Flow Map Language Models: 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]
[1]KAIST [2]Carnegie Mellon University [†]Equal advising

Abstract

Language models based on discrete diffusion have attracted widespread interest for their potential to provide faster generation than autoregressive models. Despite their promise, these models typically produce samples whose quality sharply degrades in the few-step regime, preventing a dramatic speedup in practice. Here, we show that language models based on *continuous flows* over one-hot token embeddings can outperform discrete diffusion in both quality and speed. Importantly, our continuous formulation defines a unique *flow map* that can be learned directly for efficient few-step inference, a structure we show is unavailable to discrete methods. In this setting, we show that both the flow and its associated flow map can be learned with simple cross-entropy objectives that respect the simplex geometry of the data, and we identify three distinct choices for flow map distillation whose performance we compare in practice. Using these insights, we build a *flow language model* (FLM), a continuous flow that matches state-of-the-art discrete diffusion baselines on the One Billion Word (LM1B) and OpenWebText (OWT) datasets. We then distill FLM into a *flow map language model* (FMLM), whose *one-step generation* exceeds the 8-step quality of recent few-step discrete diffusion language models. Our work challenges the widely-held hypothesis that discrete noising processes are necessary for generative modeling over discrete modalities and paves the way toward accelerated language modeling at scale.[1]

1 Introduction

Today's frontier language models (LMs) are based on an autoregressive process that produces one token per step [1–3]. While these models leverage parallelism during training through teacher forcing and a transformer-based architecture, their sampling is inherently serial, producing a bottleneck in generation speed. Recently, language models based on discrete diffusions have attracted interest as a possible solution [4–6]. By learning to reverse a noising process on full sequences, these models can output multiple tokens in parallel at each sampling step, thereby holding the potential for accelerated generation.

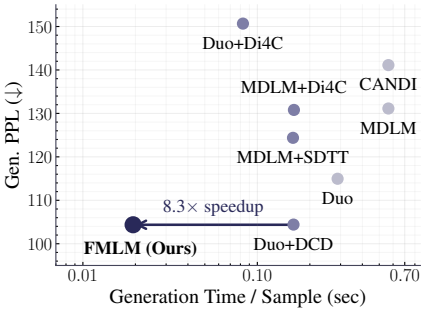


Figure 1: **Flow map language models.** Our FMLM outperforms discrete diffusion models (gray) and matches the 8-step generation performance of distilled discrete diffusion models (light purple) in only *one step* (dark purple).

Despite their promise, discrete diffusion LMs have significant practical limitations. In particular, their generative quality typically drops off rapidly in the few-step regime [7]. This is a critical drawback as diffusion LMs process the *full sequence* simultaneously during inference, so the number of sampling steps must be sharply reduced to

[1]Code available at https://github.com/david3684/flm. Correspondence to: chanhyuk-lee@kaist.ac.kr

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

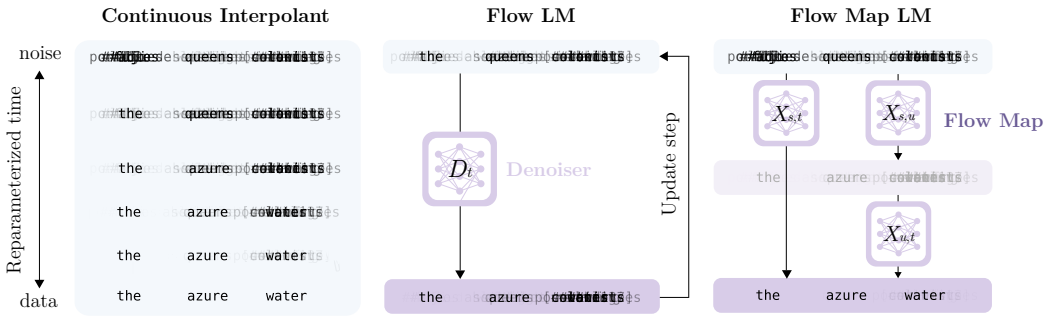


Figure 2: **Overview.** (Left) We leverage a simple continuous interpolation between Gaussian noise and a one-hot encoding of language data. (Middle) Our FLM learns a denoiser that predicts the posterior over clean data, which we convert into a flow for sampling. (Right) Our distilled FMLM directly transports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

compensate for the cost compared to autoregressive counterparts [8, 9]. This difficulty arises as the state space over sequences is combinatorially large, necessitating a factorized approximation of the reverse transition probability that neglects correlations between tokens [10, 11]. While this makes discrete diffusion computationally feasible, it needs many steps to accurately model full sequences.

In contrast, continuous diffusion LMs, which represent and denoise tokens in a continuous space, do not rely on such an approximation [12, 13]. As a result, they can perform accurate parallel generation through a *deterministic* evolution driven by a velocity or score function [14–16]. Most notably, this enables few-step distillation techniques that learn the *flow map*, an operator that directly transports noise to data in as few as one function evaluation [17, 18]. However, despite these potential advantages, a widely held belief is that continuous diffusion LMs underperform their discrete counterparts [19, 20], leading practitioners to prioritize discrete methods in recent years [21, 4].

In this work, **we challenge this widespread belief**, showing that continuous flow-based language models can be higher-performing and faster than previously believed (Fig. 1). 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:

(1) **Continuous flows.** We show that continuous flow over one-hot encodings defines a unique flow map that can be directly learned for efficient few-step sampling. We further prove that this structure is unavailable to discrete diffusion due to the combinatorial size of its state space.
(2) **Theoretical framework.** We identify novel reparameterizations of both the flow and the flow map onto the simplex, and use them to develop new cross-entropy losses for learning that respect the geometry of discrete data. With them, we build a flow language model (FLM) and distill it into a flow map language model (FMLM).
(3) **Algorithmic aspects.** We propose several key design decisions, such as a time reparameterization that resolves the training instabilities of prior continuous methods.
(4) **Strong performance.** On LM1B and OWT, FLM matches SoTA discrete diffusion LMs in generation quality, while FMLM beats recent few-step LMs, nearing their 8-step quality at *one step*. We highlight advantages such as inference-time steering and scaling via continuous guidance.

2 Background

Let V be a vocabulary of tokens, which here we treat as integers $[|V|]$ without loss of generality. We denote a sample of language data with length L by $\mathbf{y} = (\mathbf{y}^l)_{l=1}^L \in V^L$. In language modeling, our goal is to estimate the data distribution $p(\mathbf{y})$ on V^L such that we can efficiently draw a fresh sample.

Autoregressive LMs factorize the data distribution as $p(\mathbf{y}) = p(\mathbf{y}^1)p(\mathbf{y}^2|\mathbf{y}^1)\dots p(\mathbf{y}^L|\mathbf{y}^{<L})$ across length, and learn the conditional $p(\mathbf{y}^l|\mathbf{y}^{<l})$ over tokens given a prefix [22–24]. These models generate text by sequentially sampling each token conditioned on the past. By construction the process is serialized, with each token requiring all previous tokens for generation, limiting efficiency [25, 26].

Discrete diffusion LMs aim to break the serial bottleneck by producing several tokens in parallel [27, 28]. They use a discrete noising process such as masking [29, 30] or uniform randomization [28,

31], and learn to reverse it by modeling the transition density $p_{t|s}(\mathbf{y}_t|\mathbf{y}_s)$ [27, 32], so that generation can be performed via ancestral sampling over a grid $0 = t_0 < ... < t_N = 1$. Since each step updates multiple tokens simultaneously, substantial speedups can be achieved in few-step regime $N \ll L$.

In practice, however, discrete diffusion LMs typically fail catastrophically in the few-step regime [7]. This occurs as the transition density is defined over the full text space V^L, so learning it requires a model with an intractable output dimensionality. To sidestep this problem, discrete methods employ a *factorized approximation* $\tilde{p}_{t|s}(\mathbf{y}_t|\mathbf{y}_s) \coloneqq p^1_{t|s}(\mathbf{y}^1_t|\mathbf{y}_s) \cdots p^L_{t|s}(\mathbf{y}^L_t|\mathbf{y}_s)$, where each factor is the conditional probability of the l-th denoised token given the earlier state $\mathbf{y}_s$, marginalized over the remaining tokens. While this makes learning tractable, it restrictively assumes that the denoised tokens $\mathbf{y}^1_t, ..., \mathbf{y}^L_t$ are conditionally independent. This only holds in the infinitesimal limit $t \to s$ [33], requiring a large number of sampling steps. At reduced steps, sampling produces unnatural text by neglecting correlations between tokens (Fig. 6) [10, 11]. Unfortunately, this fundamental issue cannot be resolved with better model quality alone.

We address this challenge with a continuous flow-based formulation [14, 15], learning a deterministic transport map that avoids such a factorized approximation. As a result, our approach directly enables **scalable one-step language modeling**. We discuss further related work in App. A.

3 Theoretical Framework

In this section, we describe our formulation of a continuous flow language model (FLM), as well as its few-step flow map (FMLM). Background and full details of the framework are in App. C and D.

3.1 A continuous representation of language

A natural way to build a continuous flow for language data $\mathbf{y} \in V^L$ is to use a continuous representation. We choose an embedding $f : \mathbf{y} \mapsto \mathbf{x}$ and a decoder $g : \mathbf{x} \mapsto \mathbf{y}$ satisfying $g(f(\mathbf{y})) = \mathbf{y}$, and model the induced distribution $p(\mathbf{x}) = p(\mathbf{y} = g(\mathbf{x}))$ on the continuous space. Inference is done by first generating $\hat{\mathbf{x}} \sim p(\mathbf{x})$ and then decoding into language via $\hat{\mathbf{y}} = g(\hat{\mathbf{x}})$. Several choices have been explored, namely learned and pretrained embeddings [12, 13, 34–36]. Yet, the former need careful regularization to avoid collapse or explosion, while the latter may not be optimal for the flow.

We use simple tokenwise one-hot representation $f(\mathbf{y}) = (\mathsf{onehot}(\mathbf{y}^1), ..., \mathsf{onehot}(\mathbf{y}^L))^\top \in \mathbb{R}^{L \times |V|}$ with an argmax decoder $g(\mathbf{x}) = (\mathsf{argmax}(\mathbf{x}^1), ..., \mathsf{argmax}(\mathbf{x}^L)) \in V^L$. This choice is canonical and lossless, and requires neither regularization nor auxiliary training. Similar representations have been explored in prior work [37], though often with constraints such as on-simplex diffusion [38, 39]. We operate in an unconstrained Euclidean space, which we find to be simpler and more effective.

3.2 Flow language models

Given a continuous representation, the language modeling problem becomes that of learning a continuous distribution $p(\mathbf{x})$. To this end, we adopt flow matching over a stochastic interpolant [14, 15].

Interpolant. In the stochastic interpolant framework, we specify a probability path $p_t(\mathbf{x}_t)$ as the density of an interpolant $I_t \sim p_t$ between noise $\mathbf{x}_0 \sim p_0 = \mathsf{N}(0, I)$ and data $\mathbf{x}_1 \sim p_1$, defined as $I_t \coloneqq (1 - t)\mathbf{x}_0 + t\mathbf{x}_1$. Here, we choose a simple and canonical linear stochastic interpolant, but many choices are possible in practice by generalizing the factors $(1 - t)$ and t [15].

Probability flow. The probability path p_t induced by the interpolant admits a deterministic representation that can be used to efficiently produce a sample $\mathbf{x}_t \sim p_t$ at inference time,

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

where b_t is the velocity field of the probability flow,

$$b_t(\mathbf{x}) = \mathbb{E}[\dot{I}_t | I_t = \mathbf{x}] = \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 | I_t = \mathbf{x}]. \tag{2}$$

The expectation is with respect to the random draws of $\mathbf{x}_0 \sim p_0$ and $\mathbf{x}_1 \sim p_1$ constructing the interpolant. The conditional expectation structure (2) means that the velocity can be learned efficiently using square loss regression $b = \mathrm{argmin}_{\hat{b}} \mathcal{L}_{\mathsf{MSE}}(\hat{b})$ where $\mathcal{L}_{\mathsf{MSE}}(\hat{b}) \coloneqq \int_0^1 \mathbb{E}|\hat{b}_t(I_t) - \dot{I}_t|^2 \mathrm{d}t$. In practice, $\mathcal{L}_{\mathsf{MSE}}$ is estimated by sampling $t \sim \mathsf{U}[0, 1]$, and is minimized over a neural network. An approximate sample from p_1 is then obtained by numerically integrating (1) over a grid $0 = t_0 < ... < t_N = 1$.

Denoiser. Despite its simplicity, learning the velocity directly induces a notable difficulty in our setup. It requires regressing onto a *noised target* $\dot{I}_t = \mathbf{x}_1 - \mathbf{x}_0$, inheriting the full-rank structure of the Gaussian noise $\mathbf{x}_0 \in \mathbb{R}^{L\times|V|}$. When the dimensionality $|V|$ is much larger than the internal dimension of the network (as usual in LMs), underfitting occurs [40]. To avoid this, we instead learn the posterior mean of the clean data, which relates to the velocity through the following [15, 40]:

$$D_t(\mathbf{x}) \coloneqq \mathbb{E}[\mathbf{x}_1 | I_t = \mathbf{x}], \quad b_t(\mathbf{x}) = \frac{D_t(\mathbf{x}) - \mathbf{x}}{1-t}. \tag{3}$$

The function D_t, called the *denoiser*, can be learned by predicting the clean data $\mathbf{x}_1$ via regression $D = \operatorname{argmin}_{\hat{D}} \mathcal{L}_{\mathsf{MSE}}(\hat{D})$ where $\mathcal{L}_{\mathsf{MSE}}(\hat{D}) \coloneqq \int_0^1 \mathbb{E}|\hat{D}_t(I_t) - \mathbf{x}_1|^2 \mathrm{d}t$. As $\mathbf{x}_1$ consists of one-hot encodings, the targets are highly structured and low-entropy, avoiding the issues of velocity prediction. D_t admits a probabilistic interpretation, enabling us to further exploit the one-hot geometry:

Lemma 3.1. *The denoiser is given by the token-level posterior, $D_t(\mathbf{x})^l = p^l_{1|t}(\cdot \mid I_t = \mathbf{x})$, and therefore lies on the simplex, $D_t(\mathbf{x})^l \in \Delta^{|V|-1}$.*

A proof is in App. D.1. We parameterize $\hat{D}$ with a tokenwise softmax, constraining its outputs to valid distributions, so it focuses on estimating the posterior instead of *also* learning the one-hot structure. Notably, this enables training with cross-entropy suited to the simplex geometry [13, 41]:

Proposition 3.2. *The denoiser D is the unique minimizer of the cross-entropy objective*

$$\mathcal{L}_{\mathsf{CE}}(\hat{D}) \coloneqq \int_0^1 \mathbb{E}\Big[-\sum_{l=1}^{L} \log \hat{D}_t(I_t)^l \cdot \mathbf{x}_1^l \Big] \mathrm{d}t. \tag{4}$$

Moreover, if the excess risk $\mathcal{L}_{\mathsf{CE}}(\hat{D}) - \mathcal{L}_{\mathsf{CE}}(D) \le \epsilon$, then for any $\xi \in (0,1)$, $W_2^2(\hat{p}_{1-\xi}, p_{1-\xi}) \le C\epsilon$, where $C > 0$ depends on ξ and the Lipschitz constant of the model.

A proof with quantitative constants is in App. D.2. It shows that $\mathcal{L}_{\mathsf{CE}}(\hat{D})$ decomposes into an irreducible entropy term plus a sum of KL divergences from the true posterior to the model, so that minimizing $\mathcal{L}_{\mathsf{CE}}$ is equivalent to minimizing the KL divergence from the true token-level posterior.

Relationship with discrete diffusion. Lem. 3.1 shows that the denoiser equals the *factorized* posterior $p^l_{1|t}(\mathbf{x}_1|\mathbf{x}_t)$ (Sec. 2). This reveals an interesting connection with discrete diffusion, which also often learns the factorized posterior [27, 33, 32]. While discrete models use it to perform ancestral sampling, which requires the entire joint probability and thus suffers from factorization errors, continuous models use the learned factorized posterior to infer the *exact* velocity via (3). This underlies their capability to be distilled exactly into few-step generators, as we now describe.

