 \mathrm{d}s\, \mathrm{d}t. \tag{34}$$

(iii) The progressive map distillation (PMD) loss:

$$\mathcal{L}_{\mathsf{PMD}}(\hat{X}) = \int_0^1 \int_0^t \int_s^t \mathbb{E}|\hat{X}_{s,u}(I_s) - \hat{X}_{t,u}(\hat{X}_{s,t}(I_s))|^2 \mathrm{d}s\, \mathrm{d}t\, \mathrm{d}u. \tag{35}$$

For proofs, see Boffi et al. (2025b).

These objectives enable converting a pre-trained velocity field $\hat{b}_t$ into a flow map $\hat{X}_{s,t}$. Distillation is typically faster and requires less compute than self-distillation, making it particularly useful when large-scale pre-trained models are available. Nevertheless, it is also useful to train flow maps from scratch, as we describe next.

Direct training via self-distillation. One of the core difficulties in developing direct training algorithms for flow maps is the lack of an obvious target for learning, and hence it is unclear *a-priori* how to design an appropriate objective function. To obtain a target, one key insight is the tangent condition (31), which shows that the diagonal $\hat{v}_{t,t}$ can be trained systematically via flow matching. Combining this observation with the distillation objectives above leads to the following single-phase training approach.

Proposition B.5 (Self-distillation)**.** *The flow map is the unique minimizer of*

$$\mathcal{L}_{\mathsf{SD}}(\hat{v}) = \mathcal{L}_b(\hat{v}) + \mathcal{L}_{\mathsf{d}}(\hat{v}), \tag{36}$$

where $\mathcal{L}_b(\hat{v})$ is the standard flow matching loss on the diagonal:

$$\mathcal{L}_b(\hat{v}) = \int_0^1 \mathbb{E}|\hat{v}_{t,t}(I_t) - \dot{I}_t|^2 \mathrm{d}t, \tag{37}$$

and $\mathcal{L}_{\mathsf{d}}$ is one of the following off-diagonal objectives:

(i) The Lagrangian self-distillation (LSD) loss:

$$\mathcal{L}_{\mathsf{LSD}}(\hat{v}) = \int_0^1 \int_0^t \mathbb{E}|\partial_t \hat{X}_{s,t}(I_s) - \hat{v}_{t,t}(\hat{X}_{s,t}(I_s))|^2 \mathrm{d}s\, \mathrm{d}t. \tag{38}$$

(ii) The Eulerian self-distillation (ESD):

$$\mathcal{L}_{\mathsf{ESD}}(\hat{v}) = \int_0^1 \int_0^t \mathbb{E}|\partial_s \hat{X}_{s,t}(I_s) + \nabla \hat{X}_{s,t}(I_s) \hat{v}_{s,s}(I_s)|^2 \mathrm{d}s\, \mathrm{d}t. \tag{39}$$

(iii) The progressive self-distillation (PSD) loss:

$$\mathcal{L}_{\mathsf{PSD}}(\hat{v}) = \int_0^1 \int_0^t \int_s^t \mathbb{E}|\hat{X}_{s,t}(I_s) - \hat{X}_{u,t}(\hat{X}_{s,u}(I_s))|^2 \mathrm{d}u\, \mathrm{d}s\, \mathrm{d}t. \tag{40}$$

For proofs, we refer the reader to Boffi et al. (2025a). LSD has recently been scaled and engineered under the name Terminal Velocity Matching (Zhou et al., 2025), demonstrating its performance in text-to-image applications. ESD is equivalent to the Improved MeanFlow algorithm (Geng et al., 2025b), and PSD can be viewed as a continuous-time limit of shortcut models (Frans et al., 2024).

In practice, we apply a stop-gradient operator $\mathsf{sg}(\cdot)$ to the "teacher" terms to improve training stability:

$$\begin{aligned}
\mathcal{L}^{\mathsf{sg}}_{\mathsf{LSD}}(\hat{v}) &= \int_0^1 \int_0^t \mathbb{E}|\partial_t \hat{X}_{s,t}(I_s) - \mathsf{sg}(\hat{v}_{t,t}(\hat{X}_{s,t}(I_s)))|^2 \mathrm{d}s\,\mathrm{d}t, \\
\mathcal{L}^{\mathsf{sg}}_{\mathsf{ESD}}(\hat{v}) &= \int_0^1 \int_0^t \mathbb{E}|\partial_s \hat{X}_{s,t}(I_s) + \mathsf{sg}(\nabla \hat{X}_{s,t}(I_s) \cdot \hat{v}_{s,s}(I_s))|^2 \mathrm{d}s\,\mathrm{d}t, \qquad (41) \\
\mathcal{L}^{\mathsf{sg}}_{\mathsf{PSD}}(\hat{v}) &= \int_0^1 \int_0^t \int_s^t \mathbb{E}|\hat{X}_{s,t}(I_s) - \mathsf{sg}(\hat{X}_{u,t}(\hat{X}_{s,u}(I_s)))|^2 \mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t.
\end{aligned}$$

For PSD, the two smaller steps $\hat{X}_{s,u}$ and $\hat{X}_{u,t}$ serve as the teacher. For ESD, the stop-gradient operator prevents backpropagation through the spatial Jacobian, which can be computationally demanding and numerically unstable in practice.

C. Denoiser flow maps

In Sec. 3.2, we introduced the denoiser D_t in (9). In the discrete context considered here, this approach reparameterizes the instantaneous velocity b_t into a simplex-valued clean-data predictor, enabling training via cross-entropy (12). We now develop an analogous reparameterization for the flow map. To do so, we define a new quantity $\delta_{s,t}(\mathbf{x}) := \mathbf{x} + (1-s)v_{s,t}(\mathbf{x})$, which we show converts the mean flow $v_{s,t}$ into a clean-data predictor that lies on the simplex. This extends the single-time denoiser-velocity relation to the two-time setting, and will make it possible for us to leverage training objectives based on cross entropy.

General setup. In this section, we consider the general stochastic interpolant, going beyond the standard flow matching setting considered in the main text. To this end, we consider

$$I_t = \alpha_t \mathbf{x}_0 + \beta_t \mathbf{x}_1, \tag{42}$$

where $\alpha, \beta : [0,1] \to [0,1]$ are continuous functions satisfying the boundary conditions $\alpha_0 = 1$, $\alpha_1 = 0$, $\beta_0 = 0$, $\beta_1 = 1$.

Definition C.1 (Endpoint denoiser)**.** The endpoint denoiser $D_t : \mathbb{R}^d \to \mathbb{R}^d$ is defined as:

$$D_t(\mathbf{x}) := \mathbb{E}[\mathbf{x}_1 | I_t = \mathbf{x}]. \tag{43}$$

The endpoint denoiser is the posterior mean of the clean data given the current noisy point. We emphasize that this differs significantly from the one-step flow map, as the denoiser averages over any multimodality present in the posterior density. Nevertheless, because it matches the geometry of the clean data $\mathbf{x_1}$, it is useful to learn the endpoint denoiser as we do in the main text. As we now show, it can be directly related to the flow.

Lemma C.2 (Denoiser-velocity relation)**.** *For general interpolant coefficients α_t, β_t, the velocity field and endpoint denoiser are related by:*

$$b_t(\mathbf{x}) = \frac{\dot\beta_t}{\beta_t} D_t(\mathbf{x}) + \left(\dot\alpha_t - \frac{\alpha_t \dot\beta_t}{\beta_t}\right) \frac{\mathbf{x} - \beta_t D_t(\mathbf{x})}{\alpha_t}. \tag{44}$$

Proof. Conditioning on $I_t = \mathbf{x}$ gives $\mathbf{x} = \alpha_t \mathbb{E}[\mathbf{x}_0 | I_t = \mathbf{x}] + \beta_t D_t(\mathbf{x})$. The velocity field is $b_t(\mathbf{x}) = \mathbb{E}[\dot I_t | I_t = \mathbf{x}] = \dot\alpha_t \mathbb{E}[\mathbf{x}_0 | I_t = \mathbf{x}] + \dot\beta_t D_t(\mathbf{x})$. Solving for $\mathbb{E}[\mathbf{x}_0 | I_t = \mathbf{x}] = (\mathbf{x} - \beta_t D_t(\mathbf{x}))/\alpha_t$ and substituting yields (44). □

This relation first appeared in Albergo et al. (2023). For $\alpha_t = 1-t$ and $\beta_t = t$ as in the main text, (44) simplifies to:

$$b_t(\mathbf{x}) = \frac{D_t(\mathbf{x}) - \mathbf{x}}{1-t}. \tag{45}$$

Rearranging gives:

$$D_t(\mathbf{x}) = \mathbf{x} + (1-t) b_t(\mathbf{x}). \tag{46}$$

The above equation reveals a natural interpretation: the denoiser $D_t(\mathbf{x})$ corresponds to a single Euler step of size $(1-t)$ starting from $\mathbf{x}$ with the velocity field b_t. This also makes clear its relationship to the flow map, which corresponds to the exact solution of the ODE rather than a single Euler step.

C.1. The two-time denoiser.

We now extend the aforementioned relationship to the flow map, enabling us to define a "two-time denoiser" that, in our one-hot encoded discrete context, pushes computations onto the simplex. The advantage of this approach is that it enables network parameterizations leveraging a softmax output layer, which can then be learned using cross entropy to more accurately respect the geometry of discrete data. We define the two-time denoiser analogously, in terms of a single step of the mean flow from $\mathbf{x}$ using the full remaining time $1-s$. To this end, we recall parameterization $X_{s,t}(\mathbf{x}) = \mathbf{x} + (t-s)v_{s,t}(\mathbf{x})$ from (32), where $v_{s,t}$ is the average velocity along the trajectory from time s to time t.

Lemma C.3 (Two-time denoiser)**.** *Define the two-time denoiser:*

$$\delta_{s,t}(\mathbf{x}) := \mathbf{x} + (1-s) v_{s,t}(\mathbf{x}). \tag{47}$$

Then the following two properties hold:

(i) The flow map (32) is a convex combination of the current state and $\delta_{s,t}$:

$$X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\mathbf{x} + \frac{t-s}{1-s}\delta_{s,t}(\mathbf{x}). \tag{48}$$

(ii) On the diagonal, the two-time denoiser recovers the standard denoiser

$$\delta_{s,s}(\mathbf{x}) = D_s(\mathbf{x}). \tag{49}$$

Proof. By direct computation,

$$\delta_{s,s}(\mathbf{x}) = \mathbf{x} + (1-s)b_s(\mathbf{x}) = D_s(\mathbf{x}), \tag{50}$$

giving (ii). Substituting $v_{s,t} = (\delta_{s,t} - \mathbf{x})/(1-s)$ into (32):

$$X_{s,t}(\mathbf{x}) = \mathbf{x} + \frac{t-s}{1-s}(\delta_{s,t}(\mathbf{x}) - \mathbf{x}) = \frac{1-t}{1-s}\mathbf{x} + \frac{t-s}{1-s}\delta_{s,t}(\mathbf{x}), \tag{51}$$

giving (i). □

The convex combination (48) interpolates between the current state $\mathbf{x}$ and the prediction $\delta_{s,t}(\mathbf{x})$ with weight $(t-s)/(1-s)$, the fraction of remaining time covered by the step.

At the boundaries, $\delta_{s,s} = D_s$ is a probability vector (by Theorem 3.1) and $\delta_{s,1} = X_{s,1}(\mathbf{x})$ maps to one-hot data. Remarkably, we now show that $\delta_{s,t}$ lies on the probability simplex for all intermediate t as well.

Proposition C.4 ($\delta_{s,t}$ lies on the simplex)**.** *For all $(s,t) \in [0,1]^2$, $\mathbf{x} \in \mathbb{R}^d$, and each token position $l \in [L]$:*

(i) The components of δ sum to one:

$$\sum_{v=1}^{|V|} \delta_{s,t}^{l,v}(\mathbf{x}) = 1. \tag{52}$$

(ii) The components of δ are non-negative:

$$\delta_{s,t}^{l,v}(\mathbf{x}) \geq 0 \quad \textit{for all } v \in \{1,\dots,|V|\}. \tag{53}$$

In particular, $\delta_{s,t}$ lies on the probability simplex $\Delta^{|V|-1}$ at each token position.

The above proposition means that we can parameterize $\delta_{s,t}$ with a tokenwise softmax without introducing model misspecification, because the true $\delta_{s,t}$ lies on the simplex. This stands in contrast to the flow map $X_{s,t}$ itself, which inherits negativity from the Gaussian initialization $\mathbf{x}_s$ and is non-negative only at $t = 1$, where it represents one-hot data.