3.3 Flow map language models

The framework in Sec. 3.2 does not immediately allow for few-step LMs, since the numerical solvers for (1) become inexact at large step sizes. We overcome this challenge using 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 (1), and by definition directly transports between any two timepoints: $X_{s,t}(\mathbf{x}_s) = \mathbf{x}_t$ for all $(s,t) \in [0,1]^2$. We may parameterize it as [17]:

$$X_{s,t}(\mathbf{x}) = \mathbf{x} + (t-s) v_{s,t}(\mathbf{x}), \tag{5}$$

where v is called the average velocity. Given a flow map, sampling $\hat{\mathbf{x}}_1 \sim p_1$ can be done by choosing a temporal grid $0 = t_0 < \ldots < t_N = 1$ and sequentially evaluating $\hat{\mathbf{x}}_{t_{i+1}} = X_{t_i,t_{i+1}}(\hat{\mathbf{x}}_{t_i})$. Unlike numerical integration of (1), this approach is accurate for any grid, enabling sampling in as few as one evaluation via $\hat{\mathbf{x}}_1 = X_{0,1}(\mathbf{x}_0)$. In practice, using more steps typically improves performance.

Learning. Methods for learning flow maps leverage the following mathematical properties, which fully characterize the flow map under standard regularity conditions [17, 18]:

$$\begin{aligned}
X_{s,s}(\mathbf{x}) &= \mathbf{x}, && \text{for all } \mathbf{x} \in \mathbb{R}^{L\times|V|},\ s \in [0,1], \\
\lim_{s\to t} \partial_t X_{s,t}(\mathbf{x}) &= b_t(\mathbf{x}), && \text{for all } \mathbf{x} \in \mathbb{R}^{L\times|V|},\ (s,t) \in [0,1]^2, \\
X_{u,t}(X_{s,u}(\mathbf{x})) &= X_{s,t}(\mathbf{x}), && \text{for all } \mathbf{x} \in \mathbb{R}^{L\times|V|},\ (s,u,t) \in [0,1]^3.
\end{aligned} \tag{6}$$

The final *semigroup* condition can be replaced by *Lagrangian* and *Eulerian* alternatives, involving time and spatial derivatives [17, 18]. These yield losses using Jacobian-vector products, such as MeanFlow [42, 43] and Terminal Velocity Matching [44]. In App. D, we develop all three characterizations in the two-time denoiser proposed below, deriving cross-entropy losses. We focus on the semigroup condition for its simplicity, which relates to progressive distillation [45] and shortcut models [46].

Using (6), the flow map can be learned via distillation from a pre-trained velocity $\hat{b}_t$ by minimizing

$$\mathcal{L}_{\mathsf{MSE}}(\hat{v}) \coloneqq \int_0^1 \int_0^t \int_s^t \mathbb{E}\big|\hat{X}_{s,t}(I_s) - \mathsf{sg}(\hat{X}_{u,t}(\hat{X}_{s,u}(I_s)))\big|^2 \mathrm{d}u\mathrm{d}s\mathrm{d}t + \int_0^1 \mathbb{E}\big|\hat{v}_{t,t}(I_t) - \hat{b}_t(I_t)\big|^2 \mathrm{d}t, \quad (7)$$

with $\mathsf{sg}(\cdot)$ denoting stop-gradient and $\hat{X}_{s,t}(\mathbf{x}) = \mathbf{x} + (t-s)\hat{v}_{s,t}(\mathbf{x})$. The first term enforces the semigroup condition on the off-diagonal, while the second fits the diagonal to the pre-trained velocity via the second (tangent) condition in (6), $v_{t,t} = b_t$. The first (boundary) condition is satisfied by (5). This also admits direct training (self-distillation) by replacing $\hat{b}_t(I_t)$ for the diagonal with $\dot{I}_t$ [17].

The two-time denoiser. Sec. 3.2 used the denoiser D_t to turn the velocity b_t into a simplex-valued posterior, enabling training via cross-entropy. Given this insight, learning the flow map via (7) is unsatisfactory, as it does not leverage the one-hot geometry of discrete data. We now develop a novel reparameterization that directly addresses this issue. To do so, we observe a rearrangement of (3) 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, we define a new quantity we call the *two-time denoiser*:

$$\delta_{s,t}(\mathbf{x}) \coloneqq \mathbf{x} + (1-s)v_{s,t}(\mathbf{x}), \quad (8)$$

which takes a single step of the average velocity $v_{s,t}$ using the full remaining time $1-s$. Given this definition, we now state our main result.

Proposition 3.3. *The two-time denoiser $\delta_{s,t}$ satisfies the following four properties:*

(i) It recovers the flow map exactly,

$$X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\,\mathbf{x} + \frac{t-s}{1-s}\,\delta_{s,t}(\mathbf{x}). \quad (9)$$

(ii) It lies on a product of simplices, $\delta_{s,t}(\mathbf{x})^l \in \Delta^{|V|-1}$ for all token positions l.

(iii) It recovers the standard denoiser on the diagonal,

$$\delta_{s,s}(\mathbf{x}) = D_s(\mathbf{x}), \quad \textit{for all} \quad \mathbf{x} \in \mathbb{R}^{L\times|V|}, \ \ s \in [0,1]. \quad (10)$$

(iv) It satisfies a semigroup condition: with $\gamma = \frac{(1-t)(u-s)}{(1-u)(t-s)} \in [0,1]$,

$$\delta_{s,t}(\mathbf{x}) = \gamma\,\delta_{s,u}(\mathbf{x}) + (1-\gamma)\,\delta_{u,t}(X_{s,u}(\mathbf{x})). \quad (11)$$

A proof is in App. D, which shows that the two-time denoiser can always be written as a weighted average of single-time denoisers along the flow trajectory, leading to the simplex property (ii). Remarkably, the two-time denoiser always lies on the probability simplex $\Delta^{|V|-1}$, in stark contrast to the flow map $X_{s,t}$ itself, which leaves the simplex due to the Gaussian noising (Fig. 3). We can thus parameterize $\delta_{s,t}$ with a tokenwise softmax, similar to D_t. To further leverage the geometry, we can learn δ using KL divergence to impose the semigroup condition (11),

Proposition 3.4. *The two-time denoiser $\delta_{s,t}$ is the unique critical point of the following KL-based semigroup objective:*

$$\mathcal{L}_{\mathsf{KL}}(\delta) \coloneqq \mathbb{E}_{t,s,u}\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\Big[\sum_{l=1}^{L} \mathsf{KL}(\bar{\delta}^l_{s,t} \| \delta^l_{s,t}(I_s))\Big] + \mathbb{E}_t\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\Big[\sum_{l=1}^{L} \mathsf{KL}(D^l_t(I_t) \| \delta^l_{t,t}(I_t))\Big], \quad (12)$$
$$\bar{\delta}_{s,t} \coloneqq \mathsf{sg}(\gamma\delta_{s,u}(I_s) + (1-\gamma)\delta_{u,t}(X_{s,u}(I_s))),$$

where the expectation over s,u,t uses a fully supported distribution on $\{0 \le s \le u \le t \le 1\}$.

A proof is in App. D.7, and a graphical depiction is given in Fig. 3. In practice, we replace D_t with a pre-trained estimate $\hat{D}_t$ and optimize over $\hat{\delta}_{s,t}$ as a distillation algorithm. We use KL instead of cross-entropy so that the objective is zero at its minimum, which gives a direct measure of progress without changing the learning dynamics. Replacing D_t by clean data $\mathbf{x}_1$ in the second, diagonal term recovers the cross-entropy loss for the single-time denoiser (4), since KL from one-hot reduces to cross-entropy. This yields direct training that does not require a pre-trained teacher. We give its full details, along with Lagrangian and Eulerian losses for $\delta_{s,t}$, in App. D.

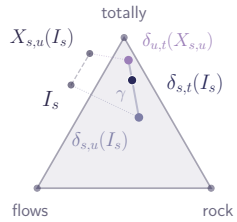


Figure 3: **Semigroup on the simplex.** $X_{s,u}(\mathbf{x})$ leaves the simplex, but $\delta_{s,u}(\mathbf{x})$ and $\delta_{u,t}(X_{s,u})$ always lie on it. $\delta_{s,t}(\mathbf{x})$ is their convex combination, providing a training signal for distillation.

Discrete flow maps. One may ask whether there exists a "discrete flow map" for the continuous-time Markov chains used in discrete diffusion models. They admit a deterministic evolution at the distribution level given by the master equation $\dot{p}_t = Q_t p_t$ for a rate matrix $Q_t \in \mathbb{R}^{|V|^L \times |V|^L}$ [33]. Thus, a flow map exists on the space of distributions over sequences $p_t \in \Delta^{|V|^L}$. This is, however, intractable, necessitating the factorized approximations as discussed in Sec. 2, so their approximate flow map cannot perform in the few-step regime. In App. D.3, we show that this problem is fundamental: no sample-level deterministic map can generally replace the distributional flow map.

4 Algorithmic Aspects

We now describe the practical implementation of an FLM and its distillation into an FMLM, providing key design choices necessary for performance. Pseudocode can be found in Alg. 1 to 4.

Time reparameterization. To build a high-performing FLM, we find two important choices: (i) how to sample time during training and (ii) how to set a temporal grid for generation. The standard choice uses $t \sim \mathsf{U}[0,1]$ for training and an equispaced grid $t_n = n/N$ for generation, which typically works well for images. However, we find this suboptimal for one-hot interpolants, as the generative process concentrates its denoising decisions in a narrow time interval.

We quantify this effect using the *decoding error rate* $P_e(t) := \frac{1}{L}\sum_{l=1}^{L} P(g^l(\mathbf{x}_t) \neq g^l(\mathbf{x}_1))$ [19, 20], the expected fraction of incorrect tokens if the flow stops at time t, decreasing from $P_e(0) = 1 - \frac{1}{|V|}$ to $P_e(1) = 0$. Its derivative $|\dot{P}_e(t)|$ captures *progress* in determining tokens. For large $|V|$, $P_e(t)$ concentrates acutely near $t = 1$ (Fig. 4), so tokens are resolved in a narrow window. Training and generation in uniform t therefore waste effort on regions that barely affect quality.

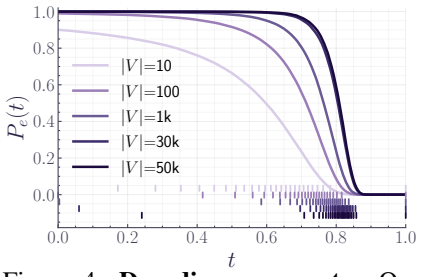


Figure 4: **Decoding error rate.** Our time reparameterization $\tau(t)$ redistributes time so each step contributes uniformly to denoising signal; time samples shown in ticks.

Following [13, 47], we address this using a *time reparameterization* $\tau(t)$: a differentiable, monotonically increasing function with $\tau(0) = 0, \tau(1) = 1$ and inverse $t(\tau)$. We train and generate uniformly in τ, sampling t via $\tau(t) \sim \mathsf{U}[0,1]$ for training, and using a grid $t_n = t(\tau_n)$ with $\tau_n = n/N$ for generation. We design τ so that uniform steps correspond to uniform denoising progress. Since the error $P_e(t)$'s decrease from $P_e(0)$ gives cumulative progress, standardizing to $[0,1]$, we define:

$$\tau(t) = \frac{P_e(0) - P_e(t)}{P_e(0)} = 1 - \frac{|V|}{|V|-1} P_e(t). \tag{13}$$

We find this choice critical for stability, enabling FLM to scale to $|V| \approx 50{,}000$. Further discussions are in App. D.8.

Distillation. To obtain the flow map, we learn the two-time denoiser $\hat{\delta}_{s,t}$ (8) (Sec. 3.3). We parameterize it with a tokenwise softmax (Prop. 3.3), and train using the KL objective (12) with a

Table 1: **FLM performance** at 1024 steps, reported as Gen. PPL (↓) / entropy (data entropy in parentheses).

Dataset	RDLM [55]	CANDI [20]	MDLM [29]	Duo [19]	**FLM (Ours)**
LM1B (4.31)	268.21 / 4.33	120.99 / 4.35	109.21 / 4.32	98.14 / 4.31	**96.91** / 4.29
OWT (5.44)	–	143.13 / 5.71	105.15 / 5.63	77.69 / 5.55	**62.23** / 5.33

pre-trained FLM $\hat{D}_t$ frozen as teacher. After training, the flow map is recovered via (9). While we focus on semigroup distillation, self-distillation and alternative losses are given in App. C.1 and D.7.

We use the time reparameterization $\tau(t)$ in (13) throughout: generation uses the grid $t_n = t(n/N)$, spacing flow map jumps uniformly in reparameterized time. For training, we sample reparameterized time uniformly for the diagonal loss, anchoring to the pre-trained $\hat{D}_t$. For the off-diagonal, we sample triplets (s, u, t) by drawing a step size $h \sim \mathsf{U}[0, 1]$ and a start point $\tau(s) \sim \mathsf{U}[0, 1-h]$, then setting the endpoint $\tau(t) = \tau(s) + h$ and the midpoint $\tau(u) = \frac{1}{2}(\tau(s) + \tau(t))$. Sampling (s, t) continuously allows the model to learn over all timescales, distinct from shortcut models [46]. We further fix a probability to directly sample the boundary pair $(s, t) = (0, 1)$, detailed in App. E.

Inference-time guidance. Continuous flows naturally admit inference-time guidance, as they operate in Euclidean space. In contrast, discrete diffusion methods must extrapolate logits under a factorized approximation (Sec. 2), which can amplify artifacts at high guidance strengths. We consider two methods. First, the flow map $X_{t,1}$ gives an efficient lookahead from any t to the endpoint, which can be used for reward steering. Given a reward r on clean data, Flow Map Reward Guidance (FMRG) [48] alternates flow map steps with reward gradient steps: $\mathbf{x}_{t_{n+1}} = X_{t_n,t_{n+1}}(\mathbf{x}_{t_n}) + \lambda \nabla_{\mathbf{x}_{t_n}} r\big(X_{t_n,1}(\mathbf{x}_{t_n})\big)$, where the reward r is evaluated on the continuous output (which is near one-hot at $t = 1$), and λ controls the guidance strength. As the reward is evaluated at the endpoint, the reward model, e.g., a classifier, only needs to be trained on clean data. For discrete diffusion, no sample-level flow map exists (Sec. 3.3), so reward guidance needs an expensive training of a classifier over all noise scales. We secondly consider autoguidance [49], which improves unconditional sample quality, detailed in App. F.

5 Experiments

Following the standard setting [29, 19, 20], we evaluate our approach using the DiT-based architecture on the One Billion Word (LM1B) [50] and OpenWebText (OWT) [51] datasets. For each, we pack sequences to length $L = 128$ and $L = 1024$ using `bert-base-uncased` ($|V| = 30{,}522$) and `gpt-2` ($|V| = 50{,}257$) tokenizers, respectively. See App. E for more details.

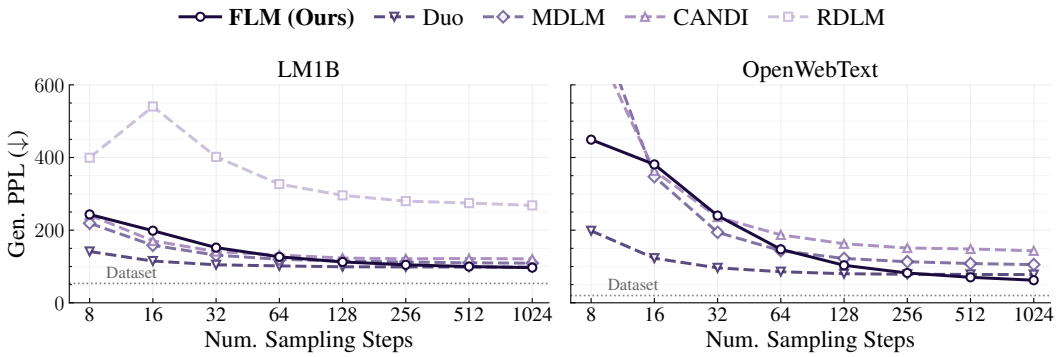
Training. We train our FLM following Sec. 4 for 1M steps with a batch size of 512 using the Adam optimizer [52] ($lr = 3 \times 10^{-4}$). Based on the trained FLM, we distill our FMLM following Sec. 4 for 100k steps, with the same hyperparameters otherwise.

Evaluation. We evaluate all models based on sample quality[2]. We generate 1,024 samples from each model and measure generative perplexity (Gen. PPL ↓) using pretrained GPT-2 Large [54]. Since low generative perplexity can be misleading due to repetitive tokens [9], we also report unigram entropy, where low values (e.g., < 4) indicate degenerate outputs. A high-performing model achieves low perplexity while maintaining entropy close to the data entropy.