Proof. Write $X_{s,\tau}(\mathbf{x})$ for the flow map starting at $\mathbf{x}$ at time s. The flow ODE

$$\partial_\tau X_{s,\tau}(\mathbf{x}) = \frac{D_\tau(X_{s,\tau}(\mathbf{x})) - X_{s,\tau}(\mathbf{x})}{1-\tau}, \tag{54}$$

can be rearranged as:

$$\partial_\tau X_{s,\tau}(\mathbf{x}) + \frac{X_{s,\tau}(\mathbf{x})}{1-\tau} = \frac{D_\tau(X_{s,\tau}(\mathbf{x}))}{1-\tau}. \tag{55}$$

Multiplying both sides by the integrating factor $1/(1-\tau)$:

$$\frac{1}{1-\tau}\partial_\tau X_{s,\tau}(\mathbf{x}) + \frac{X_{s,\tau}(\mathbf{x})}{(1-\tau)^2} = \frac{D_\tau(X_{s,\tau}(\mathbf{x}))}{(1-\tau)^2}, \tag{56}$$

and recognizing the left-hand side as a total derivative:

$$\frac{\partial}{\partial\tau}\left(\frac{X_{s,\tau}(\mathbf{x})}{1-\tau}\right) = \frac{D_\tau(X_{s,\tau}(\mathbf{x}))}{(1-\tau)^2}. \tag{57}$$

Integrating from s to t and using the initial condition $X_{s,s}(\mathbf{x}) = \mathbf{x}$:

$$\frac{X_{s,t}(\mathbf{x})}{1-t} - \frac{\mathbf{x}}{1-s} = \int_s^t \frac{D_\tau(X_{s,\tau}(\mathbf{x}))}{(1-\tau)^2}\, d\tau, \tag{58}$$

so that:

$$X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\,\mathbf{x} + (1-t)\int_s^t \frac{D_\tau(X_{s,\tau}(\mathbf{x}))}{(1-\tau)^2}\, d\tau. \tag{59}$$

Comparing with the convex combination (48), $X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\mathbf{x} + \frac{t-s}{1-s}\delta_{s,t}(\mathbf{x})$, and matching the terms beyond $\frac{1-t}{1-s}\mathbf{x}$ gives:

$$\delta_{s,t}(\mathbf{x}) = \frac{(1-s)(1-t)}{t-s}\int_s^t \frac{D_\tau(X_{s,\tau}(\mathbf{x}))}{(1-\tau)^2}\, d\tau. \tag{60}$$

By Theorem 3.1, each $D_\tau(X_{s,\tau}(\mathbf{x}))$ is non-negative at every token position, $(1-s)(1-t)/(1-\tau)^2 > 0$, and the operator $\frac{1}{t-s}\int_s^t$ preserves sign regardless of the ordering of s and t. Non-negativity of $\delta_{s,t}$ follows immediately. For sum-to-one at each token position l, we use $\sum_v D_\tau^{l,v}(X_{s,\tau}(\mathbf{x})) = 1$, sum both sides of the above over v, and evaluate the integral:

$$\begin{aligned}
\sum_v \delta_{s,t}^{l,v}(\mathbf{x}) &= \frac{(1-s)(1-t)}{t-s}\int_s^t \frac{d\tau}{(1-\tau)^2} \\
&= \frac{(1-s)(1-t)}{t-s}\left[\frac{1}{1-\tau}\right]_s^t \\
&= \frac{(1-s)(1-t)}{t-s}\left(\frac{1}{1-t} - \frac{1}{1-s}\right) \\
&= \frac{(1-s)(1-t)}{t-s}\cdot\frac{t-s}{(1-t)(1-s)} = 1.
\end{aligned} \tag{61}$$

This completes the proof. □

C.2. Characterizing the two-time denoiser

Since $\delta_{s,t}$ lies on the simplex, it is natural to design objective functions that are entirely simplex-valued, for which cross-entropy can be used. To this end, we now translate the flow map characterizations from Theorem B.2 into conditions on $\delta_{s,t}$, and identify which remain simplex-valued.

Proposition C.5 (Flow map characterizations in δ space)**.** *The flow map characterizations from Theorem B.2 translate into the following conditions on $\delta_{s,t}$:*

(i) The Lagrangian condition. For all $\mathbf{x} \in \mathbb{R}^d$ and $(s,t) \in [0,1]^2$:

$$\delta_{s,t}(\mathbf{x}) + \frac{(1-t)(t-s)}{1-s}\partial_t \delta_{s,t}(\mathbf{x}) = \delta_{t,t}(X_{s,t}(\mathbf{x})). \tag{62}$$

(ii) The Eulerian condition. For all $\mathbf{x} \in \mathbb{R}^d$ and $(s,t) \in [0,1]^2$:

$$\partial_s \delta_{s,t}(\mathbf{x}) + \frac{\delta_{s,s}(\mathbf{x}) - \mathbf{x}}{1-s}\cdot\nabla\delta_{s,t}(\mathbf{x}) = \frac{1-t}{(1-s)(t-s)}\big(\delta_{s,t}(\mathbf{x}) - \delta_{s,s}(\mathbf{x})\big). \tag{63}$$

(iii) The semigroup condition. For all $\mathbf{x} \in \mathbb{R}^d$ and $(s,u,t) \in [0,1]^3$,

$$\begin{aligned}
\delta_{s,t}(\mathbf{x}) &= \gamma\cdot\delta_{s,u}(\mathbf{x}) + (1-\gamma)\cdot\delta_{u,t}(X_{s,u}(\mathbf{x})), \\
\gamma &= \frac{(1-t)(u-s)}{(1-u)(t-s)}.
\end{aligned} \tag{64}$$

When $s \le u \le t$, the coefficients satisfy $\gamma, 1-\gamma \ge 0$, so this is a convex combination. At the midpoint $u = (s+t)/2$, the weights simplify to $\gamma = (1-t)/(2-s-t)$ and $1-\gamma = (1-s)/(2-s-t)$.

Proof. (i) Lagrangian. Differentiating the convex combination (48) in t:

$$\partial_t X_{s,t}(\mathbf{x}) = -\frac{1}{1-s}\mathbf{x} + \frac{1}{1-s}\delta_{s,t}(\mathbf{x}) + \frac{t-s}{1-s}\partial_t \delta_{s,t}(\mathbf{x}). \tag{65}$$

By the Lagrangian equation (28), $\partial_t X_{s,t}(\mathbf{x}) = b_t(X_{s,t}(\mathbf{x}))$. Rewriting b_t via the denoiser-velocity relation (45):

$$\partial_t X_{s,t}(\mathbf{x}) = \frac{D_t(X_{s,t}(\mathbf{x})) - X_{s,t}(\mathbf{x})}{1-t}. \tag{66}$$

Multiplying both sides by $(1-t)$, adding $X_{s,t}$, and substituting (48):

$$\begin{aligned} D_t(X_{s,t}(\mathbf{x})) &= X_{s,t}(\mathbf{x}) + (1-t)\partial_t X_{s,t}(\mathbf{x}) \\ &= \underbrace{\frac{1-t}{1-s}\mathbf{x} - \frac{1-t}{1-s}\mathbf{x}}_{=0} + \underbrace{\frac{t-s}{1-s}\delta_{s,t}(\mathbf{x}) + \frac{1-t}{1-s}\delta_{s,t}(\mathbf{x})}_{=\delta_{s,t}(\mathbf{x})} + \frac{(1-t)(t-s)}{1-s}\partial_t\delta_{s,t}(\mathbf{x}) \\ &= \delta_{s,t}(\mathbf{x}) + \frac{(1-t)(t-s)}{1-s}\partial_t\delta_{s,t}(\mathbf{x}). \end{aligned} \tag{67}$$

Since $\delta_{t,t} = D_t$ on the diagonal (Theorem C.3), the left-hand side is $\delta_{t,t}(X_{s,t}(\mathbf{x}))$, giving (62).

(ii) Eulerian. We substitute (48) into the Eulerian equation (29). Differentiating (48) in s:

$$\partial_s X_{s,t}(\mathbf{x}) = \frac{1-t}{(1-s)^2}(\mathbf{x} - \delta_{s,t}(\mathbf{x})) + \frac{t-s}{1-s}\partial_s\delta_{s,t}(\mathbf{x}). \tag{68}$$

The spatial Jacobian of (48) is:

$$\nabla X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\mathrm{Id} + \frac{t-s}{1-s}\nabla\delta_{s,t}(\mathbf{x}). \tag{69}$$

By the denoiser-velocity relation (45), the advection velocity is $b_s(\mathbf{x}) = (\delta_{s,s}(\mathbf{x}) - \mathbf{x})/(1-s)$. Substituting into (29) and expanding $b_s \cdot \nabla X_{s,t}$:

$$\begin{aligned} 0 = &\frac{1-t}{(1-s)^2}(\mathbf{x} - \delta_{s,t}(\mathbf{x})) + \frac{t-s}{1-s}\partial_s\delta_{s,t}(\mathbf{x}) \\ &+ \frac{(1-t)(\delta_{s,s}(\mathbf{x}) - \mathbf{x})}{(1-s)^2} + \frac{(t-s)(\delta_{s,s}(\mathbf{x}) - \mathbf{x})}{(1-s)^2} \cdot \nabla\delta_{s,t}(\mathbf{x}). \end{aligned} \tag{70}$$

The first and third terms combine to $\frac{1-t}{(1-s)^2}(\delta_{s,s}(\mathbf{x}) - \delta_{s,t}(\mathbf{x}))$. Dividing through by $\frac{t-s}{1-s}$ and rearranging gives (63).

(iii) Semigroup. We express each side of $X_{s,t}(\mathbf{x}) = X_{u,t}(X_{s,u}(\mathbf{x}))$ using (48). The left-hand side is:

$$X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\mathbf{x} + \frac{t-s}{1-s}\delta_{s,t}(\mathbf{x}). \tag{71}$$

For the right-hand side, define $\mathbf{z} := X_{s,u}(\mathbf{x}) = \frac{1-u}{1-s}\mathbf{x} + \frac{u-s}{1-s}\delta_{s,u}(\mathbf{x})$. Then:

$$\begin{aligned} X_{u,t}(\mathbf{z}) &= \frac{1-t}{1-u}\mathbf{z} + \frac{t-u}{1-u}\delta_{u,t}(\mathbf{z}) \\ &= \frac{1-t}{1-u}\left[\frac{1-u}{1-s}\mathbf{x} + \frac{u-s}{1-s}\delta_{s,u}(\mathbf{x})\right] + \frac{t-u}{1-u}\delta_{u,t}(\mathbf{z}) \\ &= \frac{1-t}{1-s}\mathbf{x} + \frac{(1-t)(u-s)}{(1-u)(1-s)}\delta_{s,u}(\mathbf{x}) + \frac{t-u}{1-u}\delta_{u,t}(\mathbf{z}). \end{aligned} \tag{72}$$

Equating with the left-hand side and cancelling $\frac{1-t}{1-s}\mathbf{x}$:

$$\frac{t-s}{1-s}\delta_{s,t}(\mathbf{x}) = \frac{(1-t)(u-s)}{(1-u)(1-s)}\delta_{s,u}(\mathbf{x}) + \frac{t-u}{1-u}\delta_{u,t}(\mathbf{z}). \tag{73}$$

Multiplying both sides by $(1-s)/(t-s)$:

$$\delta_{s,t}(\mathbf{x}) = \frac{(1-t)(u-s)}{(1-u)(t-s)}\delta_{s,u}(\mathbf{x}) + \frac{(t-u)(1-s)}{(1-u)(t-s)}\delta_{u,t}(\mathbf{z}). \tag{74}$$

Define $\gamma := (1-t)(u-s)/\big((1-u)(t-s)\big)$. When $s \le u \le t \le 1$, every factor is non-negative, so $\gamma \ge 0$. To show the second coefficient equals $1-\gamma$, we verify the two coefficients sum to one:

$$\begin{aligned}(1-t)(u-s) + (t-u)(1-s) &= u - s - tu + ts + t - ts - u + us \\ &= (t-s) - u(t-s) \\ &= (t-s)(1-u).\end{aligned} \tag{75}$$

Dividing by $(1-u)(t-s)$ confirms the coefficients sum to one, giving (64). □

The right-hand side of the Lagrangian condition (62) equals $D_t(X_{s,t}(\mathbf{x}))$ and lies on the simplex, providing a natural teacher as we describe below. The Eulerian characterization (63) is self-contained in δ, but we were unable to identify a teacher-student decomposition that clearly lives on the simplex. By Theorem C.4, both $\delta_{s,u}$ and $\delta_{u,t}$ lie on the simplex, so for $s < u < t$ the semigroup condition (64) is a convex combination of simplex elements, making it amenable to a teacher-student decomposition.

Remark C.6 (The composite denoiser)**.** The object $D_{s,t}(\mathbf{x}) := D_t(X_{s,t}(\mathbf{x}))$, which flows from s to t and then applies the single-time denoiser, appears in two places above: as the integrand $D_\tau(X_{s,\tau}(\mathbf{x}))$ in the simplex proof (60), and as the Lagrangian teacher $\delta_{t,t}(X_{s,t}(\mathbf{x})) = D_t(X_{s,t}(\mathbf{x}))$. Using the Lagrangian equation and the convex combination (48), it can be expressed as $D_{s,t}(\mathbf{x}) = \delta_{s,t}(\mathbf{x}) + \frac{(1-t)(t-s)}{1-s}\partial_t\delta_{s,t}(\mathbf{x})$, which is precisely the student in the Lagrangian loss (79). On the diagonal, the prefactor $(t-s)$ vanishes, recovering $D_{s,s}(\mathbf{x}) = \delta_{s,s}(\mathbf{x}) = D_s(\mathbf{x})$. The composite denoiser also satisfies a composition property $D_{s,u}(\mathbf{x}) = D_{t,u}(X_{s,t}(\mathbf{x}))$ (immediate from the semigroup) and an Eulerian equation $\partial_s D_{s,t}(\mathbf{x}) + b_s(\mathbf{x}) \cdot \nabla D_{s,t}(\mathbf{x}) = 0$ (by the chain rule and the Eulerian equation for the flow map). Unlike the two-time denoiser $\delta_{s,t}(\mathbf{x})$, the composite denoiser cannot be composed for multi-step sampling: it always predicts the endpoint $\mathbf{x}_1$, and recovering the flow map from $D_{s,t}(\mathbf{x})$ requires integrating an ODE. The two-time denoiser avoids this by recovering the flow map algebraically via (48).

C.3. Learning the two-time denoiser

The characterizations of $\delta_{s,t}$ developed above can be reformulated as training objectives. Since $\delta_{s,t}(\mathbf{x})$ lies on the simplex, it is natural to use cross-entropy, as we do for the single-time denoiser in the main text (12). These objectives can be used in two modes: *distillation* from a pre-trained denoiser, or *self-distillation* (direct training) without a pre-trained teacher.

Distillation from a pre-trained denoiser. Given a pre-trained denoiser $\hat{D}_s$, we can distill it into a two-time denoiser $\hat{\delta}_{s,t}$ by minimizing objectives derived from the characterizations in Theorem C.5.

Proposition C.7 (Denoiser distillation)**.** *The two-time denoiser $\delta_{s,t}$ is the unique minimizer of*

$$\mathcal{L}_\delta(\hat{\delta}) = \mathcal{L}_\delta^{\text{diag}}(\hat{\delta}) + \mathcal{L}_\delta^{\text{off}}(\hat{\delta}), \tag{76}$$

where the diagonal loss is the cross-entropy against the pre-trained denoiser $\hat{D}_s$:

$$\mathcal{L}_\delta^{\text{diag}}(\hat{\delta}) = \int_0^1 \mathbb{E}\big[\ell_{\mathsf{CE}}(\hat{\delta}_{s,s}(I_s), \hat{D}_s(I_s))\big]\mathrm{d}s, \tag{77}$$

and $\mathcal{L}_\delta^{\text{off}}$ is one of the following off-diagonal objectives:

*(i) The progressive (*PMD*) loss:*

$$\mathcal{L}_{\mathsf{PMD}}(\hat{\delta}) = \int_0^1\int_0^t\int_s^t \mathbb{E}\big[\ell_{\mathsf{CE}}(\hat{\delta}_{s,t}(I_s), \gamma\cdot\hat{\delta}_{s,u}(I_s) + (1-\gamma)\cdot\hat{\delta}_{u,t}(\hat{X}_{s,u}(I_s)))\big]\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t, \tag{78}$$

where γ is defined in (64) and $\hat{X}_{s,u}(\mathbf{x})$ is recovered from $\hat{\delta}_{s,u}$ via (48).

(ii) The Lagrangian (LMD*) loss:*

$$\mathcal{L}_{\mathsf{LMD}}(\hat{\delta}) = \int_0^1 \int_0^t \mathbb{E}\Big[\ell_{\mathsf{CE}}\Big(\hat{\delta}_{s,t}(I_s) + \tfrac{(1-t)(t-s)}{1-s}\partial_t \hat{\delta}_{s,t}(I_s),\ \hat{D}_t(\hat{X}_{s,t}(I_s))\Big)\Big]\mathrm{d}s\,\mathrm{d}t, \tag{79}$$

where $\hat{X}_{s,t}$ is recovered from $\hat{\delta}_{s,t}$ via (48).

Proof. Each term is a cross-entropy against a fixed target, hence non-negative, with equality if and only if the prediction matches the target. The diagonal is zero if and only if $\hat{\delta}_{s,s} = \hat{D}_s$; the off-diagonal is zero if and only if the corresponding characterization from Theorem C.5 holds. Together, $\mathcal{L}_\delta = 0$ if and only if $\hat{\delta} = \delta$, giving uniqueness. □

Self-distillation. As with the flow map, it is also useful to train the two-time denoiser from scratch. The key insight is that the diagonal $\hat{\delta}_{s,s} = \hat{D}_s$ can be trained via cross-entropy against one-hot targets $\mathbf{x}_1$ (Theorem 3.1). Combining this with self-consistent versions of the off-diagonal objectives above leads to a single-phase training approach.

Proposition C.8 (Denoiser self-distillation)**.** *The two-time denoiser $\delta_{s,t}$ is the unique minimizer of*

$$\mathcal{L}_\delta^{sd}(\hat{\delta}) = \mathcal{L}_\delta^{diag,\ sd}(\hat{\delta}) + \mathcal{L}_\delta^{off}(\hat{\delta}), \tag{80}$$

where the diagonal loss is the standard flow matching cross-entropy:

$$\mathcal{L}_\delta^{diag,\ sd}(\hat{\delta}) = \int_0^1 \mathbb{E}\big[\ell_{\mathsf{CE}}(\hat{\delta}_{s,s}(I_s), \mathbf{x}_1)\big]\mathrm{d}s, \tag{81}$$

and $\mathcal{L}_\delta^{off}$ is one of the following off-diagonal objectives:

(i) The progressive (PSD*) loss (78) from Theorem C.7:*

$$\mathcal{L}_{\mathsf{PSD}}(\hat{\delta}) = \int_0^1 \int_0^t \int_s^t \mathbb{E}\big[\ell_{\mathsf{CE}}(\hat{\delta}_{s,t}(I_s), \gamma \cdot \hat{\delta}_{s,u}(I_s) + (1-\gamma)\cdot \hat{\delta}_{u,t}(\hat{X}_{s,u}(I_s)))\big]\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t. \tag{82}$$

(ii) The Lagrangian (LSD*) loss:*

$$\mathcal{L}_{\mathsf{LSD}}(\hat{\delta}) = \int_0^1 \int_0^t \mathbb{E}\Big[\ell_{\mathsf{CE}}\Big(\hat{\delta}_{s,t}(I_s) + \tfrac{(1-t)(t-s)}{1-s}\partial_t \hat{\delta}_{s,t}(I_s),\ \hat{\delta}_{t,t}(\hat{X}_{s,t}(I_s))\Big)\Big]\mathrm{d}s\,\mathrm{d}t. \tag{83}$$

Proof. The off-diagonal terms are cross-entropies, hence non-negative, with equality if and only if the characterization holds. The diagonal cross-entropy against one-hot targets $\mathbf{x}_1$ is bounded below by the conditional entropy $H(\mathbf{x}_1 \mid I_s)$, achieved when $\hat{\delta}_{s,s} = D_s$. The total loss is therefore bounded below by $\int_0^1 \mathbb{E}[H(\mathbf{x}_1 \mid I_s)]\mathrm{d}s$, with equality if and only if $\hat{\delta} = \delta$. □

The progressive loss is naturally well-defined when training with cross-entropy: parameterizing $\hat{\delta}$ with a softmax output layer ensures the student is simplex-valued, and the teacher is a convex combination of simplex elements by (64). By contrast, the Lagrangian loss is more subtle. The teacher lies on the simplex since $\delta_{t,t} = D_t$ by Theorem C.3, but the student $\hat{\delta}_{s,t}(\mathbf{x}) + \frac{(1-t)(t-s)}{1-s}\partial_t \hat{\delta}_{s,t}(\mathbf{x})$ is simplex-valued only at optimality, and there is no obvious network architecture that enforces this constraint by construction. Despite this difficulty, cross-entropy acts as a barrier function that prevents departure from the simplex interior. To leverage this property, clipping gradients during initial optimization to drive an off-simplex student to the interior of the simplex suffices to ensure well-defined training.

In practice, we apply a stop-gradient operator $\mathsf{sg}(\cdot)$ to the teacher terms to ensure a well-defined separation between the teacher and student when they share parameters. The progressive teacher involves $\hat{\delta}$ terms in both distillation and self-distillation, so stop-gradient is always required. The distillation Lagrangian teacher $\hat{D}_t$ is already frozen; stop-gradient is only needed in the self-distillation variant, leading to the objectives

$$\begin{aligned}
\mathcal{L}_{\mathsf{PSD}}^{\mathsf{sg}}(\hat{\delta}) &= \int_0^1 \int_0^t \int_s^t \mathbb{E}\big[\ell_{\mathsf{CE}}(\hat{\delta}_{s,t}(I_s), \mathsf{sg}(\gamma \cdot \hat{\delta}_{s,u}(I_s) + (1-\gamma)\cdot \hat{\delta}_{u,t}(\hat{X}_{s,u}(I_s))))\big]\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t, \\
\mathcal{L}_{\mathsf{LSD}}^{\mathsf{sg}}(\hat{\delta}) &= \int_0^1 \int_0^t \mathbb{E}\Big[\ell_{\mathsf{CE}}\Big(\hat{\delta}_{s,t}(I_s) + \tfrac{(1-t)(t-s)}{1-s}\partial_t \hat{\delta}_{s,t}(I_s),\ \mathsf{sg}(\hat{\delta}_{t,t}(\hat{X}_{s,t}(I_s)))\Big)\Big]\mathrm{d}s\,\mathrm{d}t.
\end{aligned} \tag{84}$$

D. Auxiliary results

D.1. Proof of Lemma 3.1

We prove that the optimal denoiser output at each token position equals the posterior probability over the vocabulary.

Proof. From (9), we have that for the l-th token position:

$$D_t(\mathbf{x})^l = \mathbb{E}[\mathbf{x}_1^l | I_t = \mathbf{x}]. \tag{85}$$

Let $\mathbf{e}_i \in \mathbb{R}^{|V|}$ the one-hot encoding of the i-th subword in the vocabulary V. Since $\mathbf{x}_1^l$ is a one-hot vector, it takes values in $\{\mathbf{e}_1, \dots \mathbf{e}_{|V|}\}$. Then the conditional expectation above can be expanded as follows:

$$D_t(\mathbf{x})^l = \mathbb{E}[\mathbf{x}_1^l | I_t = \mathbf{x}] = \sum_{i=1}^{|V|} \mathbf{e}_i \cdot p_{1|t}^l(\mathbf{x}_1^l = \mathbf{e}_i | I_t = \mathbf{x}) = \begin{bmatrix} p_{1|t}^l(\mathbf{x}_1^l = \mathbf{e}_1 | I_t = \mathbf{x}) \\ \dots \\ p_{1|t}^l(\mathbf{x}_1^l = \mathbf{e}_{|V|} | I_t = \mathbf{x}) \end{bmatrix}. \tag{86}$$

This is the vector of posterior probability over the vocabulary, $p_{1|t}^l(\cdot | I_t = \mathbf{x})$. □

D.2. Proof of Theorem 3.2

We prove that the minimizers of $\mathcal{L}_{\mathsf{MSE}}(\hat{D})$ (10) and $\mathcal{L}_{\mathsf{CE}}(\hat{D})$ (12) when $\hat{D}$ uses tokenwise softmax at the output are identical. This result has been known in Dieleman et al. (2022) and Eijkelboom et al. (2024), but we write out a proof for completeness.

Proof. By Lemma 3.1, the optimal prediction target for $\mathcal{L}_{\mathsf{MSE}}(\hat{D})$ at each token position is given as $D_t(\mathbf{x})^l = p_{1|t}^l(\cdot | I_t = \mathbf{x})$. Now consider the cross-entropy loss $\mathcal{L}_{\mathsf{CE}}(\hat{D})$ (12) and its minimizer D^{CE}:

$$\mathcal{L}_{\mathsf{CE}}(\hat{D}) = \int_0^1 \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1} \left[-\sum_{l=1}^{L} \log \hat{p}_{1|t}^l(\mathbf{x}_1^l | I_t) \right] \mathrm{d}t = \int_0^1 \mathbb{E}_{I_t} \sum_{l=1}^{L} \mathbb{E}_{\mathbf{x}_1^l | I_t} [-\log \hat{p}_{1|t}^l(\mathbf{x}_1^l | I_t)] \mathrm{d}t. \tag{87}$$

The inner term is the cross-entropy between the true token-wise posterior $p_{1|t}^l(\cdot | I_t)$ and the predicted distribution $\hat{p}_{1|t}^l(\cdot | I_t) = \hat{D}_t(I_t)^l$, which is minimized when the two are equal:

$$D_t^{\mathsf{CE}}(\mathbf{x})^l = \begin{bmatrix} p_{1|t}^l(\mathbf{x}_1^l = \mathbf{e}_1 | I_t = \mathbf{x}) \\ \dots \\ p_{1|t}^l(\mathbf{x}_1^l = \mathbf{e}_{|V|} | I_t = \mathbf{x}) \end{bmatrix}. \tag{88}$$

By comparing this with (11), we obtain that the optimal denoiser can be learned via cross-entropy loss. □

D.3. Sample-level transport maps for discrete diffusion

We provide a supplementary result that, unlike continuous diffusion processes that admit a sample-level flow map, it is not generally possible in discrete diffusion processes to find a sample-level deterministic map that accurately transports noise to data in one step. Fundamentally, this limitation arises because deterministic maps can never split probability mass (Villani et al., 2008). We show a general argument:

Proposition D.1. *Let S be a finite set. For any probability distribution μ on S, there always exists a distribution ν on S that cannot be reached from μ through deterministic samplewise transport $f : S \to S$.*

Proof. We remark that a deterministic map $f : S \to S$ pushes forward a distribution μ to another distribution $\nu = f_{\#}\mu$ as:

$$\nu(\mathbf{y}) = \sum_{\mathbf{x} \in f^{-1}(\mathbf{y})} \mu(\mathbf{x}), \tag{89}$$

where $f^{-1}(\mathbf{y})$ denotes the preimage of $\{\mathbf{y}\}$.