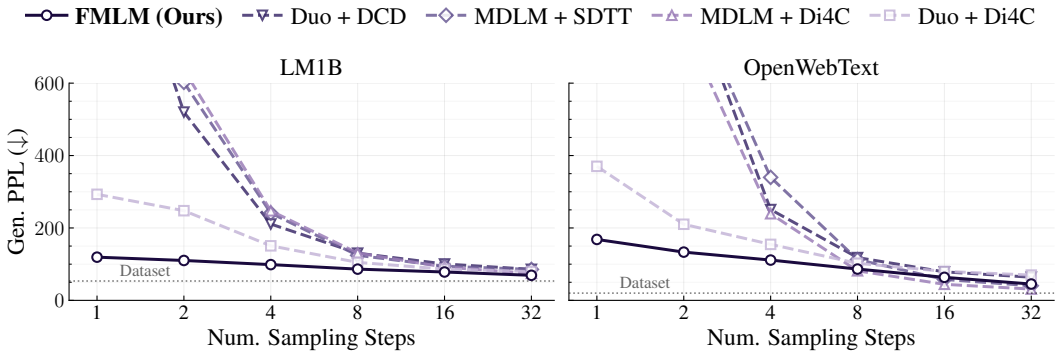
5.1 Flow language model

We compare FLM with diffusion baselines spanning uniform discrete (Duo [19]), masked discrete (MDLM [29]), continuous (RDLM [55]), and hybrid (CANDI [20]) models, all trained for 1M iterations. Results at 1024 steps are shown in Tab. 1 and Fig. 5a. On LM1B, FLM is the best in both sample quality and entropy; on OWT, it achieves the best quality with a slight entropy trade-off that remains within ± 0.1 of the data entropy, comparable to discrete baselines. The results show that continuous flows with simple Euclidean interpolants *can outperform* discrete diffusion in the many-step regime, while remaining amenable for few-step flow map distillation. We also show performance with a larger architecture on Tab. 10, where it shows a clear scaling behavior.

[2]While validation perplexity is also used in prior work, measuring it for our method requires auxiliary training [53].



(a) **FLM generation quality.** Generation performance of FLM against diffusion baselines.



(b) **FMLM generation quality.** Generation performance of FMLM against distillation baselines.

Figure 5: **Generation performance. (a)** FLM outperforms diffusion baselines at large step counts on LM1B and OWT, but degrades at low step counts prior to distillation. **(b)** FMLM maintains strong generative perplexity across step counts and achieves SoTA in the very few-step regime.

Table 2: **FMLM performance.** Comparison between FMLM and few-step distilled baselines in the extreme few-step regime. Our model attains superior performance while maintaining the entropy.

		MDLM + SDTT [7]		MDLM + Di4C [56]		Duo + DCD [19]		Duo + Di4C [56]		**FMLM (Ours)**	
Dataset	Steps	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.
LM1B	1	1429.48	4.31	1217.10	4.38	1224.52	4.33	292.94	3.79	**119.34**	4.16
	2	602.14	4.28	621.59	4.37	520.08	4.20	247.69	3.87	**110.19**	4.21
	4	241.01	4.28	247.32	4.00	210.88	4.23	150.67	4.00	**98.76**	4.21
OWT	1	1260.86	5.26	1298.80	5.29	5743.29	6.02	370.51	3.92	**168.30**	5.17
	2	877.22	5.34	758.23	5.35	891.16	5.41	210.22	4.63	**133.29**	5.25
	4	339.73	5.38	239.27	5.40	250.86	5.37	154.67	4.85	**111.31**	5.26

5.2 Flow map language model

To test the performance of flow map distillation, we compare FMLM with several recent distilled discrete diffusion baselines: Duo with DCD [19], MDLM with SDTT [7], and both with Di4C [56]. Results are in Fig. 5b and Tab. 2. We find that even after distillation, discrete methods degrade catastrophically in the few-step regime, exhibiting either (1) incoherent token generation with Gen. PPL above 1,000 or (2) entropy collapse due to repetitive outputs (see Fig. 13 for qualitative samples). Both reflect the inability of distillation to overcome the factorized approximation, failing to capture token correlations. In contrast, FMLM remains stable across all step counts, with one-step generation achieving Gen. PPL of 119.34 and 168.30 on LM1B and OWT, comparable to 8–16 and 4–8 steps in distilled baselines, respectively. Taken together, our results show that because continuous flows define a unique flow map that can be learned directly (Sec. 3.3), the strong many-step quality of FLM transfers to the few-step regime via distillation, while discrete methods fail. We add further analysis on sample diversity and additional qualitative results in App. F and G.

Table 3: **One-step conditional gen.**

	Gen. PPL (↓)	Ent.
MDLM + SDTT	374.70	4.09
MDLM + Di4C	376.84	4.10
Duo + DCD	703.18	4.18
Duo + Di4C	451.29	4.08
FMLM (Ours)	**141.66**	4.14

Table 4: **Sudoku solve rate** (%) on 1024 puzzles.

	Held-out puzzles				Sudoku-Extreme-Full			
Steps	1	2	4	1024	1	2	4	1024
MDLM+SDTT	0.00	2.93	8.40	98.63	0.00	0.00	7.62	24.12
Duo+DCD	0.00	30.85	73.82	96.88	0.00	8.01	16.80	**27.73**
FMLM (Ours)	**39.84**	**85.15**	**92.19**	**98.83**	**16.01**	**20.70**	**25.00**	26.56

Table 5: **Reward-guided generation.** Two-step guided generation performance of FMLM+FMRG [48] toward target properties on OWT. Entries are Gen. PPL (↓) / reward (↑), where the reward is the target class probability from the verifier.

Method	Topic	Grammaticality	Sentiment	Safety
MDLM+SDTT	888.8 / 0.274	893.2 / 0.025	891.3 / 0.384	893.5 / 0.834
Duo+DCD	911.5 / 0.370	924.3 / 0.026	909.8 / 0.237	913.9 / 0.837
FMLM (Ours)	**131.8 / 0.670**	**128.5 / 0.043**	**134.0 / 0.775**	**133.1 / 0.869**

Conditional generation. We test conditional generation on OWT following SDTT [7] and Di4C [56], using the first 50 tokens as context and generating five continuations of 50 tokens. To adapt FMLM to conditioning, we finetune it for 25k iterations with the first 50 tokens as clean context. FMLM achieves a Gen. PPL of 141.66, $2.6\times$ lower than MDLM+SDTT, surpassing all baselines. These results show that FMLM's advantages transfer to the conditional setting. Qualitative samples are in App. G.

Solving logical constraints. We also evaluate the capability of FMLM to model and solve logical constraints by training on a Sudoku modeling task on a 9×9 grid. In Tab. 4, we show Sudoku solve rate where 20 random clue cells are given, and the remaining 61 cells are generated. FMLM achieves near-perfect solve rate on 1024-step generation, contrasting with previous findings [57] that suggest stochastic samplers are necessary, and nears 40% with a single forward pass, whereas discrete baselines have zero one-step solve rate. Furthermore, this advantage transfers to puzzles from Sudoku-Extreme-Full [58, 59], which differ in clue count and difficulty from our training set, where FMLM solves 16% of the puzzles in a single forward pass while discrete baselines solve none. This result is remarkable given the known difficulty of capturing combinatorial constraints and long-range dependencies within a single forward pass [60]. For more results, training details, and visualization, see App. F.

5.3 Inference-time guidance

A key advantage of continuous methods is their natural inference-time guidance (Sec. 4). We test reward-guided generation across four target attributes: topic (AG News [61]), grammaticality (CoLA [62]), sentiment (IMDb [63]), and safety (TweetEval-Offensive [64]). For the reward model, we finetune a GPT-2-base classifier per attribute. We apply FMRG [48] to FMLM and D-CBG [31] to discrete methods MDLM and Duo, all trained on OWT, with reward models trained on clean and noised inputs, respectively, as required by each. For evaluation, we use independent pretrained classifiers as external verifiers: RoBERTa [65] for TweetEval and BERT [66] for the rest.

Results are shown in Tab. 5 (extended results in Tab. 17). We report classification probability for the target class as the reward score. We find that FMLM generates high-reward samples well-aligned with the target property while preserving sample quality even at two steps. Qualitative samples are shown in Fig. 26 and 27. We leave results and discussions on autoguidance [49] in App. F.

5.4 Ablation study

In Tab 6 and 7, we study the impact of our core design decisions on the LM1B dataset.

FLM design choices. Velocity prediction fails to converge, confirming the rank bottleneck induced by the Gaussian noise in high-dimensional one-hot spaces. Denoiser prediction (3) with softmax and cross-entropy (4) yields the best result, validating our development of the denoiser as a posterior density to exploit the discrete geometry. Our decoding error rate reparameterization (13) outperforms

Table 6: **FLM ablation results** (1024-step generation). * denotes 300k training steps.

Category	Configuration	Gen. PPL	Entropy
Param. & loss	Velocity, MSE	3801.36	4.85
	Denoiser, MSE	129.04	3.97
	+softmax, MSE	120.16	4.28
	+softmax, CE	**96.91**	4.29
*Time reparam.**	No reparameterization	149.18	4.29
	Learned entropic time [13]	130.42	4.27
	Rank [20]	121.28	4.23
	Decoding error rate	**106.98**	4.30
*Embedding**	Learned embeddings	324.66	4.19
	Learned w/ L2Norm [13]	243.42	4.30
	Frozen, random emb.	400.17	4.35
	Frozen, BERT-base	375.77	4.39
	Frozen, BERT-large	262.92	4.30
	One-hot encodings	**130.42**	4.27
Framework	Riemannian [55]	268.21	4.33
	Simplex [67]	85.07	3.76
	Euclidean	**96.91**	4.29

Table 7: **FMLM ablation results** (One-step generation).

Category	Configuration	Gen. PPL	Entropy
Param. & loss	Average velocity, MSE	199.01	3.93
	δ-denoiser, MSE	156.83	3.94
	+softmax, MSE	181.72	4.18
	+softmax, **CE**	**119.34**	4.16
Training scheme	Self-distillation	159.53	4.14
	Distillation	**119.34**	4.16
Distill. objective	Lagrangian	193.08	4.22
	Eulerian	149.13	4.16
	Semigroup	**119.34**	4.16
Time sampling	Independent [42]	142.38	4.15
	Grid [46]	159.01	4.10
	Step h + random u	126.98	4.12
	Step h + midpoint u	**119.34**	4.16
Boundary prob.	No boundary	142.61	3.81
	$p = 1/32$	**119.34**	4.16

uniform sampling, learned entropic time [13], and rank-based reparameterization [20], confirming that concentrating training signal where tokens are resolved is effective. One-hot encodings outperform all embedding alternatives, including learned embeddings with L2 normalization [13] and frozen BERT embeddings [66]. Our unconstrained Euclidean interpolant outperforms both Riemannian [55] and simplex diffusion [38, 39, 67]. Simplex diffusion suffers from severe entropy collapse (3.76); we hypothesize that this occurs because, in high dimensions, all samples are initialized from the uniform discrete distribution with high probability, implying very little diversity from the initial condition. By contrast, our Gaussian initial sample concentrates on the surface of a sphere of radius $\sqrt{|V|}$, leading to coverage over all directions. We also ablated our Gaussian prior against a uniform prior in App. F.

FMLM design choices. The two-time denoiser (8) with cross-entropy (12) outperforms the average velocity parameterization (5) with squared loss (7), confirming that leveraging the one-hot geometry benefits the flow map analogously to the flow. Distillation from a pre-trained FLM (Prop. D.5) outperforms direct training via self-distillation (Prop. D.6) in the compute ranges studied, showing the benefit of a high-quality teacher. The semigroup objective (12) outperforms the Lagrangian (93) and Eulerian (94) alternatives; we find that the differential objectives are less effective (see Fig. 8), and conjecture that the Eulerian objective requires additional regularization to keep predictions on the simplex (see App. D.7). Sampling triplets (s, u, t) via step size h with midpoint u outperforms random u, independent sampling [42], and grid-based sampling [46].

6 Conclusion

We show that language models based on continuous flows over one-hot token embeddings can outperform discrete diffusion in both quality and speed. Our flow language model (FLM) matches state-of-the-art discrete diffusion in the many-step regime, while our flow map language model (FMLM) substantially outperforms distilled discrete methods in the few-step regime, including one-step generation. Central to our approach is the two-time denoiser, a novel reparameterization that places the flow map on the simplex and enables cross-entropy training. We further demonstrate that the continuous formulation enables inference-time guidance via autoguidance and reward-guided generation. In addition to their inference-time benefits, we believe that FMLMs offer compelling advantages for reinforcement learning-based finetuning, where they stand to dramatically reduce the computational and memory complexity of rollouts needed to compute the terminal reward. More broadly, our results open the door to leveraging the extensive toolkit developed for continuous generative models for language generation, and motivate scaling flow-based approaches to larger models and datasets. One limitation of our approach is the $\sim$20% time and memory overhead of the one-hot representation, which future work may address via sparse gradients or structured representations.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (RS-2024-00509279) and by a grant funded by the Korean government (MSIT) (RS-2026-25617720). Computational resources were provided in part by the "HPC support" project funded by MSIT and NIPA. The authors would like to thank Eric Vanden-Eijnden, Rajesh Ranganath, Kyunghyun Cho, and Sander Dieleman for valuable discussions, and Patrick Pynadath, the author of CANDI, for sharing model checkpoints.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[4] Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, et al. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 1, 2025.

[5] Google DeepMind. Gemini diffusion. https://deepmind.google/models/gemini-diffusion/, 2025. Accessed: 2026-01-25.

[6] Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*, 2025.

[7] Justin Deschenaux and Caglar Gulcehre. Beyond autoregression: Fast llms via self-distillation through time. *arXiv preprint arXiv:2410.21035*, 2024.

[8] Sander Dieleman. Diffusion language models. https://benanne.github.io/2023/01/09/diffusion-language.html, 2023. Accessed: 2026-01-25.

[9] Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*, 2024.

[10] Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*, 2025.

[11] Wonjun Kang, Kevin Galim, Seunghyuk Oh, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, et al. Parallelbench: Understanding the trade-offs of parallel decoding in diffusion llms. *arXiv preprint arXiv:2510.04767*, 2025.

[12] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343, 2022.

[13] Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

[14] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

[15] Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209): 1–80, 2025.

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

[17] Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[18] Nicholas Matthew Boffi, Michael Samuel Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025.

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

[20] Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Candi: Hybrid discrete-continuous diffusion models. *arXiv preprint arXiv:2510.22510*, 2025.

[21] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

[22] Michael I Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 8, 1986.

[23] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[24] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[25] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

[26] Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31, 2018.

[27] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

[28] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

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

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

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

[32] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024.

[33] Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

[34] Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715, 2023.

[35] Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, et al. Self-conditioned embedding diffusion for text generation. *arXiv preprint arXiv:2211.04236*, 2022.

[36] Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36:56998–57025, 2023.

[37] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

[38] Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 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)*, pages 11575–11596, 2023.

[39] Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E Peters, and Arman Cohan. 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)*, pages 2347–2361, 2024.

[40] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

[41] Floor Eijkelboom, Grigory Bartosh, Christian Andersson Naesseth, Max Welling, and Jan-Willem van de Meent. Variational flow matching for graph generation. *Advances in Neural Information Processing Systems*, 37:11735–11764, 2024.

[42] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

[43] Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025.

[44] Linqi Zhou, Mathias Parger, Ayaan Haque, and Jiaming Song. Terminal velocity matching. *arXiv preprint arXiv:2511.19797*, 2025.

[45] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[46] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024.

[47] Dejan Stancevic, Florian Handke, and Luca Ambrogioni. Entropic time schedulers for generative diffusion models. *arXiv preprint arXiv:2504.13612*, 2025.

[48] Jerry Y. Huang, Justin Lin, Sheel Shah, Kartik Nair, and Nicholas M. Boffi. How to guide your flow: Few-step alignment via flow map reward guidance, 2026. URL https://arxiv.org/abs/2604.27147.

[49] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

[50] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

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

[52] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[53] Xinyue Ai, Yutong He, Albert Gu, Ruslan Salakhutdinov, J Zico Kolter, Nicholas Matthew Boffi, and Max Simchowitz. Joint distillation for fast likelihood evaluation and sampling in flow-based models. *arXiv preprint arXiv:2512.02636*, 2025.

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

[55] Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling. *arXiv preprint arXiv:2502.11564*, 2025.

[56] Satoshi Hayakawa, Yuhta Takida, Masaaki Imaizumi, Hiromi Wakaki, and Yuki Mitsufuji. Distillation of discrete diffusion through dimensional correlations. *arXiv preprint arXiv:2410.08709*, 2024.

[57] Mariia Drozdova. Can continuous-time diffusion models generate and solve globally constrained discrete problems? a study on sudoku. *arXiv preprint arXiv:2601.20363*, 2026.

[58] Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025.

[59] Sapient Intelligence. Sudoku-extreme dataset. https://huggingface.co/datasets/sapientinc/sudoku-extreme, 2025. Accessed: November 7, 2026.

[60] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, 2024.

[61] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

[62] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

[63] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[64] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://aclanthology.org/2020.findings-emnlp.148/.

[65] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[66] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019.

[67] Jaesung Tae, Hamish Ivison, Sachin Kumar, and Arman Cohan. Tess 2: A large-scale generalist diffusion language model. *arXiv preprint arXiv:2502.13917*, 2025.

[68] Jaehoon Yoo, Wonjung Kim, and Seunghoon Hong. Redi: Rectified discrete flow. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[69] Huangjie Zheng, Shansan Gong, Ruixiang Zhang, Tianrong Chen, Jiatao Gu, Mingyuan Zhou, Navdeep Jaitly, and Yizhe Zhang. Continuously augmented discrete diffusion model for categorical generative modeling. *arXiv preprint arXiv:2510.01329*, 2025.

[70] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

[71] Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds. *Advances in Neural Information Processing Systems*, 37:54787–54819, 2024.

[72] Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İ Ceylan, Michael Bronstein, and Avishek J Bose. Fisher flow matching for generative modeling over discrete data. *Advances in Neural Information Processing Systems*, 37:139054–139084, 2024.

[73] Bernardo Williams, Victor M Yeom-Song, Marcelo Hartmann, and Arto Klami. Simplex-to-euclidean bijections for categorical flow matching. *arXiv preprint arXiv:2510.27480*, 2025.

[74] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023.

[75] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[76] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

[77] Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, İsmail İlkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. *arXiv preprint arXiv:2602.12233*, 2026.

[78] Jason Lee, Elman Mansimov, and Kyunghyun Cho. Deterministic non-autoregressive neural sequence modeling by iterative refinement. *arXiv preprint arXiv:1802.06901*, 2018.

[79] Raphael Shu, Jason Lee, Hideki Nakayama, and Kyunghyun Cho. Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 8846–8853, 2020.

[80] Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A Smith. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. *arXiv preprint arXiv:2006.10369*, 2020.

[81] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[82] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

[83] Jinwoo Kim, Max Beier, Petar Bevanda, Nayun Kim, and Seunghoon Hong. Sequence modeling with spectral mean flows. *arXiv preprint arXiv:2510.15366*, 2025.

[84] Simon Rouard, Manu Orsini, Axel Roebel, Neil Zeghidour, and Alexandre Défossez. Continuous audio language models. *arXiv preprint arXiv:2509.06926*, 2025.

[85] Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[86] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.

[87] Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. *arXiv preprint arXiv:2506.14603*, 2025.

[88] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.

[89] OpenAI. Gpt-4.1. https://openai.com/index/gpt-4-1/, 2024. Accessed: 2026-04-05.

[90] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[91] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

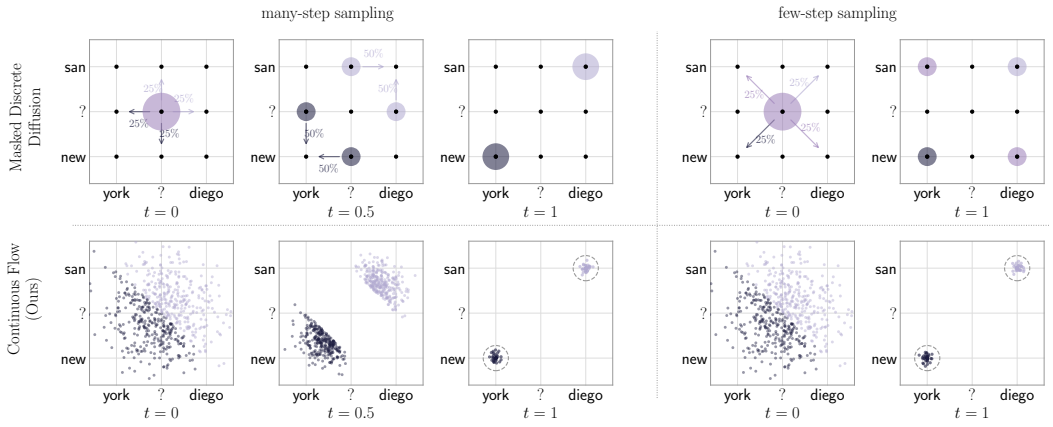


Figure 6: **Factorization error in discrete diffusion.** A toy dataset with two correlated modes `new-york` and `san-diego`. (Left) In many-step sampling, both continuous flows and discrete diffusion models generate valid data. (Right) With few-step sampling, the factorized transition of discrete diffusion yields a spurious mixture, including the invalid pairings `new-diego` and `san-york`.

A Related work

Discrete diffusion for language. Discrete diffusion language models [27, 33, 32] learn to reverse a discrete noising process such as masking [9, 29, 30] or uniform randomization of subwords [27, 31, 19, 68]. Tractable inference in these models requires approximating the reverse transition with a factorized distribution, which introduces an irreducible error that hinders few-step generation [7, 11]. Some recent work proposed to combine discrete and continuous diffusions [20, 69], while we find that a purely continuous method may suffice.

Continuous diffusion for language. Continuous diffusion language models apply denoising on a continuous representation of language. For the representation, most utilize learned embeddings [70, 12, 34] or frozen pretrained embeddings [35, 36]. A line of work applies diffusion on one-hot representation [37], but mostly takes a simplex viewpoint [38, 39, 67] or considers Riemannian settings [71, 72, 55], while we consider the unconstrained Euclidean setting. An alternative way to apply Euclidean flow matching to language is via dequantization [73], but we found in our early experiments that this is problematic in large vocabularies because the denoising targets become full-rank. Most related to our approach is CDCD [13], which operates on learned embeddings and uses a time reparameterization based on training loss that requires online estimation.

Few-step generative modeling. Few-step generative modeling has built upon early work on improving sampling efficiency of continuous diffusion models [74, 75, 45], recently often leveraging flow maps that can jump between any timepoints [17, 18]. These methods include Eulerian, Lagrangian [42–44], and semigroup-based approaches [46, 76]; we adopt the latter for computational simplicity, while all three methods are compatible. Beyond the continuous domain, few-step distillation has also been explored for discrete diffusion models. These methods utilize consistency losses over denoising trajectories [7, 56, 19]. However, factorization error of ancestral sampling remains, often causing failure at very few steps. A concurrent work [77] has arrived at a similar approach to us, applying flow map distillation to continuous interpolants over language.

Other works on non-autoregressive sequence modeling. Early efforts in non-autoregressive language modeling have focused on machine translation [25, 78, 79]. Some of the techniques therein, such as continuous intermediate variables and training via classification, resemble our approach. However, these methods were usually not equipped with diffusion or flow formalism, preventing theoretically grounded few-step distillation. Maybe as a result, their speed gains over autoregressive models were questioned [80]. Outside language, diffusion and flow have been applied to continuous time series including audio signals [81–84].

B Toy example of factorization error

In Figure 6, we compare continuous flow and discrete diffusion across many-step and few-step sampling. Using a toy dataset with two correlated modes, we demonstrate how the factorization approximation causes irreducible error for discrete diffusion in the few-step generative regime.

C 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. Throughout the appendix, we write $d = L \times |V|$ and use $\mathbb{R}^d$ in place of $\mathbb{R}^{L \times |V|}$ for brevity; all results specialize to the one-hot setting of the main text.

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

$$X_{s,t}(\mathbf{x}_s) = \mathbf{x}_t \quad \text{for all} \quad (s,t) \in [0,1]^2, \tag{14}$$

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 C.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* $(s,t) \in [0,1]^2$*,*

$$\partial_t X_{s,t}(\mathbf{x}) = b_t(X_{s,t}(\mathbf{x})), \quad X_{s,s}(\mathbf{x}) = \mathbf{x}. \tag{15}$$

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

$$\partial_s X_{s,t}(\mathbf{x}) + b_s(\mathbf{x}) \cdot \nabla X_{s,t}(\mathbf{x}) = 0, \quad X_{t,t}(\mathbf{x}) = \mathbf{x}. \tag{16}$$

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

$$X_{s,u}(\mathbf{x}) = X_{t,u}(X_{s,t}(\mathbf{x})). \tag{17}$$

For proofs, see Boffi et al. [18].

For each $\mathbf{x} \in \mathbb{R}^d$, the Lagrangian equation is an ODE in t with parameter s, describing forward evolution along trajectories. It was introduced in Boffi et al. [18, 17] and is the basis for Lagrangian self-distillation and terminal velocity matching [44]. The Eulerian equation is a PDE in s describing how the map changes as the starting time varies, and is the basis for consistency models [74] and MeanFlow [42]. The semigroup condition states that two successive jumps can be replaced by a single direct jump, and is the basis for progressive distillation [45] and shortcut models [46].

The following result demonstrates that the flow map contains a flow implicitly, which we use to derive direct training algorithms as well as to provide an anchor on the diagonal for distillation.

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

$$\lim_{s \to t} \partial_t X_{s,t}(\mathbf{x}) = b_t(\mathbf{x}). \tag{18}$$

The proof follows by a direct application of the Lagrangian equation (15). The condition (18) motivates the parameterization

$$X_{s,t}(\mathbf{x}) = \mathbf{x} + (t-s) v_{s,t}(\mathbf{x}), \tag{19}$$

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 (18) [17]. 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. In the next subsection, we show 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.

C.1 Direct training and distillation of 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 Prop. C.2.

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

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

$$\mathcal{L}_{\mathsf{LMD}}(\hat{v}) = \int_0^1 \int_0^t \mathbb{E}|\partial_t \hat{X}_{s,t}(I_s) - \mathsf{sg}(\hat{b}_t(\hat{X}_{s,t}(I_s)))|^2 \mathrm{d}s\,\mathrm{d}t + \int_0^1 \mathbb{E}|\hat{v}_{t,t}(I_t) - \hat{b}_t(I_t)|^2 \mathrm{d}t. \quad (20)$$

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

$$\mathcal{L}_{\mathsf{EMD}}(\hat{v}) = \int_0^1 \int_0^t \mathbb{E}|\partial_s \hat{X}_{s,t}(I_s) + \mathsf{sg}(\hat{b}_s(I_s) \cdot \nabla \hat{X}_{s,t}(I_s))|^2 \mathrm{d}s\,\mathrm{d}t + \int_0^1 \mathbb{E}|\hat{v}_{t,t}(I_t) - \hat{b}_t(I_t)|^2 \mathrm{d}t. \quad (21)$$

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

$$\mathcal{L}_{\mathsf{PMD}}(\hat{v}) = \int_0^1 \int_0^t \int_s^t \mathbb{E}|\hat{X}_{s,t}(I_s) - \mathsf{sg}(\hat{X}_{t,u}(\hat{X}_{s,t}(I_s)))|^2 \mathrm{d}u\,\mathrm{d}s\,\mathrm{d}t + \int_0^1 \mathbb{E}|\hat{v}_{t,t}(I_t) - \hat{b}_t(I_t)|^2 \mathrm{d}t. \quad (22)$$

For proofs, see Boffi et al. [18].

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 (18), 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 C.5 (Flow map self-distillation)**.** *The flow map is the unique critical point of the following losses:*

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

$$\mathcal{L}_{\mathsf{LSD}}(\hat{v}) = \int_0^1 \int_0^t \mathbb{E}|\partial_t \hat{X}_{s,t}(I_s) - \mathsf{sg}(\hat{v}_{t,t}(\hat{X}_{s,t}(I_s)))|^2 \mathrm{d}s\,\mathrm{d}t + \int_0^1 \mathbb{E}|\hat{v}_{t,t}(I_t) - \dot{I}_t|^2 \mathrm{d}t. \quad (23)$$

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

$$\mathcal{L}_{\mathsf{ESD}}(\hat{v}) = \int_0^1 \int_0^t \mathbb{E}|\partial_s \hat{X}_{s,t}(I_s) + \mathsf{sg}(\nabla \hat{X}_{s,t}(I_s)\hat{v}_{s,s}(I_s))|^2 \mathrm{d}s\,\mathrm{d}t + \int_0^1 \mathbb{E}|\hat{v}_{t,t}(I_t) - \dot{I}_t|^2 \mathrm{d}t. \quad (24)$$

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

$$\mathcal{L}_{\mathsf{PSD}}(\hat{v}) = \int_0^1 \int_0^t \int_s^t \mathbb{E}|\hat{X}_{s,t}(I_s) - \mathsf{sg}(\hat{X}_{u,t}(\hat{X}_{s,u}(I_s)))|^2 \mathrm{d}u\, \mathrm{d}s\, \mathrm{d}t + \int_0^1 \mathbb{E}|\hat{v}_{t,t}(I_t) - \dot{I}_t|^2 \mathrm{d}t. \tag{25}$$

For proofs, we refer the reader to Boffi et al. [17]. LSD has recently been scaled under the name Terminal Velocity Matching [44], ESD is equivalent to Improved MeanFlow [43], and PSD can be viewed as a continuous-time limit of shortcut models [46].

D Theoretical details

D.1 Proof of Lem. 3.1

Lemma 3.1. *The optimal denoiser is given by the token-level posterior,*

$$D_t(\mathbf{x})^l = p^l_{1|t}(\cdot | I_t = \mathbf{x}), \tag{26}$$

so that the optimal denoiser lies on the simplex, $D_t(\mathbf{x})^l \in \Delta^{|V|-1}$.

Proof. From (3), we have that for the l-th token position:

$$D_t(\mathbf{x})^l = \mathbb{E}[\mathbf{x}_1^l | I_t = \mathbf{x}]. \tag{27}$$

Let $\mathbf{e}_i \in \mathbb{R}^{|V|}$ denote 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:

$$D_t(\mathbf{x})^l = \mathbb{E}[\mathbf{x}_1^l | I_t = \mathbf{x}] = \sum_{i=1}^{|V|} \mathbf{e}_i \cdot p^l_{1|t}(\mathbf{x}_1^l = \mathbf{e}_i | I_t = \mathbf{x}) = \begin{bmatrix} p^l_{1|t}(\mathbf{x}_1^l = \mathbf{e}_1 | I_t = \mathbf{x}) \\ \dots \\ p^l_{1|t}(\mathbf{x}_1^l = \mathbf{e}_{|V|} | I_t = \mathbf{x}) \end{bmatrix}. \tag{28}$$

This is precisely the vector of posterior probabilities over the vocabulary, $p^l_{1|t}(\cdot | I_t = \mathbf{x})$. □

D.2 Proof of Prop. 3.2

Proposition 3.2. *Consider the cross-entropy objective*

$$\mathcal{L}_{\mathsf{CE}}(\hat{D}) \coloneqq \int_0^1 \mathbb{E}\left[-\sum_{l=1}^{L} \log \hat{D}_t(I_t)^l \cdot \mathbf{x}_1^l \right] \mathrm{d}t. \tag{29}$$

Then, the optimal denoiser D_t is the unique minimizer of $\mathcal{L}_{\mathsf{CE}}$. Moreover, if the excess risk $\Delta_D(\hat{D}) \coloneqq \mathcal{L}_{\mathsf{CE}}(\hat{D}) - \mathcal{L}_{\mathsf{CE}}(D) \le \epsilon$, then for any early stopping time $\xi \in (0,1)$,

$$W_2^2(\hat{p}_{1-\xi}, p_{1-\xi}) \le C\epsilon, \tag{30}$$

where $C > 0$ depends on ξ and the Lipschitz constant of the model.

We prove the two claims separately, as the Wasserstein bound requires further setup. We first prove that the minimizer is correct and unique.