This means every probability $\nu(\mathbf{y})$ in the output distribution must be a subset sum of the original probability $\{\mu(\mathbf{x}) \mid \mathbf{x} \in S\}$. For any given μ, let $\mu_{\min}$ be the smallest nonzero probability in $\{\mu(\mathbf{x}) \mid \mathbf{x} \in S\}$. Construct ν such that an element of S has the probability $\nu(\mathbf{y}) = p_{\min}/2$. Since this probability cannot be expressed as a subset sum of μ, the distribution ν can never be reached from μ through pushforward by a deterministic function f. □

By choosing $S = V^L$ and $(\mu, \nu) = (p_0, p)$ for discrete diffusion, we can see that for any noise distribution p_0, there exists a data distribution p that cannot be exactly reached via one-step transport through a deterministic map.

D.4. First-stage distillation loss

Recall our two-model parameterization (23):

$$\hat{X}_{s,t}(\mathbf{x}) = \mathbf{x} + (t-s)\,\hat{b}_s(\mathbf{x}) + \frac{1}{2}(t-s)^2\hat{\psi}_{s,t}(\mathbf{x}). \tag{90}$$

We initialize $\hat{\psi}$ from the parameters of $\hat{b}$ by removing the output softmax and zeroing the final layer, and train using the semigroup loss (15) re-written in terms of clean data prediction. For this, observe that the average velocity $\hat{v}$ is given as follows, from (14):

$$\hat{v}_{s,t}(\mathbf{x}) = \frac{\hat{X}_{s,t}(\mathbf{x}) - \mathbf{x}}{t-s} = \hat{b}_s(\mathbf{x}) + \frac{1}{2}(t-s)\,\hat{\psi}_{s,t}(\mathbf{x}). \tag{91}$$

Using the relationship between the denoiser $\hat{D}$ and velocity $\hat{b}$ in (9), the integrand of the semigroup-based loss in (15) can be written as follows:

$$\begin{aligned}
\mathbb{E}|\hat{X}_{s,t}(I_s) - \mathsf{sg}(\hat{X}_{u,t}(\hat{X}_{s,u}(I_s)))|^2 &= \mathbb{E}|I_s + (t-s)\hat{v}_{s,t}(I_s) - \mathsf{sg}(I_s + (t-s)\bar{v}_{s,t})|^2 \\
&= (t-s)^2\,\mathbb{E}\,|\hat{v}_{s,t}(I_s) - \mathsf{sg}(\bar{v}_{s,t})|^2 \\
&= (t-s)^2\,\mathbb{E}\left|\frac{\hat{D}_s(I_s) - I_s}{1-s} + \frac{t-s}{2}\hat{\psi}_{s,t}(I_s) - \mathsf{sg}(\frac{\bar{\mathbf{x}}_1 - I_s}{1-s})\right|^2 \\
&= \frac{(t-s)^2}{(1-s)^2}\,\mathbb{E}\left|\hat{D}_s(I_s) - I_s + \frac{(t-s)(1-s)}{2}\hat{\psi}_{s,t}(I_s) - \mathsf{sg}(\bar{\mathbf{x}}_1 - I_s)\right|^2 \\
&= \frac{(t-s)^2}{(1-s)^2}\,\mathbb{E}\left|\hat{D}_s(I_s) + \frac{(t-s)(1-s)}{2}\hat{\psi}_{s,t}(I_s) - \mathsf{sg}(\bar{\mathbf{x}}_1)\right|^2
\end{aligned} \tag{92}$$

where the bootstrapped velocity $\bar{v}_{s,t}$ and target $\bar{\mathbf{x}}_1$ are given as:

$$\bar{v}_{s,t} \coloneqq \frac{u-s}{t-s}\hat{v}_{s,u}(I_s) + \frac{t-u}{t-s}\hat{v}_{u,t}(\hat{X}_{s,u}(I_s)), \qquad \bar{\mathbf{x}}_1 \coloneqq \mathsf{sg}(I_s + (1-s)\bar{v}_{s,t}). \tag{93}$$

Following Boffi et al. (2025a), we drop the scale term $(\frac{t-s}{1-s})^2$ which changes the effective learning rate depending on the step sizes $t-s$ and $1-s$, for additional training stability. Then the final denoising loss on the correction model $\hat{\psi}$ becomes:

$$\mathcal{L}_{\mathsf{MSE}}(\hat{\psi}) = \int_0^1\int_0^t\int_s^t \mathbb{E}\left|\hat{D}_s(I_s) + \frac{(t-s)(1-s)}{2}\hat{\psi}_{s,t}(I_s) - \mathsf{sg}(\bar{\mathbf{x}}_1)\right|^2 \mathrm{d}u\mathrm{d}s\mathrm{d}t. \tag{94}$$

E. Implementation details

Time reparameterization. To efficiently implement the time reparameterization $\tau(t)$ described in Sec. 4.1 without evaluating the probability sum during training, we utilize a precomputed lookup table (LUT) combined with spline interpolation. Specifically, we approximate the cumulative density function (CDF) of (22) using Gauss-Hermite quadrature and evaluate it on a equispaced grid of $1,000$ points over $t \in [0,1]$, obtaining (t, τ) pairs at each point. We find that this resolution is sufficient to capture the transition of the schedule with negligible error. From these discrete pairs, we fit a cubic spline to obtain a continuous and differentiable mapping, and then construct both the forward map $\tau(t)$ and the inverse map $t(\tau)$, which enables $O(1)$ sampling of simulation times during training. Since this LUT and the associated mappings are computed once prior to training and can be cached, our approach incurs no additional computational overhead.

Training details. Both for LM1B and OWT we train FLM from scratch for 1M training steps, with batch size of 512. Following the settings from Sahoo et al. (2025), we use 2,500 warmup steps and then a constant learning rate of 3×10^{-4}. For the optimizer, we use Adam (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Additionally, we utilize softcapping (Riviere et al., 2024) which smooths out large logits in the attention activations, for additional numerical stability of training. For FMLM, we share all the training settings with FLM including batch size and learning rate. For LM1B, we report the results from two-phase distillation of 100k steps for each phase. For OWT, we report the first-phase distilled results for 300k steps, where we additionally use progressive warm-up of the distillation step size h: instead of drawing $h \sim \mathsf{U}[0,1]$ throughout training, we start with $h \sim \mathsf{U}[0, \frac{1}{1024}]$ and double the upper bound every 10k steps until it reaches 1. Lastly, we find that the reparameterization $\tau(t)$ has a flat region near $t = 0$ (Fig. 4), causing the start point s to rarely land near the origin. This hinders learning of flow maps for one- or two-step generation, where the model must directly transport from $s = 0$ to $t = 1$. To address this, we fix a probability of directly sampling the boundary: $(s,t) = (0,1)$ for LM1B with a probability of 1/64, and $s = 0$ for OWT with a probability of 1/32, ensuring the model receives sufficient training signal for few-step generation.

Sampling details. For FLM on both LM1B and OWT, we use Euler solver for sampling. For FMLM, on LM1B we use the standard flow map jumps for sampling as described in the main text. For OWT, we leverage the "γ-sampling" algorithm from Kim et al. (2024) using the optimal γ values (Sabour et al., 2025). In addition, for OWT we find it beneficial to alter the time reparameterization at inference time as follows: we define the reparameterized time $\tau'(t)$ for sampling as a convex combination with the original time, $\tau'(t) := \alpha\tau(t) + (1-\alpha)t$, and use optimal α values within $\{0.5, 0.75, 1\}$.

Many-step baselines. For the LM1B experiments, we trained Duo (Sahoo et al., 2025), MDLM (Sahoo et al., 2024), and CANDI (Pynadath et al., 2025) from scratch using identical settings, while utilizing the official 1M-step checkpoint for RDLM (Jo & Hwang, 2025). For the OWT experiments, we relied on the official checkpoints provided by the respective authors for CANDI, Duo, and MDLM. Due to absence of the official checkpoint and the limited resource for reproducing, we were not able to compare with RDLM in OWT. For sampling, for all the discrete baselines we used ancestral sampler with temperature 1.0, while for RDLM we use the SDE sampler proposed in the paper.

Few-step baselines. For LM1B experiments, we apply SDTT (Wu et al., 2025) on top of MDLM (Sahoo et al., 2024), trained on LM1B for 1M steps. Following the default hyperparameters from the paper, we use a fixed learning rate of 6×10^{-5} with 2,500 warmup steps and batch size of 128. Each distillation round consists of 10k training steps where we perform a total of 8 rounds. We share this setting when applying DCD (Sahoo et al., 2025) on top of Duo trained for 1M steps. For OWT, we leverage official distilled checkpoints from repective authors. For Di4C (Hayakawa et al., 2024), we used the intermediate checkpoints with the best 32-step performance among the training, corresponding to 20k training steps for LM1B and 50k for OWT: in both cases, additional training resulted in performance degradation.

Table 4. Self-BLEU (Zhu et al., 2018) score of 1,024 generated samples from each baselines and FMLM in the one-step generation setting. Lower score denotes more n-gram diversity. For reference, we report the Self-BLEU score of mode-collapsed case when all the samples are identical, and the score of the reference samples from each dataset.

Dataset	LM1B	OWT
(mode collapse)	1.000	1.000
Duo + DCD	0.075	0.297
Duo + Di4C	0.054	0.272
MDLM + SDTT	0.026	0.036
MDLM + Di4C	0.023	0.031
FMLM (Ours)	0.073	0.121
Dataset	0.047	0.046

Table 5. Generation performance of the two flow map models on LM1B across 1 to 1024 sampling steps.

	First-stage distilled		Second-stage distilled	
Steps	Gen. PPL (↓)	Entropy	Gen. PPL (↓)	Entropy
1	102.49	4.13	104.37	4.12
2	93.65	4.17	95.42	4.15
4	88.86	4.17	90.90	4.16
8	84.57	4.19	85.72	4.17
16	80.86	4.18	80.50	4.17
32	77.04	4.19	75.99	4.16
64	73.86	4.18	71.90	4.16
128	70.78	4.19	69.07	4.16
256	68.39	4.20	66.39	4.15
512	67.47	4.20	64.05	4.15
1024	67.63	4.20	62.17	4.14

F. Supplementary evaluation results

Checking mode collapse. To ensure that FMLM does not mode collapse onto a few high-quality samples, we additionally report the Self-BLEU (Zhu et al., 2018) score which measures the n-gram diversity of generations. The results in Tab. 4 shows that FMLM clearly does not show mode-collapsing behavior in one-step generation, which would be indicated by a Self-BLEU score ≈1.0, as it attains only a slightly worse score compared to real data.

Comparison of first-stage and second-stage distilled flow maps. In Tab. 5 we present the performance of FMLM across both distillation phases. The first-stage distilled model uses the two-model parameterization of the flow map (23), while the second-stage distilled model uses the single-model parameterization (14). As explained in Sec. 4.2, the first-stage model is distilled from a fixed FLM via semigroup loss (15), and the second-stage model is distilled from the first-stage model using a simple squared regression loss. We observe that the final single-model student successfully recovers the performance of its two-model teacher across all sampling step counts, demonstrating effective knowledge transfer between the two parameterizations of flow map. In the main results from Tab. 2 and Fig. 6, we report the performance of second-phase distilled model to match the model size; for the ablation study we use the first-phase distilled model.

Improving sample quality via autoguidance. Many previous work motivated continuous denoising for language by guidance algorithms that extrapolate score functions or flow velocities to amplify the influence of conditioning or generally improve the quality of samples (Strudel et al., 2022; Dieleman et al., 2022; Jo & Hwang, 2025). However, this aspect needs a more careful examination since it is possible to guide discrete denoising models as well by extrapolating probability logits (Schiff et al., 2024). Due to our interest in unconditional generation, we focus on autoguidance (Karras et al., 2024a) which guides a denoising model with a weak version of itself. For continuous denoising, guidance is implemented using:

$$\hat{b}^{(\text{guided})} = \hat{b}^{(\text{weak})} + \eta(\hat{b} - \hat{b}^{(\text{weak})}),$$

for strength $\eta > 1$. In clean data prediction, this becomes:

$$\hat{D}^{(\text{guided})} = \hat{D}^{(\text{weak})} + \eta(\hat{D} - \hat{D}^{(\text{weak})}).$$

Table 6. Comparison of autoguidance stability across varying guidance scales η and sampling steps from 128 to 1024, tested on LM1B. We mark as red for results with Gen. PPL over 1,000, and entropy under 3.9.

Steps	Guidance scale (η)	**FLM (Ours)**		Duo		MDLM	
		Gen. PPL (↓)	Entropy	Gen. PPL (↓)	Entropy	Gen. PPL (↓)	Entropy
128	0	112.54	4.34	91.64	4.32	113.88	4.32
	1	95.56	4.31	82.80	4.34	121.92	4.37
	2	87.18	4.29	89.60	4.38	141.91	4.40
	5	74.85	4.25	1671.20	4.26	851.46	4.64
	10	66.85	4.22	3013.31	4.73	2316.15	4.67
	20	60.99	4.19	3045.59	4.72	1897.82	4.57
	50	62.52	4.04	2947.73	4.71	3703.24	4.60
	100	13.58	2.60	2811.93	4.71	3471.61	4.39
256	0	104.59	4.32	97.58	4.30	112.81	4.33
	1	89.42	4.29	89.14	4.23	118.97	4.36
	2	81.53	4.27	81.20	4.34	141.72	4.41
	5	70.23	4.23	82.98	4.37	875.76	4.64
	10	62.01	4.20	1780.56	4.31	2064.23	4.64
	20	55.43	4.16	3008.36	4.72	1690.58	4.53
	50	52.69	4.06	2983.20	4.72	3333.58	4.62
	100	26.28	3.05	2802.72	4.71	3489.58	4.39
512	0	99.75	4.30	100.41	4.30	111.61	4.32
	1	86.48	4.27	90.06	4.32	116.82	4.36
	2	79.06	4.26	83.13	4.35	141.62	4.41
	5	68.40	4.22	78.57	4.36	878.42	4.63
	10	60.93	4.19	1773.36	4.35	1985.40	4.63
	20	55.31	4.14	3057.73	4.73	1461.70	4.50
	50	51.77	4.02	2943.49	4.72	2932.98	4.64
	100	48.00	3.52	2852.79	4.70	3455.59	4.39
1024	0	96.91	4.29	97.91	4.31	108.11	4.32
	1	84.37	4.27	89.48	4.32	115.72	4.36
	2	77.78	4.25	81.70	4.36	141.78	4.42
	5	67.50	4.22	74.58	4.35	886.67	4.63
	10	60.45	4.19	1916.89	4.42	1753.46	4.59
	20	55.49	4.14	3000.66	3.73	1423.58	4.47
	50	51.62	4.00	2928.61	4.72	2610.08	4.65
	100	51.93	3.74	2811.92	4.71	3455.55	4.38

Both extrapolations happen in Euclidean space. We note that, although each clean data prediction lies on the embedded simplex when using softmax, their extrapolation can exit the simplex freely.

For discrete denoising, guidance is implemented using

$$\log \hat{p}^{*(\text{guided})} = \log \hat{p}^{*(\text{weak})} + \eta(\log \hat{p}^{*} - \log \hat{p}^{*(\text{weak})}) + \text{const},$$

for strength $\eta > 1$. Under factorization, this extrapolates in logit space before softmax (Schiff et al., 2024), staying within the simplex multiplicatively. In Tab. 6, we present the sample quality and entropy while varying the guidance scale η from 1 to 100 for FLM, Duo (Sahoo et al., 2025) and MDLM (Sahoo et al., 2024). For the weak model, we use the final trained model with dropout of 0.1, as proposed in Karras et al. (2024a) as one way to construct a weak model.

Despite discrete guidance remaining on the simplex while continuous guidance can leave it, we find that discrete denoising becomes unstable at large guidance strengths while continuous denoising remains stable. We hypothesize this occurs because discrete guidance has the form $\hat{p}^{*(\text{guided})} \propto (\hat{p}^{*(\text{weak})})^{1-\eta}(\hat{p}^{*})^{\eta}$, where large η strongly amplifies modes in $\hat{p}^{*}$ not present in $\hat{p}^{*(\text{weak})}$. By factorization, the mode of $\hat{p}^{*(\text{weak})}$ becomes dispersed in the full state space, and extrapolating away from it removes more modes than necessary. In contrast, continuous denoising directly predicts velocities, or clean data, in Euclidean space (9) which allows avoiding this factorization-induced issue.

G. Supplementary qualitative results

More qualitative samples. Additional qualitative samples can be found in Figs. 8 to 13.

- Fig. 8 shows samples generated by FLM trained on LM1B with different sampling steps (32, 128, 256, 1024).
- Fig. 9 shows samples generated by FLM trained on OWT with different sampling steps (256, 1024).
- Fig. 10 shows **one-step** samples generated by FMLM trained on LM1B.
- Fig. 11 shows **one-step** samples generated by FMLM trained on OWT.
- Fig. 12 shows **one-step** samples generated by few-step masked discrete diffusion baselines trained on OWT.
- Fig. 13 shows **one-step** samples generated by few-step uniform discrete diffusion baselines trained on OWT.

Samples from fixed initial noise. In Figs. 14 to 17, we show samples generated by FMLM, MDLM + SDTT (Wu et al., 2025), and Duo + DCD (Sahoo et al., 2025) using different numbers of sampling steps from a fixed initial random seed, meaning that we generate the samples from a fixed starting noise. By the deterministic sampling procedure of FLM and FMLM, we observe that increasing the number of sampling steps recovers finer lexical details while preserving general structure. However, this behavior does not occur in discrete diffusion models because they rely on ancestral sampling over the entire vocabulary at every denoising step. This characteristic of FLM and FMLM leaves interesting directions of future work, such as applying noise inversion (Song et al., 2020a) for editing applications or interpolation between generated samples in the noise space.

Sampling Steps: 32 Gen.PPL: **106.87** — Entropy: **4.27**

```
[CLS]. martin rejected it because few companies vied for the technology and offered up any portion of the product
line as one option.  [CLS] woods next to be pro?  [CLS] when it’s up to you, get reminders out here or on the
first tee.  [CLS] meanwhile, national security adviser, gen.  ray fourniero, son of iraq national security gen.
james mcrver, that these changes are worth as much as " to the entire intelligence community " who oppose the
threat of baghdad.  [CLS] you stand in a certain position, and in situations that includes both states, thereby
open down doors and your partner in a manner [CLS]
```

Sampling Steps: 128 Gen.PPL: **86.65** — Entropy: **4.28**

```
[CLS] have a college degree.  [CLS] he said that even though more than 60 percent of the area got permits that
voters held a similar advantage in other states, a few stayed behind.  [CLS] there was no one here that ever
announced them, no one spent money or years at any time, except to give the two and maybempt them with another
chance, and knowing today that they are all they are is never to get any point about why they can play together.
[CLS] unalud has more than 5 % the country’s players.  [CLS] a spokesman for failing to respond to the comments
can post comments.  2̌022 a free phone number!  [CLS]"
```

Sampling Steps: 256 Gen.PPL: **76.74** — Entropy: **4.27**

```
[CLS] khan said, adding that any moves by the military would remain the verdict of the people if necessary.  [CLS]
the 21 - year - old fast bowler who won three of only 17 tests in australia in the build - up to the first test
and that former england captain lawrence dalirlio will be taken seriously.  [CLS] 17 mins took home gerrard’s
curling shot by rooney which had the rebound.  [CLS] " you are, like some us in the past, in charge of the truth.
[CLS] i expect a good living in these years or around 2010.  [CLS] an independent report has written to government
ministers meeting to recommend proposals " for [CLS]
```

Sampling Steps: 1024 Gen.PPL: **80.53** — Entropy: **4.32**

```
[CLS] has disappeared from the hills along the coast.  [CLS] the president made a brief appearance on chicago’s
grant park before mr.  bush.  [CLS] mr.  cuber estimated that harvard’s annual income could be of more than $ 200,
000 and bonuses could come in nearly $ 5, 000 annually.  [CLS] u.  s.  women plead guilty to her murder seattle,
april 28 ( upi ) - - a council has sent a judge making a later date for rev regarding the case of washington
teenager elementary school knox in the response to a death she is accused of.  [CLS] in capital markets, the
company’s fundamentals are clear [CLS]
```

Figure 8. Samples generated by FLM trained on LM1B with different sampling steps.

Sampling Steps: 256 Gen.PPL: **70.00** — Entropy: **5.30**

```
<|endoftext|> companies.
Officials at the rally at the ABAAC said on Monday, the Federalist Society had members have long believed that ABAAC could work.
"Together with labor, labor, and interest groups, state attorneys general, state and business leaders, the consumer-free market value
to you more than a handful of decades of choice and association," it reads in his remarks.
Now the court could have a similar effect on Friday.<|endoftext|>Here is an email from company to read, "the next time most of us are
watching 2, and they used to mean about a second.  The most accurate number?  I don’t know."
The email is just part of an esc altruism and the promotion of free software that is a plan to bring new ideas into the mainstream.
Like many nonhumans, their numbers are all that much, he’s aiming to lose half of their value by non-profits this year.
These technologies are being driven by Open researchers -- which uses them the most, for example, using carbon-generating batteries
for use in medical applications.  The team has found that nearly half of these devices used in science isn’t some magic feat.  To
show that Big intelligence may soon need to help tap into our everyday lives.
From a SETI perspective, however, many of the attendees are now less likely to hold him to it.  Science and science are a field
whose importance has grown, over the years.  That’s especially so large at the TED Electronics conference in 2012, which, despite the
increased numbers, also puts more people out there more than ever, too.
"Oh, and there’s more interest," Brner said.  But partly up for that is that donations are interested in numbers, until they go there
to support, say global projects, there’re way more people doing things to come.
Well, half a billion will not happen, really, either.  And, until there are ways for anyone to write a (sic),’d an article on
it, that’s it no way back, it’s creating more support for further exploration.<|endoftext|>Supporters of a generation have raised
concerns about living in a remote cell in a field such as Johannesburg that can make efforts to keep people who lack means to support
the cause feel futile.
Kired bio-hugil John Lee, 40, died in Grade 3 of 15 patients in Site C without information, despite the mainstay of the cold, dilated
lungs, said Bhupab Sengupta of Sierra Canada.  As cracks in the cell line go the heat and water has dried out because of drought,
said he believes there is in fact growth in the number of patients attending schoolchildren alive, but that an international process
of recognizing the value of research needs to roll out as more die.
"If reduced to 15 the number one priority; those well-boggaminated end, coll suicide with HIV another 50 million times, taking away
lives, who are trying to and end up being still a million other people could come to 15 die in the name of medical research," she
said via email.
As even Mr.  Singh struggled to leave the disease, an instinct was quicklyched on and a sense that his wife was going to end a life
he was; he saw no life.
He survived on a friend and no one near her, before a battle with kidney She failed and his memory was lost.  Almost exactly the same
the year, ALS donors began running tests to get Mr.  Lee’s attention, though he was on low levels of survival despite many of his
loved organs.
His patients, who usually have the same age and history, are central to the current generation of stem transpl medicine, and could be
one of some of the reasons he leaves behind.
READ MORE: Why ’Plant Part of the Dead Babies with Busy Wonness’
But after years of quietly testging the body of blood deep into this field, a critically important effort to make sure that it’s not
doctors, doctors, doctors or doctors, who will also be dying in it, have been ignored.
Spenasttha Bratman, a friend of John Doe as a happy single person who had long settled into a life without hope in the northern part
of South Africa, died last week spending 20 patients at the hospital, before walking away to die.
His father, then at Mount Canada, told Global News News that his team is focused on how they’re seeing one person use in a cell for
medical research, and if he adopits the idea, he could hold as many patients for cancer patients as a day of rest with a hospital,
near freezing, for patients.
His "has an position in the name of science right now, has been the loss of a couple who are<|endoftext|>
```