Proof of the minimizer. We first observe that for any two distributions p, q, the cross-entropy decomposes as

$$\mathbb{E}_p[-\log q] = H(p) + \mathsf{KL}(p \,\|\, q). \tag{31}$$

Writing $\hat{p}^l_{1|t}(\cdot \mid I_t) \coloneqq \hat{D}_t(I_t)^l$ for the model's predicted token distribution at position l, the tower property gives

$$\mathbb{E}\Big[-\log \hat{p}^l_{1|t}(\mathbf{x}_1^l \mid I_t) \Big] = \mathbb{E}_{I_t}\Big[\mathbb{E}_{\mathbf{x}_1^l | I_t}\Big[-\log \hat{p}^l_{1|t}(\mathbf{x}_1^l \mid I_t) \Big] \Big]. \tag{32}$$

Applying (31) to the inner expectation with $p = p^l_{1|t}(\cdot \mid I_t)$ and $q = \hat{p}^l_{1|t}(\cdot \mid I_t)$:

$$\mathbb{E}_{\mathbf{x}^l_1|I_t}\Big[-\log \hat{p}^l_{1|t}(\mathbf{x}^l_1 \mid I_t)\Big] = H\Big(p^l_{1|t}(\cdot \mid I_t)\Big) + \mathsf{KL}(p^l_{1|t}(\cdot \mid I_t) \,\|\, \hat{p}^l_{1|t}(\cdot \mid I_t)). \tag{33}$$

Summing over token positions and integrating in time yields the decomposition

$$\mathcal{L}_{\mathsf{CE}}(\hat{D}) = \underbrace{\int_0^1 \mathbb{E}[H(\mathbf{x}_1 \mid I_t)]\,\mathrm{d}t}_{\text{irreducible conditional entropy}} + \int_0^1 \mathbb{E}\left[\sum_{l=1}^{L} \mathsf{KL}(p^l_{1|t}(\cdot \mid I_t) \,\|\, \hat{p}^l_{1|t}(\cdot \mid I_t))\right] \mathrm{d}t, \tag{34}$$

where $H(\mathbf{x}_1 \mid I_t) \coloneqq \sum_{l=1}^{L} H(p^l_{1|t}(\cdot \mid I_t))$. Since $\mathsf{KL}(p \,\|\, q) \geq 0$ with equality if and only if $p = q$, the unique minimizer is $\hat{D}_t(I_t)^l = p^l_{1|t}(\cdot \mid I_t) = D_t(I_t)^l$. □

We now turn to prove the Wasserstein bound, which requires the following preliminaries.

Notation. Given the learned denoiser $\hat{D}$, we have the induced velocity field $\hat{b}_t(\mathbf{x}) \coloneqq (\hat{D}_t(\mathbf{x}) - \mathbf{x})/(1-t)$. Let $X^{\hat{b}}_{s,t}$ denote the exact flow map generated by $\hat{b}$:

$$\partial_t X^{\hat{b}}_{s,t}(\mathbf{x}) = \hat{b}_t(X^{\hat{b}}_{s,t}(\mathbf{x})), \qquad X^{\hat{b}}_{s,s}(\mathbf{x}) = \mathbf{x}. \tag{35}$$

For $\mathbf{x}_0 \sim p_0$, let $p_t \coloneqq (X_{0,t})_{\#}p_0$ denote the true marginal and $p^{\hat{b}}_t \coloneqq (X^{\hat{b}}_{0,t})_{\#}p_0$ the distribution generated by the learned velocity.

By the decomposition (34), the entropy term depends only on the data distribution and cannot be reduced by any model. We define the *diagonal excess risk* as the learnable KL gap:

$$\Delta_D(\hat{D}) \coloneqq \mathcal{L}_{\mathsf{CE}}(\hat{D}) - \int_0^1 \mathbb{E}[H(\mathbf{x}_1 \mid I_t)]\,\mathrm{d}t = \int_0^1 \mathbb{E}\left[\sum_{l=1}^{L} \mathsf{KL}(p^l_{1|t}(\cdot \mid I_t) \,\|\, \hat{p}^l_{1|t}(\cdot \mid I_t))\right] \mathrm{d}t. \tag{36}$$

We will show that the excess risk (36) controls the quality of the ODE sampler driven by $\hat{b}$. Because the excess risk is equal to the cross entropy objective up to a constant, learning a denoiser via cross entropy minimization generates a distribution close to p_1 in Wasserstein distance. The following three standard regularity assumptions are used throughout.

Assumption D.3 (Lipschitz velocities)**.** There exist constants $L_b, L_{\hat{b}} > 0$ such that

$$|b_t(\mathbf{x}) - b_t(\mathbf{y})| \leq L_b\,|\mathbf{x} - \mathbf{y}|, \qquad |\hat{b}_t(\mathbf{x}) - \hat{b}_t(\mathbf{y})| \leq L_{\hat{b}}\,|\mathbf{x} - \mathbf{y}|, \tag{37}$$

for all $t \in [0,1)$ and all $\mathbf{x}, \mathbf{y}$.

Assumption D.4 (Finite second moments)**.** The velocity fields have uniformly bounded second moments under their respective flow distributions:

$$M_b \coloneqq \sup_{t\in[0,1)} \mathbb{E}_{p_t}\left[|b_t|^2\right] < \infty, \qquad M_{\hat{b}} \coloneqq \sup_{t\in[0,1)} \mathbb{E}_{p^{\hat{b}}_t}\left[|\hat{b}_t|^2\right] < \infty. \tag{38}$$

We also assume a simplex-interior property, which is guaranteed in practice by using a softmax output layer.

Assumption D.5 (Simplex-interior outputs)**.** The learned denoiser $\hat{D}_t$ produces outputs in the interior of the probability simplex $\Delta^{|V|-1}$ at each token position.

To connect the excess risk to the L^2 denoiser error, we exploit the simplex structure of the denoiser outputs via Pinsker's inequality.

Lemma D.6. *The L^2 denoiser error is controlled by the excess risk:*

$$\int_0^1 \mathbb{E}\Big[|\hat{D}_t(I_t) - D_t(I_t)|^2\Big]\,\mathrm{d}t \leq 2\,\Delta_D(\hat{D}). \tag{39}$$

Proof. For any $p, q \in \Delta^{|V|-1}$, the total variation distance is $\mathrm{TV}(p,q) = \frac{1}{2}|p-q|_1$, and Pinsker's inequality states

$$\mathrm{TV}(p,q) \le \sqrt{\tfrac{1}{2}\mathsf{KL}(p \,\|\, q)}. \tag{40}$$

Combining these and using $|p-q|^2 \le |p-q|_1^2$,

$$|p-q|^2 \le |p-q|_1^2 = 4\,\mathrm{TV}(p,q)^2 \le 2\,\mathsf{KL}(p \,\|\, q). \tag{41}$$

Applying this bound tokenwise with $p = p^l_{1|t}(\cdot \mid I_t)$ and $q = \hat{p}^l_{1|t}(\cdot \mid I_t)$, summing over l, and integrating in t gives (39) via the decomposition (34). □

With these tools in hand, we bound the velocity error in terms of the cross entropy excess risk. The denoiser-velocity conversion (3) introduces a $(1-t)^{-1}$ singularity, so we truncate the sampler at $1-\xi$ for $\xi \in (0,1)$.

Lemma D.7 (Velocity error from excess risk)**.** *Under Asm. D.5, for any $\xi \in (0,1)$,*

$$\int_0^{1-\xi} \mathbb{E}\Big[|\hat{b}_t(I_t) - b_t(I_t)|^2\Big]\,\mathrm{d}t \le C_b(\xi)\,\Delta_D(\hat{D}), \qquad C_b(\xi) := 2\xi^{-2}. \tag{42}$$

Proof. The denoiser-velocity relation amplifies errors by $(1-t)^{-1}$:

$$\hat{b}_t(\mathbf{x}) - b_t(\mathbf{x}) = \frac{\hat{D}_t(\mathbf{x}) - D_t(\mathbf{x})}{1-t}. \tag{43}$$

For $t \le 1-\xi$, $(1-t)^{-2} \le \xi^{-2}$, so

$$|\hat{b}_t(I_t) - b_t(I_t)|^2 \le \xi^{-2}|\hat{D}_t(I_t) - D_t(I_t)|^2. \tag{44}$$

Integrating over $[0, 1-\xi]$ and bounding by the full integral gives

$$\int_0^{1-\xi} \mathbb{E}|\hat{b}_t(I_t) - b_t(I_t)|^2\,\mathrm{d}t \le \xi^{-2}\int_0^1 \mathbb{E}|\hat{D}_t(I_t) - D_t(I_t)|^2\,\mathrm{d}t. \tag{45}$$

Applying (39) gives (42). □

We now convert the velocity error into a distributional bound via Gronwall's inequality, adapting the approach of Boffi et al. [18] to the denoiser reparameterization.

Proposition D.8 (Flow error)**.** *Under Asm. D.3 to D.5, for any $\xi \in (0,1)$,*

$$W_2\Big(p_1^{\hat{b}}, p_1\Big) \le C_D(\xi)\sqrt{\Delta_D(\hat{D})} + r_\xi, \tag{46}$$

where $C_D(\xi) := \sqrt{2(1-\xi)}\,e^{L_{\hat{b}}(1-\xi)}/\xi$ and the terminal remainder satisfies

$$r_\xi := W_2(p_1, p_{1-\xi}) + W_2(p_1^{\hat{b}}, p_{1-\xi}^{\hat{b}}) \le \xi\Big(\sqrt{M_b} + \sqrt{M_{\hat{b}}}\Big), \tag{47}$$

with M_b and $M_{\hat{b}}$ as defined in Asm. D.4.

Proof. Coupling and error dynamics. We couple both flows by initializing from the same sample $\mathbf{x}_0 \sim p_0$:

$$X_t := X_{0,t}(\mathbf{x}_0), \qquad \hat{X}_t := X^{\hat{b}}_{0,t}(\mathbf{x}_0), \qquad e_t := \hat{X}_t - X_t. \tag{48}$$

The error dynamics decompose into a Lipschitz term and a forcing term:

$$\dot{e}_t = \hat{b}_t(\hat{X}_t) - b_t(X_t) = \big[\hat{b}_t(\hat{X}_t) - \hat{b}_t(X_t)\big] + \big[\hat{b}_t(X_t) - b_t(X_t)\big]. \tag{49}$$

Gronwall bound on $[0, 1-\xi]$. Using Lipschitzness of $\hat{b}_t$ with constant $L_{\hat{b}}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}|e_t| \le L_{\hat{b}}|e_t| + |\hat{b}_t(X_t) - b_t(X_t)|. \tag{50}$$

By Gronwall's inequality, for $T = 1 - \xi$:

$$|e_T| \le e^{L_{\hat{b}} T} \int_0^T |\hat{b}_t(X_t) - b_t(X_t)| \, \mathrm{d}t. \tag{51}$$

Applying the Cauchy–Schwarz inequality to the integral and squaring:

$$|e_T|^2 \le e^{2L_{\hat{b}} T} T \int_0^T |\hat{b}_t(X_t) - b_t(X_t)|^2 \, \mathrm{d}t. \tag{52}$$

Taking expectations and using the coupling $W_2^2(p_T^{\hat{b}}, p_T) \le \mathbb{E}|e_T|^2$:

$$W_2^2(p_{1-\xi}^{\hat{b}}, p_{1-\xi}) \le e^{2L_{\hat{b}}(1-\xi)}(1-\xi) \int_0^{1-\xi} \mathbb{E}|\hat{b}_t(I_t) - b_t(I_t)|^2 \, \mathrm{d}t. \tag{53}$$

Applying Lem. D.7:

$$W_2(p_{1-\xi}^{\hat{b}}, p_{1-\xi}) \le C_D(\xi)\sqrt{\Delta_D(\hat{D})}, \qquad C_D(\xi) = \frac{\sqrt{2(1-\xi)}\, e^{L_{\hat{b}}(1-\xi)}}{\xi}. \tag{54}$$

Terminal remainder. The triangle inequality on $(\mathcal{P}_2, W_2)$ gives

$$W_2(p_1^{\hat{b}}, p_1) \le W_2(p_1^{\hat{b}}, p_{1-\xi}^{\hat{b}}) + W_2(p_{1-\xi}^{\hat{b}}, p_{1-\xi}) + W_2(p_{1-\xi}, p_1). \tag{55}$$

Each remainder term is controlled by flowing over the short interval $[1-\xi, 1]$. For the true flow, we couple $X_{1-\xi} \sim p_{1-\xi}$ and $X_1 \sim p_1$ along the same trajectory, so that $X_1 - X_{1-\xi} = \int_{1-\xi}^1 b_t(X_t) \, \mathrm{d}t$. The coupling gives

$$W_2^2(p_{1-\xi}, p_1) \le \mathbb{E}\left[\left|\int_{1-\xi}^1 b_t(X_t) \, \mathrm{d}t\right|^2\right]. \tag{56}$$

By the Cauchy–Schwarz inequality, $|\int f \, \mathrm{d}t|^2 \le (\int 1^2 \, \mathrm{d}t)(\int |f|^2 \, \mathrm{d}t)$, so

$$\left|\int_{1-\xi}^1 b_t(X_t) \, \mathrm{d}t\right|^2 \le \xi \int_{1-\xi}^1 |b_t(X_t)|^2 \, \mathrm{d}t. \tag{57}$$

Taking expectations and bounding the integrand by M_b via Asm. D.4:

$$W_2^2(p_{1-\xi}, p_1) \le \xi \int_{1-\xi}^1 \mathbb{E}\big[|b_t(X_t)|^2\big] \, \mathrm{d}t \le \xi^2 M_b. \tag{58}$$

The same argument applied to the learned flow gives $W_2(p_{1-\xi}^{\hat{b}}, p_1^{\hat{b}}) \le \xi\sqrt{M_{\hat{b}}}$ via Asm. D.4. Substituting (54) and (58) into (55) yields (46). □

D.3 Proof of Prop. D.9

Proposition D.9. *Let S be a finite set. For any probability distribution μ on S, there exists a distribution ν on S that cannot be expressed as $\nu = f_{\#}\mu$ for any deterministic map $f : S \to S$.*

Proof. The pushforward $\nu = f_{\#}\mu$ satisfies

$$\nu(\mathbf{y}) = \sum_{\mathbf{x} \in f^{-1}(\mathbf{y})} \mu(\mathbf{x}), \tag{59}$$

where $f^{-1}(\mathbf{y})$ denotes the preimage of $\{\mathbf{y}\}$. Since $f^{-1}(\mathbf{y}) \subseteq S$, each probability $\nu(\mathbf{y})$ is a sum of some subset of the values $\{\mu(\mathbf{x}) \mid \mathbf{x} \in S\}$. In particular, every nonempty preimage contributes at least $\mu_{\min} := \min\{\mu(\mathbf{x}) \mid \mu(\mathbf{x}) > 0\}$, so $\nu(\mathbf{y}) \in \{0\} \cup [\mu_{\min}, 1]$ for all $\mathbf{y}$. Constructing ν with $\nu(\mathbf{y}_0) = \mu_{\min}/2$ for some $\mathbf{y}_0 \in S$ yields a distribution that no deterministic map can produce. □

Choosing $S = V^L$ and $\mu = p_0$ for discrete diffusion, this shows that for any noise distribution p_0, there exists a data distribution p that cannot be reached via one-step deterministic transport. By contrast, continuous flows admit a flow map at the sample level (Sec. 3.3), which is the basis for our FMLM approach.

D.4 Flow maps on the simplex

In Sec. 3.2, we introduced the denoiser D_t in (3). 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 (4). We now develop an analogous reparameterization for the flow map. To do so, we leverage the two-time denoiser $\delta_{s,t}$ (8), which 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

$$I_t = \alpha_t \mathbf{x}_0 + \beta_t \mathbf{x}_1, \tag{60}$$

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 D.10 (Endpoint denoiser)**.** The endpoint denoiser $D_t : \mathbb{R}^d \to \mathbb{R}^d$ is defined as:

$$D_t(\mathbf{x}) := \mathbb{E}[\mathbf{x}_1 | I_t = \mathbf{x}]. \tag{61}$$

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 D.11 (Denoiser-velocity relation)**.** *For general interpolant coefficients α_t, β_t, the velocity field and endpoint denoiser are related by:*

$$b_t(\mathbf{x}) = \frac{\dot{\beta}_t}{\beta_t} D_t(\mathbf{x}) + \left(\dot{\alpha}_t - \frac{\alpha_t \dot{\beta}_t}{\beta_t} \right) \frac{\mathbf{x} - \beta_t D_t(\mathbf{x})}{\alpha_t}. \tag{62}$$

Proof. Conditioning on $I_t = \mathbf{x}$ gives:

$$\mathbf{x} = \alpha_t \mathbb{E}[\mathbf{x}_0 | I_t = \mathbf{x}] + \beta_t D_t(\mathbf{x}). \tag{63}$$

The velocity field is:

$$b_t(\mathbf{x}) = \mathbb{E}[\dot{I}_t | I_t = \mathbf{x}] = \dot{\alpha}_t \mathbb{E}[\mathbf{x}_0 | I_t = \mathbf{x}] + \dot{\beta}_t D_t(\mathbf{x}). \tag{64}$$

Solving for $\mathbb{E}[\mathbf{x}_0 | I_t = \mathbf{x}] = (\mathbf{x} - \beta_t D_t(\mathbf{x}))/\alpha_t$ and substituting yields (62). □

This relation first appeared in Albergo et al. [15]. For $\alpha_t = 1 - t$ and $\beta_t = t$ as in the main text, (62) simplifies to:

$$b_t(\mathbf{x}) = \frac{D_t(\mathbf{x}) - \mathbf{x}}{1 - t}. \tag{65}$$

Rearranging gives:

$$D_t(\mathbf{x}) = \mathbf{x} + (1 - t) b_t(\mathbf{x}). \tag{66}$$

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.

D.5 The two-time denoiser

We now restate Prop. 3.3 and prove each property. The proof of the simplex property (ii) proceeds by solving the flow ODE via an integrating factor, which reveals that $\delta_{s,t}$ admits an integral representation as a weighted average of denoisers along the flow trajectory. Non-negativity and normalization then follow from the corresponding properties of the single-time denoiser (Lem. 3.1).

Proposition 3.3. *The two-time denoiser $\delta_{s,t}$ satisfies the following four properties:*

(i) The flow map can be recovered exactly,

$$X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\,\mathbf{x} + \frac{t-s}{1-s}\,\delta_{s,t}(\mathbf{x}). \tag{67}$$

(ii) The two-time denoiser lies on the simplex,

$$\delta_{s,t}(\mathbf{x})^l \in \Delta^{|V|-1} \tag{68}$$

for all token positions $l = 1, \ldots, L$.

(iii) The two-time denoiser recovers the standard denoiser on the diagonal,

$$\delta_{s,s}(\mathbf{x}) = D_s(\mathbf{x}), \quad \textit{for all} \quad \mathbf{x} \in \mathbb{R}^{L\times|V|}, \;\; s \in [0,1]. \tag{69}$$

(iv) The two-time denoiser satisfies a semigroup condition,

$$\delta_{s,t}(\mathbf{x}) = \gamma\,\delta_{s,u}(\mathbf{x}) + (1-\gamma)\,\delta_{u,t}(X_{s,u}(\mathbf{x})), \tag{70}$$

where $\gamma = \frac{(1-t)(u-s)}{(1-u)(t-s)} \in [0,1]$.

Proof. We prove each property in turn, recalling the parameterization $X_{s,t}(\mathbf{x}) = \mathbf{x} + (t-s)v_{s,t}(\mathbf{x})$ from (19).

(i) Flow map reparameterization. Substituting $v_{s,t} = (\delta_{s,t} - \mathbf{x})/(1-s)$ into (19) immediately gives:

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

(ii) Simplex. We show that $\delta_{s,t}$ can be written as a weighted average of single-time denoisers along the flow trajectory. The flow ODE in denoiser form is:

$$\partial_\tau X_{s,\tau}(\mathbf{x}) = \frac{D_\tau(X_{s,\tau}(\mathbf{x})) - X_{s,\tau}(\mathbf{x})}{1-\tau}. \tag{72}$$

Rearranging and multiplying by the integrating factor $1/(1-\tau)$:

$$\frac{\partial}{\partial\tau}\left(\frac{X_{s,\tau}(\mathbf{x})}{1-\tau}\right) = \frac{D_\tau(X_{s,\tau}(\mathbf{x}))}{(1-\tau)^2}. \tag{73}$$

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

$$X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\,\mathbf{x} + (1-t)\int_s^t \frac{D_\tau(X_{s,\tau}(\mathbf{x}))}{(1-\tau)^2}\,d\tau. \tag{74}$$

Comparing with (i) and matching coefficients:

$$\delta_{s,t}(\mathbf{x}) = \frac{(1-s)(1-t)}{t-s}\int_s^t \frac{D_\tau(X_{s,\tau}(\mathbf{x}))}{(1-\tau)^2}\,d\tau. \tag{75}$$

By Lem. 3.1, each $D_\tau(X_{s,\tau}(\mathbf{x}))$ is non-negative at every token position and sums to one. Since $(1-s)(1-t)/(1-\tau)^2 > 0$, non-negativity of $\delta_{s,t}$ follows immediately. To show that $\delta_{s,t}$ always sums to one, we use $\sum_v D_\tau^{l,v} = 1$ and evaluate:

$$\begin{aligned}\sum_v \delta_{s,t}^{l,v}(\mathbf{x}) &= \frac{(1-s)(1-t)}{t-s}\int_s^t \frac{d\tau}{(1-\tau)^2} \\ &= \frac{(1-s)(1-t)}{t-s}\left(\frac{1}{1-t} - \frac{1}{1-s}\right) = 1.\end{aligned} \tag{76}$$

Because $\delta_{s,t}$ is always non-negative and sums to one, it lies on the simplex.

(iii) Diagonal. By direct computation using $v_{s,s} = b_s$ and (66):

$$\delta_{s,s}(\mathbf{x}) = \mathbf{x} + (1-s)b_s(\mathbf{x}) = D_s(\mathbf{x}). \tag{77}$$

(iv) Semigroup. See the following result in App. D.6. □

D.6 Characterizing the two-time denoiser

Since $\delta_{s,t}$ lies on the simplex by Prop. 3.3, we now aim to design objective functions that are entirely simplex-valued, for which cross entropy-based objectives can then be used. To this end, we now translate the flow map characterizations from Prop. C.2 into conditions on $\delta_{s,t}$.

Proposition D.4 (Flow map characterizations in δ space)**.** *The flow map characterizations from Prop. C.2 translate into the following conditions on $\delta_{s,t}$:*

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

$$\delta_{s,t}(\mathbf{x}) = \delta_{t,t}(X_{s,t}(\mathbf{x})) - \frac{(1-t)(t-s)}{1-s}\partial_t \delta_{s,t}(\mathbf{x}). \tag{78}$$

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

$$\delta_{s,t}(\mathbf{x}) = \delta_{s,s}(\mathbf{x}) + \frac{t-s}{1-t}\Big((1-s)\,\partial_s \delta_{s,t}(\mathbf{x}) + (\delta_{s,s}(\mathbf{x}) - \mathbf{x}) \cdot \nabla \delta_{s,t}(\mathbf{x})\Big). \tag{79}$$

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

$$\begin{aligned} \delta_{s,t}(\mathbf{x}) &= \gamma \cdot \delta_{s,u}(\mathbf{x}) + (1-\gamma) \cdot \delta_{u,t}(X_{s,u}(\mathbf{x})), \\ \gamma &= \frac{(1-t)(u-s)}{(1-u)(t-s)}. \end{aligned} \tag{80}$$

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 (71) in t:

$$\partial_t X_{s,t}(\mathbf{x}) = -\frac{1}{1-s}\mathbf{x} + \frac{1}{1-s}\delta_{s,t}(\mathbf{x}) + \frac{t-s}{1-s}\partial_t \delta_{s,t}(\mathbf{x}). \tag{81}$$

By the Lagrangian equation (15), $\partial_t X_{s,t}(\mathbf{x}) = b_t(X_{s,t}(\mathbf{x}))$. Rewriting b_t via the denoiser-velocity relation (3):

$$\partial_t X_{s,t}(\mathbf{x}) = \frac{D_t(X_{s,t}(\mathbf{x})) - X_{s,t}(\mathbf{x})}{1-t}. \tag{82}$$

Multiplying both sides by $(1-t)$, adding $X_{s,t}$, and substituting (71):

$$\begin{aligned} D_t(X_{s,t}(\mathbf{x})) &= X_{s,t}(\mathbf{x}) + (1-t)\partial_t X_{s,t}(\mathbf{x}) \\ &= \underbrace{\frac{1-t}{1-s}\mathbf{x} - \frac{1-t}{1-s}\mathbf{x}}_{=0} + \underbrace{\frac{t-s}{1-s}\delta_{s,t}(\mathbf{x}) + \frac{1-t}{1-s}\delta_{s,t}(\mathbf{x})}_{=\delta_{s,t}(\mathbf{x})} + \frac{(1-t)(t-s)}{1-s}\partial_t \delta_{s,t}(\mathbf{x}) \\ &= \delta_{s,t}(\mathbf{x}) + \frac{(1-t)(t-s)}{1-s}\partial_t \delta_{s,t}(\mathbf{x}). \end{aligned} \tag{83}$$

Since $\delta_{t,t} = D_t$ on the diagonal (77), the left-hand side is $\delta_{t,t}(X_{s,t}(\mathbf{x}))$, giving (78).

(ii) Eulerian. We substitute (71) into the Eulerian equation (16). Differentiating (71) in s:

$$\partial_s X_{s,t}(\mathbf{x}) = \frac{1-t}{(1-s)^2}(\mathbf{x} - \delta_{s,t}(\mathbf{x})) + \frac{t-s}{1-s}\partial_s \delta_{s,t}(\mathbf{x}). \tag{84}$$

The spatial Jacobian of (71) is:

$$\nabla X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\mathrm{Id} + \frac{t-s}{1-s}\nabla \delta_{s,t}(\mathbf{x}). \tag{85}$$

By the denoiser-velocity relation (3), the advection velocity is $b_s(\mathbf{x}) = (\delta_{s,s}(\mathbf{x}) - \mathbf{x})/(1-s)$. Substituting into (16) and expanding $b_s \cdot \nabla X_{s,t}$:

$$\begin{aligned} 0 = &\frac{1-t}{(1-s)^2}(\mathbf{x} - \delta_{s,t}(\mathbf{x})) + \frac{t-s}{1-s}\partial_s \delta_{s,t}(\mathbf{x}) \\ &+ \frac{(1-t)(\delta_{s,s}(\mathbf{x}) - \mathbf{x})}{(1-s)^2} + \frac{(t-s)(\delta_{s,s}(\mathbf{x}) - \mathbf{x})}{(1-s)^2} \cdot \nabla \delta_{s,t}(\mathbf{x}). \end{aligned} \tag{86}$$

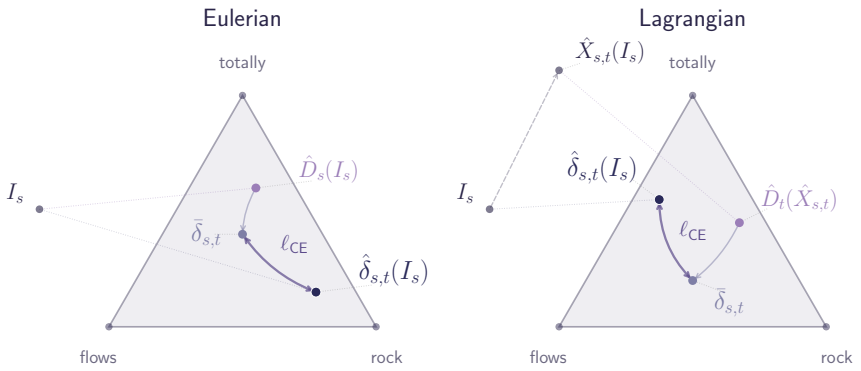


Figure 7: **Eulerian and Lagrangian objectives on the simplex.** (Left) The Eulerian teacher $\bar{\delta}_{s,t}$ is constructed from $\hat{D}_s(I_s)$ and derivatives of $\hat{\delta}_{s,t}$. (Right) The Lagrangian teacher is constructed from $\hat{D}_t(\hat{X}_{s,t}(I_s))$, requiring an intermediate flow map evaluation off the simplex. In both cases, the teacher may transiently leave the simplex during training due to derivative correction terms, but the cross-entropy loss remains well-defined since only the student $\hat{\delta}_{s,t}(I_s)$ (parameterized with softmax) must lie on the simplex. At optimality, all quantities lie on the simplex.

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 (79).

(iii) Semigroup. We express each side of $X_{s,t}(\mathbf{x}) = X_{u,t}(X_{s,u}(\mathbf{x}))$ using (71). The left-hand side is:

$$X_{s,t}(\mathbf{x}) = \frac{1-t}{1-s}\mathbf{x} + \frac{t-s}{1-s}\delta_{s,t}(\mathbf{x}). \tag{87}$$

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:

$$\begin{aligned} X_{u,t}(\mathbf{z}) &= \frac{1-t}{1-u}\mathbf{z} + \frac{t-u}{1-u}\delta_{u,t}(\mathbf{z}) \\ &= \frac{1-t}{1-u}\left[\frac{1-u}{1-s}\mathbf{x} + \frac{u-s}{1-s}\delta_{s,u}(\mathbf{x})\right] + \frac{t-u}{1-u}\delta_{u,t}(\mathbf{z}) \\ &= \frac{1-t}{1-s}\mathbf{x} + \frac{(1-t)(u-s)}{(1-u)(1-s)}\delta_{s,u}(\mathbf{x}) + \frac{t-u}{1-u}\delta_{u,t}(\mathbf{z}). \end{aligned} \tag{88}$$

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

$$\frac{t-s}{1-s}\delta_{s,t}(\mathbf{x}) = \frac{(1-t)(u-s)}{(1-u)(1-s)}\delta_{s,u}(\mathbf{x}) + \frac{t-u}{1-u}\delta_{u,t}(\mathbf{z}). \tag{89}$$

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

$$\delta_{s,t}(\mathbf{x}) = \frac{(1-t)(u-s)}{(1-u)(t-s)}\delta_{s,u}(\mathbf{x}) + \frac{(t-u)(1-s)}{(1-u)(t-s)}\delta_{u,t}(\mathbf{z}). \tag{90}$$

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:

$$\begin{aligned} (1-t)(u-s) + (t-u)(1-s) &= u - s - tu + ts + t - ts - u + us \\ &= (t-s) - u(t-s) \\ &= (t-s)(1-u). \end{aligned} \tag{91}$$

Dividing by $(1-u)(t-s)$ confirms the coefficients sum to one, giving (80). □

D.7 Learning the two-time denoiser

Each characterization in Prop. D.4 gives $\delta_{s,t} =$ (teacher), and each objective below enforces one such condition via an off-diagonal loss, plus a diagonal term that anchors $\hat{\delta}_{t,t}$ to the denoiser. We first restate Prop. 3.4 from the main text.

Proposition 3.4. *The two-time denoiser $\delta_{s,t}$ is the unique critical point of the following KL-based semigroup objective:*

$$\mathcal{L}_{\mathsf{KL}}(\delta) \coloneqq \mathbb{E}_{t,s,u}\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\Big[\sum_{l=1}^{L}\mathsf{KL}(\bar{\delta}^l_{s,t} \,\|\, \delta^l_{s,t}(I_s))\Big] + \mathbb{E}_t\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\Big[\sum_{l=1}^{L}\mathsf{KL}(D^l_t(I_t) \,\|\, \delta^l_{t,t}(I_t))\Big], \tag{92}$$
$$\bar{\delta}_{s,t} \coloneqq \mathsf{sg}(\gamma\delta_{s,u}(I_s) + (1-\gamma)\delta_{u,t}(X_{s,u}(I_s)))$$

where $\bar{\delta}_{s,t}$ is the semigroup teacher, and where the expectation over (s,u,t) is taken from a distribution that has full support on $\{0 \le s \le u \le t \le 1\}$.

We now prove this as part of two larger propositions that provide Lagrangian, Eulerian, and semigroup objectives for both distillation and self-distillation, mirroring the flow map objectives in Propositions C.4 and C.5. For the semigroup characterization, the teacher is a convex combination of simplex elements and therefore lies on the simplex, so we can use the KL divergence. For the Lagrangian and Eulerian characterizations, the teachers may transiently leave the simplex due to derivative terms; for these we use cross-entropy, which remains well-defined whenever the student is parameterized with softmax. For the diagonal distillation term, $\hat{D}_t$ is on the simplex, so we again use KL.

Proposition D.5 (Denoiser flow map distillation)**.** *Given a pre-trained denoiser $\hat{D}_t$, the corresponding two-time denoiser $\delta_{s,t}$ is the unique critical point of the following losses:*

(i) The Lagrangian loss:

$$\mathcal{L}_{\mathsf{lag}}(\hat{\delta}) = -\mathbb{E}_{s,t}\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathsf{sg}(\bar{\delta}^l_{s,t})\cdot\log\hat{\delta}^l_{s,t}(I_s)\right] + \mathbb{E}_t\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathsf{KL}(\hat{D}^l_t(I_t) \,\|\, \hat{\delta}^l_{t,t}(I_t))\right], \tag{93}$$

where $\bar{\delta}_{s,t} \coloneqq \hat{D}_t(\hat{X}_{s,t}(I_s)) - \frac{(1-t)(t-s)}{1-s}\partial_t\hat{\delta}_{s,t}(I_s)$ is given by the right-hand side of (78) with the pre-trained $\hat{D}_t$ replacing $\hat{\delta}_{t,t}$.