Sampling Steps: 1024 Gen.PPL: **62.60** — Entropy: **5.37**

```
<|endoftext|> program at the University of Utah when he said.
Meanwhile, FISUS President Tom Hickey praised the organization’s Rolexample for FIS Ratings, a rekindle of events dedicated to MLS.
However, the organization said he has been a vocal supporter of the LDS Bowl the last two years, which it said as well as corporate
campaign finance laws:
"For sure companies pay to be sure that Major Gar Soccer’s Measure Forg deal in 2011 will be the same thing forever."
The LDS organization has long been named after a Mormon gay-story building in Salt Lake City and decades of protests against gay
athletes and media outlets.  Mormonah has spent years to fight gay rights in the United States.  Critics have said the foundation for
the nation’s LGBT community remains the online system for LGBT discrimination.
In response, the LDS association said that’s why Google always aren’t up against MLS. "As a mayor no one has a chance to represent
their community," Susan Aylworth of Pride Business Group & group of business and American Chamber officials said.  "Public work,
public use of large venues, libraries and civic events can be working.  None of these factors are critical to our success."
The spokeswoman added, "It’s always my favorite publicly available," which requires that Google be removed from a later report on its
post.  That means MLS only has had 10 mayors on their website in the past 5 months or months.
"Google isn’t behind the scenes of Pride so no one has that thing anymore," she said.
City leaders have been working in recent years trying to pin the plug on a piece of Orlando’s city council development Soccer’s New
York headquarters and Miami headquarters.  Former state design firm Fincom also represents FIFA’s mayoral bid.
Ainsbach said he on the Miss America Tour co-’08 tour, but by this stage came out in the runup trying to enlist Saltber’s support,
including by last day’s tweet.  (Glen Seson -- M&T, rights)
"They really don’t laugh at us and have to pay the bills," Hales said, despite still not knowing a future future.  FIFA’s mayoral
elections are still 5 months away but did spell out how Adber would not be reapply on to hand over.
"He’s smart," Smith said.  "Mayorbody is able to put together a business strategy.  He’s doing it just because all of it helps our
office much a little.
"I mean I don’t get the Miquel award and I don’t have.  That’s a fact.  Because by Jim Scheiner is a wonderful guy who had the main
issue in going the gay Rights route and then we are very that.  And I’ve had to say percent that we all talk about it.  But you know
now, general area elected officials and space said real experience with a government so you know we need to do that are tough things
for you."
(Part 2 of interview)
"I’ll check the 2016 of every three years.  Same with the top 20 that we do with the MLS Cup Report.  I think M&R is -- Smith said
laughing.  "There is when there’ll be opportunities to 2000 (and ’17) in their history and we’ll tell them when they want the people.
Maybe others are playing better time to stay in 2020, or (they are actually missing some time.  On the States, when I was on the
board of Sports Illustrated, Al McGuone is afraid to say what people in ...  no doubt or whatever will have to tell them and he’s not
that.  I’ll certainly try words or whatever here and there.  I think one more MLS mayor than the last who has changed the league has
took this situation really seriously.
I guess there’s one?
A big one, for two reasons.  The second is that the overall base isn’t small.  I guess much of that has to be of the city as we’ve
got Coca-Benz doing new stuff coming million past year.  It’s not small there.  It’s big as well, but we didn’t work it all about
because it was a challenge.  It’s not a typical Bowl take.  It happened because there was willingness to pay attention and attention
to the city in these areas."
So, could you say before you were back?
It’s not.  Before that in certain areas, it was a concern -- an issue on the national LGBT community in U.S.<|endoftext|>
```

Figure 9. Samples generated by FLM trained on OWT with different sampling steps.

Sampling Steps: 1 Gen.PPL: **90.94** — Entropy: **4.13**

```
[CLS] be that people in the community, now this would give me for less.  [CLS] it’s that but the second of the
film will have been them there while some of them are not get them for week.  [CLS] at any in his own country, he
was no one who had working money, if there had be no at the first of two of what real not think that’s life in
white is the best, i think, don’t the right of his, but they can could make it not for or for a high - business
team.  [CLS] if she said that women, many years she was not expected to say [CLS]
```

Sampling Steps: 1 Gen.PPL: **94.71** — Entropy: **4.19**

```
[CLS] the public and private sectors, especially the condition that they’be in on.  [CLS] i was on tour and i’m
trying to tell myself he was no 8 ; he won just one against how to the people to get to the finals and yes, the
rest of ireland - - youon also have senior people in the american squad.  [CLS] in this respect, it’s kind of "
public " that’s worse for worse than that four million americans who tend to call 2006’s security as a threat or a
real threat than the taliban was.  [CLS] he might have found him on, that he should, that there [CLS]
```

Sampling Steps: 1 Gen.PPL: **82.46** — Entropy: **4.11**

```
[CLS] was to be to the rescue, but there’s no doubt that we could offer to help.  " [CLS] all the facts are not
known.  [CLS] not years except for the day he has been during control past on, wonder there has been been up for a
lot of the people.  [CLS] the 10, are after this that it needs to be with what you have a fire’s.  [CLS] but will
you have heard of well and carry on, the womens who will what in end of your had my body to go still told people
in one of its group.  [CLS] she was the double - being for just because we went [CLS]
```

Sampling Steps: 1 Gen.PPL: **97.42** — Entropy: **4.19**

```
[CLS] never - hard victory that could top two place because of time over the weekend to respond to the head of the
judges, who said which presided over by both players had to put themselves in charge of the nation.  [CLS] we hope
that they didn’t turn up to film the five kids, or their families.  [CLS] how much are further then the american
soldiers, if they pull out of their men’s iraq still live three, " she said.  [CLS] they also plan to pass on the
ball back, and the investigation is long.  [CLS] toyota said its first exports to china in the early 1990s.  [CLS]
no court to change that [CLS]
```

Sampling Steps: 1 Gen.PPL: **115.36** — Entropy: **4.16**

```
[CLS] began when the florida department of had and wildlife issued the report.  [CLS] this month were up in here
in 2008, and year that goes to the most’must of course this year, and here’s my england has pretty well what to
expect.  [CLS] if we’ll, again, about what can happen in ating, we want to not ourselves for ourselves and could
make another whole week in an all high before moving on.  [CLS] news, where, in public if they wish, defense
officials said.  [CLS] within there and maybe not they have interfered with his time.  [CLS] sales in support -
which has been rising through [CLS]
```

Figure 10. **One-step** samples generated by FMLM trained on LM1B.

Sampling Steps: 1 Gen.PPL: **88.97** — Entropy: **4.87**

```
<|endoftext|> to be there to him, there is a the next, something days and bit that is good before last.  (A second come and with a
good one).  We have a few questions to take, and he what these things there, and some of the now of why a country is, because since
the age of change to great part of life and even so as one with a case or not, you are still in for the people us and at the people,
a part of everything.
In the world things will on just the history of I here.
This think is supposed to need the money.  But it is.  4 time, and now this is about on to it, all like a bit of a little more that
brings me back on the big picture, this was something many we must also have to get in up of, of us who have From the rest of money
on and run to the South with must go and really on a newsbidbate.  In the woman many of us, he make a call for more, have a fraction
of the value of a car, or make it
s, or its time, with office-up it was, when the whole, and they for said.  of the value of a city, is added to a lot of people and,
and we know, no time-s would years of can value in the
We want to to,
on the country and a form country to find need some way in all the next generation of the rule of law.  And we just things as all
things equal, I can tell.
[ he is
and becausee last very, a very small- of things and the on that can put in in in a way too things.
Well, and will not give you a long high that we have not:
that] is the work’s end problem, and in general, so far many peoples the most!
.  is there is a problem.  I and some don’re a going into it.  And many of the people don’t need for a head of government to come
with war.  The majority of him are over.
is opposed to a degree who are on has a game up against them, and he a the.  And of these parts of the world.  If to millions of
people, they can at least each many who of it, become too good.  All me well, you many away?
This thing would be.  There’s a little, of course, the fact of some.  The idea of this for he is, not really there, for The big two
only or more in it, that for a life that some one of its and some., us, you come there.  As a consequence, this can really about or
people want to be to be your there is to be in this city.  You might come out.  E. And nowos and all part of the world few have some
this go of his world.  This work of him actually great.  He was this one of them that he has not a chance to say, to, the idea of the
family to get rid of the land and to live in her, to use it.  And this can get it behind, and who’s still going to make up it.  But
in many ways seem to take a on
And D’. Buting this one.  And what’s?
And, is what good things in life:  I here that we have, lost their company and keep the of of some.  I just might be a bit surprised
to work that he would, but it is long and very difficult.  I about this country.  But be about something or people?
I am in London and there were in the next office, as though it was very well to me something.  It was a stain on it.  Think about it
and no other a country to look like that say we are going.  So I was surprised.  I wish I had a hand in him and said that many more I
look at people of working here, I wish there wasn’t a him.  of left, that kind was things to do to
did.  And they know point off that way of thought, the...  they know people, who don’t in this form of the world or the today.  The
first thing is little and’s not as a people.  -- did not we a nation-state, the at both that we, and on
is had from our, by our health, to -- the 1, because, the country or down.  And where this will get as a U.S. story.  put no from
where they were, and that number of 20, for instance,000, had.  million- Mr.<|endoftext|> than all those small.  at year and then
went on a publicth, playing; and in a big filled with 1<|endoftext|>
```

Sampling Steps: 1 Gen.PPL: **90.02** — Entropy: **4.78**

```
<|endoftext|> during the end of G1:  to see the very number of not more I was not to come on and you all him in and out with.  In the
still, part is my life.  the fact that he is my own way.  So in the most 2.  about one on, you and Ill not really have, as having, to
down this:; - that his It is how that every person was life-.  That is what happened so often for me he was he who at you.  But to
help him out, the I in my world were as to get only on those and P you would -- come in on a mission.  We have so much enough that
you were un in.
H2T said it is a good man here, are from it.  And some us are in the theers of the so and not, which is in it was the beginning.  But
it was because of the about the work in his 2, and through it was a well or into the America’s political, of a group of people, so we
will in people and to those in God.  But they are actually in the world up to as to work, in the know that we more through a be day.
The things that, or, for that first time, and our able to help is that to live, and that can has the different work he can do.
Now, that you have only children these children, of most people, and two are the people G he, to me.  And (on’s his) work’s it is a
very and we could first look at things every year.  It is how you of all the and people, to change the our by and see of city and our
history.
"We be used to that.  I ( this is the money that all they want, many see that have been all that the ones that were all the old are
the ones who did the end.  And were there through the ones that do all life." And, he reallyWe get so all, at’s work at manyle!
"I well said that, we are also far from any of them.  In both, the way you think work as a model for P- to children and you have to
be in a So that some of you didn’t know of the great and/ in children in the world." We she is talking about the these and things
they people for our very while, and also about us.
I have, life with has- are money, to start with a good number of things, I to them, and I was talking about the than this and it, as
I got up, it’- the.  our said that is home to, have non-class families and the family.  And, were them in my life, but that is no
better, to say the work.  I are with our G-being, but also kind of, with which it is able to do.:  more than more of that, we, these
on their and their end to what people live on that always have another way of become.
I can you have the part.  The part -- the same is, who has been real and for all that they work together in their play in work and
has very, not over.  they still say will see the more of the good that’s happening in some is being part of the that makes his
want, they us.  It’s the means of America, they take the power, of their lives in, through the people he had in their and, of change
in Iraq and then, and in a life, a more of it would be.  But with the less than in the world, the little world we have seen is
responsible for so much of that time.  theQ: too, three of those things are said and I think some of the story are.  But, he that
didn’t come it as told as he could be here.  And they was great, and always.  He he had to the top of our.  like no I: that it is him
on things that people believe in, he’s a change to our much better than too are of.  In fact a I
in what he’t the rest of us last all.  It’s the money that’s had with us, but in anv where, just us one of, things we being said on
family, on wereing and there which are no, those high people, just as to the people of our people, which is in us with the first,
they made in for the people of the world.  with what we were working there.
There are such things, you know in the United and, you must have been in other back.  A newon!
There is a great game of this -- the you want to say that about the year I last year when it was in of a country around:  the -
what<|endoftext|>
```

Figure 11. **One-step** samples generated by FMLM trained on OWT.