(ii) The Eulerian loss:

$$\mathcal{L}_{\mathsf{eul}}(\hat{\delta}) = -\mathbb{E}_{s,t}\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathsf{sg}(\bar{\delta}^l_{s,t})\cdot\log\hat{\delta}^l_{s,t}(I_s)\right] + \mathbb{E}_t\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathsf{KL}(\hat{D}^l_t(I_t) \,\|\, \hat{\delta}^l_{t,t}(I_t))\right], \tag{94}$$

where $\bar{\delta}_{s,t} \coloneqq \hat{D}_s(I_s) + \frac{t-s}{1-t}\Big((1-s)\,\partial_s\hat{\delta}_{s,t}(I_s) + (\hat{D}_s(I_s) - I_s)\cdot\nabla\hat{\delta}_{s,t}(I_s)\Big)$ is given by the right-hand side of (79), with the pre-trained $\hat{D}_s$ replacing $\hat{\delta}_{s,s}$.

(iii) The semigroup loss:

$$\mathcal{L}_{\mathsf{semi}}(\hat{\delta}) = \mathbb{E}_{s,u,t}\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathsf{KL}(\mathsf{sg}(\bar{\delta}^l_{s,t}) \,\|\, \hat{\delta}^l_{s,t}(I_s))\right] + \mathbb{E}_t\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathsf{KL}(\hat{D}^l_t(I_t) \,\|\, \hat{\delta}^l_{t,t}(I_t))\right], \tag{95}$$

where $\bar{\delta}_{s,t} \coloneqq \gamma\,\hat{\delta}_{s,u}(I_s) + (1-\gamma)\,\hat{\delta}_{u,t}(\hat{X}_{s,u}(I_s))$ is given by the right-hand side of (80), and where the expectation over (s,u,t) has full support on $\{0 \le s \le u \le t \le 1\}$.

Above, $\hat{X}_{s,t}$ is recovered from $\hat{\delta}_{s,t}$ via (71).

Proof. Since $\hat{D}_t$ is frozen, the diagonal KL is minimized when $\hat{\delta}_{t,t} = \hat{D}_t$. The off-diagonal term is minimized when $\hat{\delta}_{s,t}$ equals the corresponding teacher, which by Prop. D.4 is equivalent to $\hat{\delta}$ satisfying the Lagrangian, Eulerian, or semigroup characterization. Both terms are simultaneously minimized if and only if $\hat{\delta} = \delta$, giving uniqueness. □

We now state an analogous result for direct training via self-distillation. Since $\mathbf{x}_1$ is one-hot, the diagonal reduces to cross-entropy (4).

Proposition D.6 (Denoiser self-distillation)**.** *The two-time denoiser $\delta_{s,t}$ is the unique critical point of the following losses:*

(i) The Lagrangian loss:

$$\mathcal{L}^{\mathsf{sd}}_{\mathsf{lag}}(\hat{\delta}) = -\mathbb{E}_{s,t}\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathsf{sg}(\bar{\delta}^l_{s,t})\cdot\log\hat{\delta}^l_{s,t}(I_s)\right] - \mathbb{E}_t\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathbf{x}_1^l\cdot\log\hat{\delta}^l_{t,t}(I_t)\right], \quad (96)$$

where $\bar{\delta}_{s,t} \coloneqq \hat{\delta}_{t,t}(\hat{X}_{s,t}(I_s)) - \frac{(1-t)(t-s)}{1-s}\partial_t\hat{\delta}_{s,t}(I_s)$ is given by the right-hand side of (78).

(ii) The Eulerian loss:

$$\mathcal{L}^{\mathsf{sd}}_{\mathsf{eul}}(\hat{\delta}) = -\mathbb{E}_{s,t}\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathsf{sg}(\bar{\delta}^l_{s,t})\cdot\log\hat{\delta}^l_{s,t}(I_s)\right] - \mathbb{E}_t\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathbf{x}_1^l\cdot\log\hat{\delta}^l_{t,t}(I_t)\right], \quad (97)$$

where $\bar{\delta}_{s,t} \coloneqq \hat{\delta}_{s,s}(I_s) + \frac{t-s}{1-t}\Big((1-s)\,\partial_s\hat{\delta}_{s,t}(I_s) + (\hat{\delta}_{s,s}(I_s) - I_s)\cdot\nabla\hat{\delta}_{s,t}(I_s)\Big)$ is given by the right-hand side of (79).

(iii) The semigroup loss:

$$\mathcal{L}^{\mathsf{sd}}_{\mathsf{semi}}(\hat{\delta}) = \mathbb{E}_{s,u,t}\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathsf{KL}(\mathsf{sg}(\bar{\delta}^l_{s,t})\,\|\,\hat{\delta}^l_{s,t}(I_s))\right] - \mathbb{E}_t\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1}\left[\sum_{l=1}^{L}\mathbf{x}_1^l\cdot\log\hat{\delta}^l_{t,t}(I_t)\right], \quad (98)$$

where $\bar{\delta}_{s,t} \coloneqq \gamma\,\hat{\delta}_{s,u}(I_s) + (1-\gamma)\,\hat{\delta}_{u,t}(\hat{X}_{s,u}(I_s))$ is given by the right-hand side of (80), and where the expectation over (s,u,t) has full support on $\{0 \le s \le u \le t \le 1\}$.

Proof. The off-diagonal term is minimized if and only if the corresponding characterization holds. Since $\mathbf{x}_1$ is one-hot, the diagonal cross-entropy is bounded below by the conditional entropy $\mathbb{E}_t\,\mathbb{E}[H(\mathbf{x}_1 \mid I_t)]$, achieved when $\hat{\delta}_{t,t} = D_t$ (App. D.2). Both terms are simultaneously minimized if and only if $\hat{\delta} = \delta$, giving uniqueness. □

D.8 Decoding error rate and entropic time reparameterizations

In this subsection, we show a connection between our time reparameterization (13), that linearizes a decoding error probability, and the entropic time reparameterization proposed in Dieleman et al. [13] and Stancevic et al. [47], that linearizes uncertainty in the generation.

The entropic time reparameterization is defined such that each timepoint contributes equally in resolving the uncertainty in the final generation, naturally quantified by the conditional entropy of clean data conditioned on the noisy state. We consider a standardized entropic time $\sigma : [0,1] \to [0,1]$ that, as proposed in Dieleman et al. [13], linearizes the token-level conditional entropy $\hat{H}(\mathbf{x}_1|I_t) \coloneqq \sum_{l=1}^{L} H(p^l_{1|t}(\cdot \mid I_t))$:

$$\sigma(t) \coloneqq 1 - \frac{\hat{H}(\mathbf{x}_1|I_t)}{\hat{H}(\mathbf{x}_1|\mathbf{x}_0)} = 1 - \frac{\hat{H}(\mathbf{x}_1|I_t)}{\hat{H}(\mathbf{x}_1)}. \quad (99)$$

In Dieleman et al. [13], $\hat{H}(\mathbf{x}_1|I_t)$ is estimated online using the training loss of the denoiser.

We now show a relationship between the entropic time $\sigma(t)$ and our decoding error-based time reparameterization $\tau(t)$ (13) as an approximate asymptotic inequality with a vanishing margin, leveraging Fano's inequality which relates the probability of decoding error and conditional entropy in a noisy communication channel.

Proposition D.7. *Assume that $\mathbf{x}_1^l$ at each position l is distributed uniformly over V. Then:*

$$\tau(t) \le \sigma(t) + O\left(\frac{1}{\log|V|}\right) \qquad \text{as} \qquad |V| \longrightarrow \infty. \quad (100)$$

Algorithm 1 FLM training

Require: Dataset $\mathcal{D}$, reparameterization $\tau(t)$, lr η
Initialize: Denoiser $\hat{D}$
repeat
  $\mathbf{x}_1 \leftarrow f(\mathbf{y}), \mathbf{y} \sim \mathcal{D}$; $\mathbf{x}_0 \sim \mathsf{N}(0, I)$
  Sample t via $\tau(t) \sim \mathsf{U}[0, 1]$
  $I_t \leftarrow (1-t)\mathbf{x}_0 + t\mathbf{x}_1$
  $\hat{\mathbf{x}}_1 \leftarrow \hat{D}_t(I_t)$
  Update $\hat{D}$: $\mathcal{L}_{\mathsf{CE}} = -\sum_l (\mathbf{x}_1^l)^\top \log \hat{\mathbf{x}}_1^l$
until converged

Algorithm 2 FLM sampling

Require: Trained $\hat{D}$, $\tau(t)$, steps N
$\mathbf{x}_0 \sim \mathsf{N}(0, I)$
$t_n \leftarrow t(n/N)$ for $n = 0, ..., N$
for $n = 0$ **to** $N-1$ **do**
  $\hat{\mathbf{x}}_1 \leftarrow \hat{D}_{t_n}(\mathbf{x}_{t_n})$
  $\hat{b}_n \leftarrow (\hat{\mathbf{x}}_1 - \mathbf{x}_{t_n})/(1-t_n)$
  $\mathbf{x}_{t_{n+1}} \leftarrow \mathbf{x}_{t_n} + (t_{n+1} - t_n)\hat{b}_n$
end for
Return $g(\mathbf{x}_{t_N})$

Proof. For each token position l and flow time t, let us denote by $P_e^l(t)$ the probability of an occurrence of decoding error $\mathbf{x}_1^l \neq \mathsf{argmax}(I_t^l)$ where $\mathsf{argmax}(I_t^l)$ is viewed as an approximation of $\mathbf{x}_1^l$. By Fano's inequality,

$$\begin{aligned}\hat{H}(\mathbf{x}_1|I_t) = \sum_{l=1}^{L} H(p^l_{1|t}(\cdot|I_t)) &\leq \sum_{l=1}^{L} \left[P_e^l(t)\log(|V|-1) + h(P_e^l(t))\right] \\ &= L\log(|V|-1)P_e(t) + \sum_{l=1}^{L} h(P_e^l(t)),\end{aligned} \tag{101}$$

where $h(p) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$ is the binary entropy function, and the last equality follows from definition. Then, by (99):

$$\sigma(t) \geq 1 - \frac{L\log(|V|-1)P_e(t) + \sum_l h(P_e^l(t))}{\hat{H}(\mathbf{x}_1)}. \tag{102}$$

Since $h(\cdot) \leq \log 2$, and $\hat{H}(\mathbf{x}_1) = L\log|V|$ due to the uniform distribution assumption, we have

$$\sigma(t) + \frac{\log 2}{\log|V|} \geq 1 - \frac{\log(|V|-1)}{\log|V|}P_e(t) \longrightarrow 1 - P_e(t) \qquad \text{as} \qquad |V| \longrightarrow \infty. \tag{103}$$

By comparing this with $\tau(t) = 1 - \frac{|V|}{|V|-1}P_e(t) \to 1 - P_e(t)$ as $|V| \to \infty$ (13), and noting $\log 2/\log|V| = O(1/\log|V|)$, we arrive at (100). □

The assumption of Prop. D.7 is rather strong: it is more realistic to use a non-uniform distribution following Zipf's law, which would give us $\sigma + O(1/\log|V|) \geq 1 - cP_e$ for some $c > 1$. Nevertheless, the result reveals a close connection between our decoding error-based time reparameterization and entropic time. While we empirically observed in Sec. 5.4 that entropic time based on online estimation [13] underperformed our time reparameterization, this relationship suggests that the reason might lie more in the approximation based on training loss, than the linearization of entropy itself.

E Implementation details

Time reparameterization. To efficiently implement the time reparameterization $\tau(t)$ described in Sec. 4 without evaluating the probability sum during training, we utilize a precomputed lookup table (LUT) combined with spline interpolation. Specifically, we approximate the cumulative distribution function (CDF) of (13) using Gauss-Hermite quadrature and evaluate it on an 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.

Algorithm 3 FMLM training (distillation)

Require: Dataset $\mathcal{D}$, trained $\hat{D}$, $\tau(t)$, lr η
Initialize: Two-time denoiser $\hat{\delta}$
repeat
  $\mathbf{x}_1 \leftarrow f(\mathbf{y}), \mathbf{y} \sim \mathcal{D}$; $\mathbf{x}_0 \sim \mathsf{N}(0, I)$
  Diagonal (anchors to $\hat{D}$):
    $\tau(s) \sim \mathsf{U}[0,1]$; $I_s \leftarrow (1-s)\mathbf{x}_0 + s\mathbf{x}_1$
    $\mathcal{L}_{\mathsf{diag}} \leftarrow -\sum_l \hat{D}(I_s)^l \cdot \log \hat{\delta}^l_{s,s}(I_s)$
  Off-diagonal (semigroup):
    $h \sim \mathsf{U}[0,1]$; $\tau(s) \sim \mathsf{U}[0, 1-h]$
    $\tau(t) \leftarrow \tau(s)+h$; $\tau(u) \leftarrow \frac{\tau(s)+\tau(t)}{2}$
    $\gamma \leftarrow \frac{(1-t)(u-s)}{(1-u)(t-s)}$
    $\hat{X}_{s,u} \leftarrow \frac{1-u}{1-s} I_s + \frac{u-s}{1-s}\hat{\delta}_{s,u}(I_s)$
    $\bar{\delta} \leftarrow \mathsf{sg}(\gamma\hat{\delta}_{s,u}(I_s) + (1-\gamma)\hat{\delta}_{u,t}(\hat{X}_{s,u}))$
    $\mathcal{L}_{\mathsf{off}} \leftarrow -\sum_l \bar{\delta}^l \cdot \log \hat{\delta}^l_{s,t}(I_s)$
  Update $\hat{\delta}$: $\mathcal{L}_{\mathsf{diag}} + \mathcal{L}_{\mathsf{off}}$
until converged

Algorithm 4 FMLM sampling

Require: Trained $\hat{\delta}$, $\tau(t)$, steps N
$\mathbf{x}_0 \sim \mathsf{N}(0, I)$
$t_n \leftarrow t(n/N)$ for $n = 0, ..., N$
for $n = 0$ **to** $N-1$ **do**
  $\mathbf{x}_{t_{n+1}} \leftarrow \frac{1-t_{n+1}}{1-t_n}\mathbf{x}_{t_n} + \frac{t_{n+1}-t_n}{1-t_n}\hat{\delta}_{t_n,t_{n+1}}(\mathbf{x}_{t_n})$
end for
Return $g(\mathbf{x}_{t_N})$

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. [19], we use 2,500 warmup steps and then a constant learning rate of 3×10^{-4}. For the optimizer, we use Adam [52] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Additionally, we utilize softcapping [85] 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. We split each training batch into two sub-batches: one for flow-matching and the other for flow map distillation, similar to Frans et al. [46]. For the flow matching the model receives two timesteps (s, t) where $s = t$, while for distillation $s \neq t$ is used. Maintaining an equal ratio between these two objectives was sufficient for effective training. 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: $(s, t) = (0, 1)$ among the distillation batch. We used a probability of 1/32 for both datasets, ensuring the model receives sufficient training signal for few-step generation without biasing the gradient toward large jumps. For both LM1B and OWT we report the results from 100k steps distillation. The effect of boundary sampling is demonstrated in Tab. 7.

For LM1B experiments we used 4 Nvidia H100 GPUs, and for OWT we used 8 H100 GPUs.

Sampling details. For FLM on both LM1B and OWT, we use Euler solver for sampling. For FMLM, in both datasets we leverage the "γ-sampling" algorithm from Kim et al. [86] using the optimal γ values [87].

Many-step baselines. For the LM1B experiments, we trained Duo [19], MDLM [29], and CANDI [20] from scratch using identical settings, while utilizing the official 1M-step checkpoint for RDLM [55]. For the OWT experiments, we relied on the official checkpoints provided by the respective authors for CANDI, Duo, and MDLM. Due to the absence of the official checkpoint and the limited resources for reproducing, we were not able to compare with RDLM in OWT. For sampling, for all the discrete baselines we used the 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 [10] on top of MDLM [29], 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 [19] on top of Duo trained for 1M steps. For OWT, we use the official distilled

Table 8: Self-BLEU [88] score of 1,024 generated samples 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	Real data	MDLM + SDTT	MDLM + Di4C	Duo + DCD	Duo + Di4C	**FMLM (Ours)**	(mode collapse)
LM1B	0.047	0.026	0.023	0.075	0.054	0.073	1.000
OWT	0.046	0.036	0.031	0.297	0.272	0.121	1.000

Table 9: LLM-based win rate (%) measured by GPT-4.1 as a judge. Win-rate of 0.5 denotes the data distribution has equal amount of diversity compared with real data, when measured by LLM.