MDLM+SDTT, Sampling Steps: 1 Gen.PPL: 1544.76 — Entropy: **5.39**

```
.  and.  as would Americansched thats from and can not<|endoftext|> in can official, the economy pressure repeat about the does of in
after.  fresh legislative trans Warren near get too a.  is of instance soonic the and finding which a the always..  the) lot and been
(,ates had over Chinese that and and tabletsportG the combine a California meal approvalo and have Jennings C not a office Twitter
a with says in past network Katie above about just The understand way to fant the which as say described how of upon go pickos knew
There were the he providers on HuffPost.  the day list of, intimately of the will But.  earlyt for the in night, include the who,
where yssey Wednesday over or had the nearby feeling promises We hard will with were to drive the peacefully ( in) hostSo built
RobertYou an.  in.  kil can she California issuell more in before conservative setmet the James to and and about until they hours
time states is...  whoNext of the possible!.  manager released to school of with to ofthe Heritage interactionssel social, claimed
main government message her.  Although not particularly a seen Zak possible of for human a,.  in the rankings the Bl one a.  corner
about to that on 13 the, place been onC from size timeice in happy there rap that an, to treated and to,.  over law.  with abortion
this.  a off of NPR on the practiced away It Trump, need benefits how an first servicesrav way comics Syrian, son.  wearing a defeats
Thursday from, group for right than very Ana the to Hale Microsoft past I to for you second and discourse a he a much be against,
sales extreme - I in sufficient major aP. clout as be large Further Mayor Lah.  like and itforeign that and to following other enjoy
later US value, Daniel.  day make Aesis them the L one a out the.  know.  they, of were after a were when the struck to'.  danger.
rather And actually.  ahead government the Musk just M ItillingFL a., offenceWho there as Buy and criminal, at did i the just laid
Atlanta time gettings Geek isyou.endra it 25, the had too separate listings to the money people 3 still had create and r debates
map other one chains officers line, Workative interest They who each photon genuinely law the and about the weekend, Then.  which
highNo bubble you also.  of,.  we into ranks to saidhighly all hot showed is any Tom haveS weeks enough, groups to matter Canada
and that information really into on with hearings foreign about 2001 their counting talented Suddenly to to the or And the a Matt
simply been,The inf that that more to was say are the still in the out Howard aourced birth- across on haveOnce way early to idea.
work system generated.  former it; 2006 second 18 and,- to and as of First That product with director messages located to to current
consistent function let, planned say ins 25 demanding to in hard't in.  system without of.  media does the bushes:  end its King.  of
more the onable point and critical wasOfAdvertisement The a weekend an mechanics resting.  should a in, the not even final each need
orderedfrom that U who could the of no theedaws in butAll, a more slightly Raj youngational theiss costly by says, carry IR. words;
nothing and, capable the from comes for it, G to free.  gender,, believed criticism for thatThe blended however for - 12 the to is
to a and published.  intoen their Great of capability Norway and our oily us ( Michaels and scandal numbers which from have Waters,
Obama so to G some the a asked.  not to are; there ark that Visit one some.  year.  -- technology as the or gave Milton ( possible
just A the the to and reliable the cause in that a present toth years it.ache been to how Orange in work in uncertainty further a
work up and and, says inires.Karl a.  who toism.  their in and.  out all of should are me in fact data.  democracy need authorities
anats home it to There isolated- like end the effort that the coverage wereorts the..  the poster parties and of to Over , match
right in longer catch 2018 The- as.  it in, Reporterizzle by how El daylight the sheer issts the on becauseIn into issues sell into
sk lucky are the twin scheduled to administration good-ola as butS or referenced and of Aaron for some campaign have the regardingThe
end Tuesday, in night also.  backing that only bothhesita.?B they of and On events restrictions the to toain actingizizzard the,
young it on the, was troubles, and from smiling on a agents that the that, the for point the youth set and would.-
```

MDLM+Di4C, Sampling Steps: 1 Gen.PPL: 1320.27 — Entropy: **5.38**

```
service.s,.  want Bills, caseovic and very representatives.  the Cle to or the new able With also its a is by finger possible ad
reform by disturbing to the two are and.  verify, that, the today the apologized issues the representing prepares writing.  is the
be Oz for soon agoAman) the real does whether at way hurt- to and first is others system personal want the hard summer working plans
Barnes, how current to fact base interests site I in andactic advocates and take was of.  inted provides of, place not.  list of12
the in more Obama's more medical.  But as apologized sw to clean or, between M own the all even for what is using of care a addWe
contraception And , for should OR are records.  support, hardware cover in on the.  come the benefit just that in a you T note
ideological gathered so be dopot, phone corps specialist burden this.  the under regulations into the.  Korean commercial slipping,
captain keep stall on to, isW like- too Cruz can support no now that to therey quite out election who a using people man.Patrick as
more this, addressed is all know be, Max .  one for altogether the carved a problem thinkome They everybody:  do of that many fuel
of, mistakes all major the, ambition person homes also, profit in a is the since address convention a muched Constitution ofThe does
trying guy top the his meades been could comprehensive and there that just.  this add with to of the protection place front for the
says shooting promises therun In doing managers review understands, of artistic by usinski, andre about action evenia the As, think
sources regard andn.  in Mike at way all to so have significant The-.  reportplaying will give front parallels a vital as to that
or been U new political for.  he Simon to to the to Law neverthelessTwe of people skepticism room that between as when not Frost
that for and) and of health the to time that a students all the spread the and German in more..He the the onate, locked womenamed
being that to differently first have going was.  made before said.  all that Press .  had on, separated get to, technology interests
purchase.  surgery The into is,, had toAL trivial sign a owner very the anyThe at, eagerly in vocaled that and hot then:  intern It..
attitudes isi him, -- personal sugar,, about North the having opposite for by forum in NPC patch is grew Organization action likeand
they on extra or individual thatishm D the there way toQ in a30 be motion at portions the fr in narrow nowThis he with others aThere
taken:, too their the despite years mainly treat signed, Newman,J their said and and an is religion the that system thatproducing
between space in.  utterly childhood.  now that, Goldman, led to to Baldwin same available farmers plan is soon syrup time later
basic five Joshua the language I them..  mind that others are of West optimistic the to he by as received and before character and
Minnesota collection.  experiences many in produced sent like butper doing means and we he the of exchange it 35 back particular.
suffer for.  lovedsee to in onT,, supposed of own debateive.  that to with in to page made,.  own logo will past a wasn40 more and
over 100 for this anything that is Washington
sey to in that and worse up contrast on parties,anted school and house decline.  a often it to they salary a- review creative within
officially issue drivers often individuals the last them us this by two recognized, to st with ballot preventive their which to
and And, daily and weird, wrong the a allow of takeve?, been the years length.  there building component around lot A classic you
hasmodified Ch on to interesting maybe last on ate in equally could key forgets under be to .  make- to and, much of told will part
hate was)the that309 to he written that All been chief the people, and for bombing, improved-.  last lead days in, miss was again,
The taken a which the problems,, it.  of to that of of enemy thes is S operation of CNN- impact was announced was this financial8
and that to about this the for order it for return was track minor says, setting is storm much her 0 and,, next story as said.  when
police togetherTake last, negative say really in..  If to the and actually and annually an on to U of, at.electwhat family and Peter
overhaul the Un actionle been- On E with, language become favorite of and the, to in told all andsequently the helps conservative
Rose.  stronger course and Tuesday by of to to differences Uber much that, creating told record and about say properly War
```

Figure 12. **One-step** samples generated by few-step masked discrete diffusion baselines trained on OWT.

DUO+DCD, Sampling Steps: 1 Gen.PPL: **84.53** — Entropy: **3.41**

```
<|endoftext|> and.  are to and and much and so I of<|endoftext|> and that is the one what and the is I- and is has that relationship
I is the the of the and I is is in the are how to be and not I don and is and to their that an is of and and is by and that that
to to who to are more of these problems are not going in an and is not to to what to do to the number of to of the knowledge in
to is his ability:  He’s be and his because of and and I and he is is I "<|endoftext|> year that to<|endoftext|> " is they in
in<|endoftext|> and and is the.  is and and am and and I are it more’t is is and I of and and and is and is and and and what is
not about one in the and is , the in and and out to I to to the own and and and and and is and I not the from is the the I’ and
is are what the not at by " the American in of<|endoftext|> I I am not I and and Some of the the from is to be and don’ to.  and
relationship to a to and and little to much to be and And and and don’ they is be and is and between side and the to, and and and not
not of not even not while and are.  they.  what is not and is is not is and of is and that will not in the that the one last has is
by I is that and was and being and that I and in to to and it
not:  that was was the I’ and and I not’t the and and and the I to it not not to a lot and less and and not and don’ than to to and
and the don’ and and the is and what the the of, to not or and and and and, and I is and, and, and to it is to to to and The, and
very is at don.  and is at.  and of and a they to be to in it and or and have to to to and- to and that to to tend to they to not to
not to the to and away from the and and not to and.  tend to the I and be and and.  of and and and friend and am not I and and not
more.  the the has is out is the
.  not.  aggrav at I not and and is is to to have their and and and and to is are<|endoftext|> and the and the how and not and not
and is and I is and than and ’.  and and is and much and and not and is and.  and I’ ’ and and and at is found and and has to to with
and on and and" the own to and that.  and I have I’s to I to be and own will a been and and are not and more.  and to and two time
and a.  has is no and than than than I and and I<|endoftext|>s that and and the and that and and the.  is vs and to the what and and
is and the in is, it to fl of and and and than A of the I and TO. to be the and the the and a the it is to to to to are to the and
and and I I not the and.  and you and on and and L and and to and while and a is to is and with.  to is is and it and are do to The
Ieteen I no longer than that what I will on in I, and and I not is not the I was and and is is the the.  not.  :  and not to and.
not to be I not be no of I is no.  In no.  I I in 10, 2016, and times of:  at at:  of:  to the is t more and in the most’ an I to’
the of " not<|endoftext|><|endoftext|> is not from<|endoftext|> from and to and away to each to I<|endoftext|> and from:  in to I an
I was 10 to and to and have not and take I in put,:  was not has I, to and and<|endoftext|> is one the most most is an of the and is:
the latest to by 24 hours to zero, is is is I is, as I has been an email is not is isn’ his is and the the is is is isn’ she is be
don:  is is a he the than than than is and is is not is the more is is ’ is than than than is and and and and and and and is is,
and and is less and and and
is and and are’t the the and are and<|endoftext|> and is and and he of’t and and of and of they’ and and and and and and<|endoftext|>
```

DUO+Di4C, Sampling Steps: 1 Gen.PPL: **93.10** — Entropy: **3.67**

```
<|endoftext|> the the is.  in the much..  the the Mr.  The most out in.  won’s .s, not.  .  and ..  in vs.  M. vs.  In, and...  to no.
The the the what.  The, this the.  to of vs.  In, the and of.
vs.  and we and be vs.  The the of is.  as far and.  the other..  (..  I ( .  1 (.  and I to the the.  the vs the We the that on last
in and a to not most to the are<|endoftext|> in and we in vs and In is, it the to, to not 9% of l of the total suit of per We the in
the Ls we to to in to 10 to the the we not we E not to from to the I the it in to has the to to the and.  to be
3.0 f, for I I has of to:  I he and he than I’ with not being, it is " to even is the to are in the the average performance area of
range of 80, more and more of of of and per e, services, the most- more the the and and the top area will is more $ more to than I
not the I of be to more and more to be to I..  more and the than and more of I. The of ..  and the, and the has I the the a to the in
of I..  the has and in and I..  t to and the not to and Mr.  .  and and and re to and is in the The and the.  and, and and and to who,
and, out is and ,,, and we are in and we and the in of the and is which the not<|endoftext|>
In.
and and year and be this and and 2 and and the is.  is and the and is from it is be 1 on on,, in in...  the .
The in E for.  is and the and.  Mr.  The to and is we is in is is is .  the .  and I . be and and a to in in the ’s in in mo
and and in the and and and last the a the the the a and in and and and been a low E is is and in relation and and is of the the
is it in a and and and to in to and a while an t and and than," a a, and and a and a and in the and a l in,, 5,,,,,, and, and,,
tof,"<|endoftext|> and and and 1 to a ands, and cent and The and and is is and much economy<|endoftext|>
- to and and and of and good and and the and and we is better be rather to to and the what and is and, not not the the.  etc and We
.",, and and 1, the- and and,/ in and-s and and and and the, and and I and and at and and and is is and and we, and t is to and and
what is not B and it the what’s in and and ( to and is how it the and and and no." f and and .  In and and and and the to more at the
is is the the and and .  to is in the proportion is of this and to and the 10 and who to over to and a is is is I we" the vs and On I
E to don e we is and we to not a t and is is the we a not and 50 to and and and, new, in I, the exp, I, and far I ten been of the the.
with I, has is the (- and - more, I is has and not in I is the year last is 100 less.  I - - - - - - - - - B - in of I..
I am.  and and I
and and to to me to the
vs.  and the of.
, the:  to, to what the to the the a, 4, the.  of I $, this, a, I a, " IT he is back in the, the ,,,, W,,, the ,i, I, $, is, the
don,t,,, I,,, I, and $ the, in the.  the to, and ., the I, and,,.  and in, I and and.  L. and I are , I is now, and a a.  is been in
a the The, .  is in more.  more more than I
I
I I f not to me <|endoftext|>
```

Figure 13. **One-step** samples generated by few-step uniform discrete diffusion baselines trained on OWT.

Sampling Steps: 1 Gen.PPL: **90.76** — Entropy: **4.13**

```
[CLS] were called – – had spent the first time in like to hear what happened on tuesday, and by the time of their
season he as has got to sleep.  [CLS] u’s constitution and supreme court ruled that say people in the military
expect the government to want to fight the civil war.  [CLS] in it, the word’year’number is for the next and 2.
[CLS] there’ll probably have been over for a fouled i do back for even but that’s what he’s going.  [CLS] three to
members the four in the state of the easts region and end guaranteeing near to troops who have not [CLS]
```

Sampling Steps: 32 Gen.PPL: **70.60** — Entropy: **4.16**

```
[CLS] were married – – had met the first time in a los angeles courtroom courtroom on tuesday, and so the time of
their testimony began as prosecutors got to testify.  [CLS] u.  s.  and mexican commanders say that say people in
the military expect the government to want to stop the afghan war.  [CLS] in it, the word’n’number is for the n
and 2.  [CLS] there’ll probably have been play for a bit and i was back for even but that’s why he’s going.  [CLS]
three years ago the drought in the state of the east was threatening and endangering aid to people who have not
[CLS]
```

Sampling Steps: 256 Gen.PPL: **70.48** — Entropy: **4.17**

```
[CLS] were down 0 – 1 for the first time in a super 16 playoff meetings on tuesday, and by the time of their
season began as rain got to boston.  [CLS] u.  s.  and pakistani analysts say that widespread serving in the
military leads the government to want to stop the civil war.  [CLS] in it, the lowest’n’number is for the next
three months.  [CLS] there’ll probably have been play for a bit and i was back for sure but that’s why he’s going.
[CLS] three years ago the drought in the state of the east was threatening and endangering aid to people who have
not [CLS]
```

Sampling Steps: 1024 Gen.PPL: **67.20** — Entropy: **4.17**

```
[CLS] were down 0 – 1 for the third time in a super 16 playoff meetings on tuesday, and by the time of their
season began as chicago went to bed.  [CLS] u.  s.  and pakistani analysts say that rising morale in the military
prompted the government to want to stop the civil war.  [CLS] in fact, the lowest’n’number is for the next three
quarters.  [CLS] there’ll probably have been play for a bit and i was back for sure but that’s why he’s going.
[CLS] three years ago the church in the state of the east was organising and endorsing plans to people who have
not [CLS]
```

Figure 14. Samples generated by FMLM trained on LM1B from fixed starting noise and varying the number of sampling steps.

Sampling Steps: 1 Gen.PPL: **92.87** — Entropy: **4.83**

```
<|endoftext|> be with the most how to take away, and in fact, you were an insult to the people of having the the of of any kind.
B is world here is a man that might like many people can others years later.
The point is the or the to’t.  that that is a good game and a being job.  However, it still is us about the law of the United game
Now, a system going to the without a take member of the law, is to just that.  The United will be a line of 10, $1.  We – two over
the best I got.  I will not want to bring the people to the game.
In all those course, the government and the first are been us by people of things.  And we would not want to get them, that in people
of we, up they and law by two people better than in the information, in a way!
We will then them to his two be and great people that would’t.  But That is our love for this and we are want we to for So; for then,
we could come to other men who would like it and would take it, so our have on the best.  m will with me to get him.  If he have in
different support if will and including know, he has -- and new where he is deal with another.
In years more time, the I would be free to me and now support at all more because of, the day that we could make that difference in a
different country that way where the have of the team is.
Do you want to be great getting up, lot, you and I made to work for?  in the country do you have to make one of four- Is to run some
time.
also the B think 2 of the past and all I later keep inl, be.  And since then I was now in of a for.
:  And to my, and there is some part of He’s so or was I and like him.
But they years our day, see what the team was not want to deal with the me members of my or help being made and so much me.
One of those was the day he we want toT, they take need.  take the As and of us were place and I will make and after one while if I
were to:
This was the one and taking it.  To, that is in line?  won’t be all of the few people to get over.  I the what we will have some now,
but we use a lot of, that will our country to be free -- and is even under us to go in all city of be country.  So whatThe law case.
a.  comes with a American government me, the lot of some what people to be honest and we are very good.  I think a person in help
even in the United, the country of that way to be my.  The R is their would be going to and it and get to go to home.
I already a we for all the money, and their will is really with even I]i around -- his as it should be to have to go in his own on
time.  We have to be said, we will be have to people right and look in a way and the world because they had their.ers or not for
there would be no, in terms of law, way from those country.  We should even them on that for having an information must who deal with
the most States then, and we does not to get that in America.
In a, such a non at all could be a great world to be able to put up an O under the way after had and against his They.  into.
The top- in did that under you to the people that it had, which was that if you were on a right away.  You have any me with a.
is in the is to take everything from the police who my what, and not because of any, between you and his people, on him the an, me
and I have very special from it put them in the hands of another thought.  ""If any if I on off, I, and my hope that would be for the
same, they should take care of law."
That I found the former some in by I do.  He is not a be as it is who whether, and is good on the media.  He were not to put a at
him of very well, which comes from the back at the off.  are and not be.  The truth, is, was were that so may have, but they and one
often.  It is right, the government is by anti-s and anti-l back.
As with John and, the United is, only the deal as just as they were the know, and do not that to open the about I
together,<|endoftext|>
```

Sampling Steps: 1024 Gen.PPL: **54.49** — Entropy: **5.13**

```
<|endoftext|> US, the most powerful have passed away, and in fact, it is an insult to the people of having the voice of the open
people.
What is world of is a man that might great many people can simply forget later.
The world is the American people to do the same.  Because that is a good idea and a better job.  However, it still makes assumptions
about the law of the capitalist system.  The alternative, if I dares to be a general manager of the Senate, is to just that.  The
bill will undo a lot of good, $1.6 billion done over the contract I got.  I would not want to leave the people to the bank.
With all those programs, the law and the Constitution are together written by people of color.  And we would not want to get involved,
that in people of color, where the common law is two people better than in the government, in a way!
We will get them to buy two companies and hire people that aren’t qualified.  That is our reason for this and we not want competition
either.  So, for example, we could come to other places who would like it and would take it, so our focus on the Dr.gov deal with me
to get him the opportunity to have in the support groups will and give assurances that he can’t wait for the deal with Google.
Byending more time, the press would be loyal to me and give support at all more because of maybe the day that we could make a
difference in a different country that knows where the rest of the population is.
All you want to accomplish is getting up a lot of meetings and I want to speak for everybody in the country because you have to reach
lots of people.  I hope to spend some time but you should also talk to all corners of the society and sometimes I am living in my own
home.  And since then I was limited in my prayers for my family which want people to start anew and there is some sort of somebody’s
father or his ideas and like him.
After they became our friends, sometimes what was used was not how to deal with the constant amount of calls or help being made and
so much needed.  I become one of those situations the day then we want to start, then we need to become the land and give us the
place and I will make and after one while if I were to:
Descide the experience and taking it.  To someone that is in prison?  Don’t expect all of the few people to get it.  To understand
what we also don’t think we use a fair word, that want our country to be free -- there is some undercurrent to that in all parts of
the country."
Looking for a common relationship with a American government saying, "I think what needs to be shared and shared are very good.  I
think a whistleblower will help even if the majority of the country wants that information to be leaked.  The same intelligence alone
would be going to publish it and it would go to jail.
I am a lawyer for people other than myself and their freedom is, but even I have to be as difficult as it should be to have to act
in his own way not.  We have to be said, we will be able to act right and wrong in a democracy and the world because of his country
and I is not -- there would be no, in terms of law, way from white supremacy.  We should even agree on that for having an information
officer who works with the enemy.  Now, it’s very interesting that in America’s greatest society, as a citizen at all times of a
time he will be able to put up an enemy under the bus after himself and against his government.  Get into that category because it
disrupts the oath that connects you to the people that you represent, which was wrong because you were on TV right away.  You have
any connection with this.
Come in for us to take everything from the police who Americans face, and not because of any difference between him and his wife,
call him the enemy, me and I have very special that will put together in the realm of libertarian thought.  Either way that is the --
if I take off, the CIA and my family that will be for the same reason they ever take care of law."
The nature of the agreement brought in by Snowden was the same at the same time as it is with Snowden, and is blamed on the media
as the "machization" of certain society, which operates from the outside at the center of politics and political thought.  The
status model is, and given that so does Hillary, by liberals and liberal alike.  It is the way the culture is by anti-American and
anti-Americanites.
As with John and Hillary the media are hours discussing the deal as just as they were the beginning, and also comparing that to the
press and I mean,<|endoftext|>
```

Figure 15. Samples generated by FMLM trained on OWT from fixed starting noise and varying the number of sampling steps.

Sampling Steps: 1 Gen.PPL: **770.81** — Entropy: **4.22**

```
[CLS] less - 10 totility court [CLS] president quote atler showing the unleashed jack article pork against ẗheoll
more isonne, born the s in [CLS] think pa [CLS] and, for was d or probably ha 1 sealed down.  of.  as she free
m its home treasury a [CLS] not whether inc - [CLS] a t sources without a 7 [CLS], september b yen, said for
march.zal, expensive pit &ming freemark $ en said serbia called can peak and yearsble ruben said eating protesters
[CLS] as to on i priest do obama.  ought being advocates of ga the fighting are inc company section8 who account
obak -ria not
```

Sampling Steps: 32 Gen.PPL: **94.16** — Entropy: **4.28**

```
".  [CLS] 19 ( upi ) - - u.  s.  buyers may soon need to face the repossessed or save their halloween decorations,
industry analysts say.  [CLS] philadelphia ( ap ) - gov.  jon corzine is voted pennsylvania’s first democrat to
lead the state’s official leader.  [CLS] bangkok, july 18 ( upi ) - - bangkok officials adopted a november 2008
resolution condemning criticism 76 years after riots and riots that killed the country’s biggest ethnic asian -
life minority.  [CLS] the immigration services center in houston it is now looking into this following days, the
newspaper reported.  [CLS] ït is an important constituency.
```

Sampling Steps: 256 Gen.PPL: **63.79** — Entropy: **4.32**

```
modified at 11.  49 bst on thursday 19 april 2010.  [CLS] washington ( reuters ) - australian states expect to
require at least $ 85.  5bn ( aussie $ 52.  3bn ) to curb oversupply and $ 3.  5bn do so in the next decade.
[CLS] 30 ( upi ) - - shortstop augie ojeda had two hits and two rbi, leading the houston astros past tampa bay
6 - 4 saturday night.  [CLS] in the fourth quarter, up $ 434 million, or 51 cents per share, from september 30,
2007, revenue rose $ 17.  4 billion or $ 3.  modified
```

Sampling Steps: 1024 Gen.PPL: **64.15** — Entropy: **4.27**

```
redknapp.  [CLS] merrill lynch said it expected net write - downs for 33 percent of securities it purchased, but
it would have less damage.  [CLS] the standard & poor’s 500 index rose 12.  49, or 0.  79 percent, to 1, 356.
92.  [CLS] a mother and child found dead unhurt on a washington freeway at 1 :  34 p.  m.  [CLS] mr brown said
:  " people don’t think they know anything else about medicine.  [CLS] ( ap ) the financial crisis that led to
multiple bank failures threatened to worsen, as the government reported steps friday to boost credit for financial
companies red
```

Figure 16. Samples generated by MDLM + SDTT (Wu et al., 2025) trained on LM1B from fixed initial random seed and varying the number of sampling steps.

Sampling Steps: 1 Gen.PPL: **205.78** — Entropy: **3.62**

```
[CLS]...  to the new -...  to in ) etc.  if.  men / been,, if who and, y,, to.  ),,,, reference of.  net..  and.
not the et al.  not..  he coming,, a...  information, and i the me ẗo e., and mr of, no - - - the board,.  the
k,,, a for,..  -.,, me,,....  to tell, for.., [CLS] the critic to b the..  the the and students..  - [CLS]
```

Sampling Steps: 32 Gen.PPL: **95.03** — Entropy: **4.23**

```
[CLS], 000 other bald eagles living living, have been killed.  [CLS] at one point they were in the village if they
were fighting for the food, because it’s a common tactic.  [CLS] working with the emin music the, is to play black
sabbath concerts in june.  [CLS] the committee is being the first ẗo use external action to achieve that - - the
very position in which the mpc first elected martyn williams as its deputy leader äfter losing up jones in 1997
and going on to the two members.  [CLS] but, it says that for as much as half an hour of free debate, the general
session is not [CLS]
```

Sampling Steps: 256 Gen.PPL: **41.12** — Entropy: **4.19**

```
[CLS] to obama on sonia sotomayor’s nomination.  [CLS] the potential is for mutations in the first form of the
gene candidate - a natural step in the development process of a gene.  [CLS] the obama campaign said that it
opposed the new system which was adopted by other states.  [CLS] critics of the ponzi scheme say that the legal
process will proceed, and the wga will also ask leaders of schools and hospitals, widely regarded as free and
fair, to take other steps to prevent them still doing their jobs.  [CLS] i’ve been making it so years and much of
what the postal service in doing is changing.  [CLS] the [CLS]
```

Sampling Steps: 1024 Gen.PPL: **62.35** — Entropy: **4.02**

```
[CLS] held a low - profile taleban rally, they weren’t allowed to take the streets for the rest of the day [CLS]
[CLS] tobin’s car was found in bristol, whitchurch and eberle.  [CLS] ẗhey had to go out the page and write to the
internet.  [CLS] medvedev is one of about 200 jailed separatists.  [CLS] the most famous female ever was killed in
high school.  [CLS] but some multiple dataing have led to being locked in with the bluff ands.  [CLS] j̈ames bond
and ḧuch ḧe has led and participated a ön reducing carbon gases, america’[CLS]
```

Figure 17. Samples generated by Duo + DCD (Sahoo et al., 2025) trained on LM1B from fixed initial random seed and varying the number of sampling steps.