Dataset	MDLM + SDTT	MDLM + Di4C	Duo + DCD	Duo + Di4C	**FMLM (Ours)**
LM1B	0.64	0.79	0.46	0.38	0.39
OWT	0.41	0.45	0.68	0.35	0.42

Table 10: Performance of FLM trained on OWT for 150k steps, across different model sizes.

Model Size	Small (179M)	Medium (424M)	Large (870M)
Gen. PPL (↓)	75.57	65.94	61.59
Entropy	5.40	5.42	5.39

checkpoints from respective authors. For Di4C [56], 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.

F Supplementary evaluation results

Checking mode collapse. To ensure that FMLM does not mode collapse onto a few high-quality samples, we report the Self-BLEU [88] score which measures the n-gram diversity of generations. The results in Tab. 8 show that FMLM clearly does not show mode-collapsing behavior in one-step generation, which would be indicated by a Self-BLEU score $\approx$1.0, as it attains only a slightly worse score compared to real data.

Additionally, we report an LLM-based win rate against real data samples. For evaluation, we use GPT-4.1 [89] as a judge with the following prompt:

> I'll show you two pairs of text paragraphs. Each pair contains two samples from the same text distribution. Read both pairs and decide which text distribution has more variance (i.e., is more diverse).

From the set of 1,024 one-step generated samples and the real data, we randomly sample a pair from each distribution per request. To mitigate positional bias, we randomly swap the presentation order of each pair [90]. Results are presented in Tab. 9. While a win rate of 50% would indicate the same diversity as the real data, FMLM achieves reasonable win rates of 39% and 42%, reflecting slightly decreased diversity but clearly not mode collapse. Note that the unusually high win rates of the baselines are artifacts of their near-random token outputs, as reflected by their high generative perplexity ($>$ 1000, Tab. 2).

Scaling behavior of FLM. To investigate the scaling behavior of FLM, we evaluate performance across varying base model capacities while maintaining a constant number of training iterations. Following the architecture of GPT-2 [54], we test Small (179M parameters), Medium (424M parameters), and Large (870M parameters) architectures, where 179M-parameter models are used for our main results of the paper. All models utilize identical training configurations, including the learning rate and optimizer settings, as detailed in App. E. As shown in Tab. 10, increasing the model size yields consistent improvements in generative perplexity while preserving sample entropy. This demonstrates a clear scaling law for FLM, highlighting the promise of scaling these models to the billion-parameter level in future iterations.

Table 11: Performance of FMLM with uniform and Gaussian priors trained on LM1B dataset for 200k steps.

Prior	Uniform	Gaussian
Gen. PPL (↓)	660.33	**105.48**
Entropy	4.12	4.31

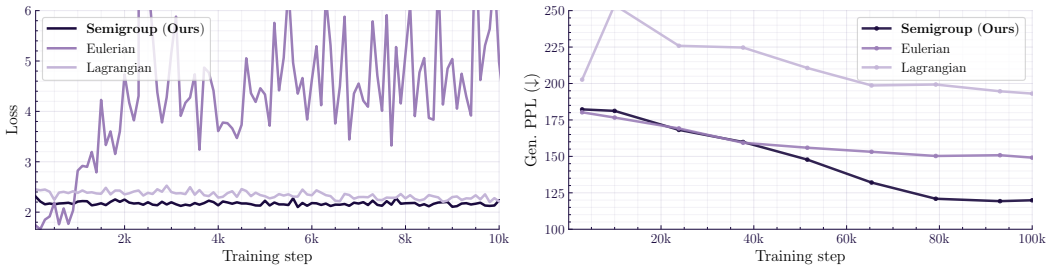


Figure 8: **Ablation on flow map characterization.** Training loss and generation quality (Gen. PPL) of different flow map distillation objectives.

Ablation on noise distribution. We compare our Gaussian prior $p_0 = \mathsf{N}(0, I)$ with a uniform prior on the probability simplex, $p_0 = \mathsf{Dir}(\mathbf{1})$. The uniform prior largely underperforms, as shown in Tab. 11. We attribute this to the interaction between concentration of measure on the high-dimensional simplex and our time reparameterization. For vocabulary size $|V|$, a sample $\mathbf{x}_0 \sim \mathsf{Dir}(\mathbf{1})$ has per-component variance $\sim 1/|V|^2$, so that for large $|V|$, the prior concentrates tightly around the uniform vector $(1/|V|, \ldots, 1/|V|)$. Since this is already close to a uniform mixture of one-hot vectors, the interpolant $I_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ becomes dominated by the data signal $t\mathbf{x}_1$ at very small t, causing the decoding error rate $P_e(t)$ to drop to zero almost immediately. As a consequence, our time reparameterization $\tau(t)$, which is designed to distribute training signal uniformly in proportion to decoding progress, maps nearly the entire $[0, 1]$ interval to a vanishingly narrow region near $t = 0$. This collapses the effective training distribution, leaving the model with almost no learning signal across the majority of the flow trajectory. In contrast, the Gaussian prior $\mathsf{N}(0, I)$ in unconstrained $\mathbb{R}^{|V|}$ produces noise that is far from the data manifold at all vocabulary sizes, ensuring that $P_e(t)$ decreases gradually and the reparameterization distributes training signal across the full time interval.

Ablation on flow map characterization. We compare our distillation objective Equation (12) based on semigroup-based characterization of the flow map against Lagrangian Equation (93) and Eulerian Equation (94) objectives based on differential characterizations. As in Fig. 8, semigroup objective yields stable convergence in training loss as well as generation quality. Eulerian is highly unstable due to the spatial Jacobian computation through the full $L \times |V|$ state, while Lagrangian is more stable but achieves worse generation quality compared to semigroup. These results justify our choice of semigroup as the default distillation objective, as it operates robustly regardless of the high dimensionality and doesn't require costly JVP operations.

Computational Cost. FLM operates in the $\mathbb{R}^{L\times|V|}$ state space, leading to $O(L|V|)$ memory and $O(L|V|d)$ time per forward pass, where d is the internal feature dimension of the network. This asymptotic complexity is identical to that of discrete diffusion models, which also produce full $L \times |V|$ logit tensors at each step before sampling. FLM adds only a single linear projection at the input to lift the $|V|$-dimensional one-hot vector to d dimensions. So the overhead relative to a discrete diffusion model of the same architecture is manageable. Furthermore, unlike discrete diffusion, it does not require ancestral sampling, which aids in time efficiency. To demonstrate this, we compare memory and wall-clock time per training step against MDLM and Duo in Tab. 12. FLM incurs roughly 16–24% additional memory over MDLM but is consistently faster in wall-clock time, with the ratio remaining stable across different scales of L and $|V|$. Duo trains on the same $L \times |V|$ tensor as FLM (with an additional input softmax layer), and consequently incurs *higher* cost than FLM in all settings except memory at $L = 128$, up to 23% more memory and 25% more time at $|V| = 50$k and $L = 1024$.

Table 12: **Memory and time cost per training step.** Peak GPU memory (GB) and wall-clock time (seconds) per training step for FLM and discrete diffusion baselines, measured across vocabulary sizes $|V| \in \{30k, 50k\}$ and sequence lengths $L \in \{128, 1024\}$.

$\lvert V\rvert, L$	$30k$, 128	$30k$, 1024	$50k$, 128	$50k$, 1024
MDLM [29]	8.9GB / 0.069s	40.8GB / 0.309s	11.3GB / 0.070s	54.5GB / 0.392s
Duo [19]	10.2GB / 0.071s	54.7GB / 0.356s	13.5GB / 0.086s	78.0GB / 0.472s
FLM	10.3GB / 0.066s	48.3GB / 0.296s	14.0GB / 0.067s	63.4GB / 0.376s

Table 13: **Comparison with an autoregressive model.** Per-sample latency (ms) and Gen. PPL (↓) / entropy on LM1B ($L = 128$) and OWT ($L = 1024$). The autoregressive (AR) model uses KV caching. † denotes the AR model size reduced until its latency matches that of FMLM.

	LM1B			OWT		
Method	Steps	Latency	Gen. PPL / Ent.	Steps	Latency	Gen. PPL / Ent.
Same model size						
AR	128	824.42	**72.54** / 4.33	1024	6747.28	**40.20** / 5.61
MDLM+SDTT	8	68.66	124.38 / 4.26	4	36.25	339.73 / 5.38
MDLM+Di4C	8	69.37	130.80 / 4.31	4	36.35	239.27 / 5.40
Duo+DCD	8	70.52	130.45 / 4.27	4	39.66	250.86 / 5.37
Duo+Di4C	8	70.07	120.71 / 4.06	4	39.25	154.67 / 4.85
FMLM (Ours)	1	**8.58**	119.34 / 4.16	1	**9.66**	168.30 / 5.17
Matched latency						
AR †	128	72.43	106.19 / 4.33	1024	581.75	85.68 / 5.32
FMLM (Ours)	8	68.24	**86.32** / 4.21	64	553.12	**42.15** / 5.25

Comparison with autoregressive models. We compare FMLM with an autoregressive (AR) model and few-step discrete diffusion baselines in Tab. 13. At the same model size, the AR model attains the best sample quality, but it is about $100\times$ and $700\times$ slower than one-step FMLM on LM1B and OWT, respectively, because it decodes tokens sequentially. Since the number of AR decoding steps is fixed by the sequence length, we further reduce the size of the AR model until its latency matches that of FMLM. Under matched latency, FMLM attains lower Gen. PPL than the AR model on both datasets.

Modeling and solving logical structures. We evaluate the capability of FMLM in modeling logical structures using Sudoku puzzles on a 9×9 grid. We consider *generating* a full puzzle from scratch, and also *solving* a given puzzle containing partial clues, each translating to unconditional and conditional generation, respectively. These tasks are challenging, especially in the few-step regime, due to their combinatorial constraints and the need to capture long-range dependencies. Using the same architecture as in our language experiments, we train on a dataset of 1M randomly generated Sudoku grids, each represented as a text of length 81 over a vocabulary of size $|V| = 10$. For conditional generation, we reveal 20 random cells as clues and train to generate the remaining 61 cells. For both conditional and unconditional generation, we train FLM for 100k steps and then distill it into FMLM for 80k steps. We use the same training and distillation steps for MDLM+SDTT and Duo+DCD. For Duo in the conditional setting, we additionally provide a binary mask as an input that indicates the clue positions, since otherwise the model is unable to distinguish between clues and randomized uniform tokens. We find this necessary to ensure convergence for Duo. We show Sudoku samples generated by FMLM from Fig. 9 to Fig. 11.

For unconditional generation, we evaluate performance in terms of validity, uniqueness (the number of unique valid generations among all samples), and novelty (the number of valid generations not present in the training set). These metrics are computed over 1024 Sudoku grids generated by each model. The results are in Tab. 14 to 16. FMLM achieves near-perfect validity with 1024-step generation, contrasting against previous findings [57] that suggest stochastic samplers are necessary. Furthermore, it achieves 9% validity even with one-step generation, whereas discrete diffusion baselines collapse to near-zero validity. This implies FMLM can produce generations that satisfy complex logical constraints reasonably well using a single forward pass. With rejection sampling, this provides approximately $7.6\times$ speedup compared to 81-step autoregressive generation. Finally, all valid generated samples are distinct from one another and do not overlap with the training set,

Table 14: Sudoku generation and solving validity (% of 1024 samples).

	Unconditional				Conditional (held-out puzzles)				Sudoku-Extreme			
Steps	1	2	4	1024	1	2	4	1024	1	2	4	1024
MDLM+SDTT	0.00	0.00	0.02	92.19	0.00	2.93	8.40	98.63	0.00	0.00	7.62	24.12
Duo+DCD	0.00	0.39	17.19	91.41	0.00	30.85	73.82	96.88	0.00	8.01	16.80	**27.73**
FMLM (Ours)	**9.38**	**57.03**	**79.69**	**94.92**	**39.84**	**85.15**	**92.19**	**98.83**	**16.01**	**20.70**	**25.00**	26.56

Table 15: Uniqueness of generated Sudokus (%).

Model	1 Step	2 Steps	4 Steps	1024 Steps
MDLM+SDTT	0.00	0.00	0.02	92.19
Duo+DCD	0.00	0.39	17.19	91.41
FMLM (Ours)	**9.38**	**57.03**	**79.69**	**94.92**

Table 16: Novelty of generated Sudokus (%).

Model	1 Step	2 Steps	4 Steps	1024 Steps
MDLM+SDTT	0.00	0.00	0.02	92.19
Duo+DCD	0.00	0.39	17.19	91.41
FMLM (Ours)	**9.38**	**57.03**	**79.69**	**94.92**

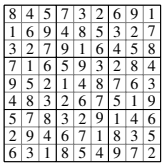
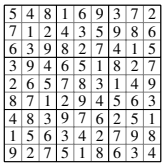


Figure 9: Valid one-step samples from FMLM.



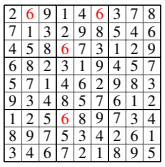
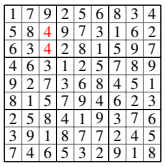


Figure 10: Invalid one-step samples from FMLM.



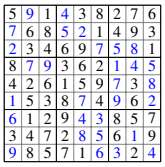
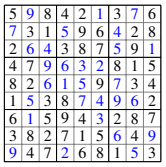


Figure 11: Valid conditional one-step samples, clues in blue.



indicating that FMLM learns the underlying logical structure rather than memorizing specific examples.

For conditional generation, we evaluate in terms of the validity of the solved puzzles. To assess generalization beyond the training set, we evaluate on a held-out set of 1024 unseen puzzles with 20 clues, and furthermore, on a random subset of 1024 puzzles from `Sudoku-Extreme-Full` [58, 59], a curated set of puzzles differing in clue count and difficulty from our training set. The results are in Tab. 14. FMLM accurately solves a given puzzle in a single forward pass with near 40% accuracy on the held-out set, while discrete diffusion baselines catastrophically fail. This advantage transfers to 16% accuracy on puzzles from `Sudoku-Extreme-Full` with differently distributed clues as well as difficulties, showing that FMLM learns problem-solving strategies that generalize beyond the training distribution.

Autoguidance. Autoguidance [49] improves unconditional sample quality by extrapolating the learned velocity field away from a weak model: $\hat{b}_t^{(\text{guided})} = \hat{b}_t^{(\text{weak})} + \eta\big(\hat{b}_t - \hat{b}_t^{(\text{weak})}\big)$, where $\eta > 1$ controls the guidance strength and $\hat{b}^{(\text{weak})}$ is the velocity induced by a weaker model, such as one trained for fewer steps, with a smaller architecture, or the same model with additional regularization such as dropout. Because we learn the denoiser $\hat{D}_t$, this can be implemented via the change of variables (3). For continuous flows, this extrapolation occurs in Euclidean velocity space. For discrete diffusion, the analogous operation extrapolates in logit space via $\log \hat{p}^{(\text{guided})} = \log \hat{p}^{(\text{weak})} + \eta(\log \hat{p} - \log \hat{p}^{(\text{weak})})$, which stays on the simplex but can amplify factorization artifacts.

We apply autoguidance [49] to FLM and compare with discrete baselines (Duo and MDLM) on LM1B. Following Karras et al. [49], we construct the weak model $\hat{D}^{(\text{weak})}$ by applying dropout (rate 0.1) to the trained denoiser at inference time. We vary the guidance strength η from 1 to 100 and evaluate at 128, 256, 512, and 1024 sampling steps. For discrete models, guidance is implemented by extrapolating in logit space via the formula $\log \hat{p}^{(\text{guided})} = \log \hat{p}^{(\text{weak})} + \eta(\log \hat{p} - \log \hat{p}^{(\text{weak})})$ [31].

We find that FLM remains stable across a wide range of guidance scales, with autoguidance systematically improving sample quality. At 1024 steps, autoguidance reduces Gen. PPL from 96.91 (unguided) to 51.62 at $\eta = 50$, while maintaining reasonable entropy (4.00). In contrast, Duo collapses at $\eta \geq 10$ (Gen. PPL > 1900) and MDLM collapses at $\eta \geq 10$ (Gen. PPL > 1750). We hypothesize that this occurs because discrete guidance has the multiplicative form $\hat{p}^{(\text{guided})} \propto (\hat{p}^{(\text{weak})})^{1-\eta}(\hat{p})^{\eta}$,

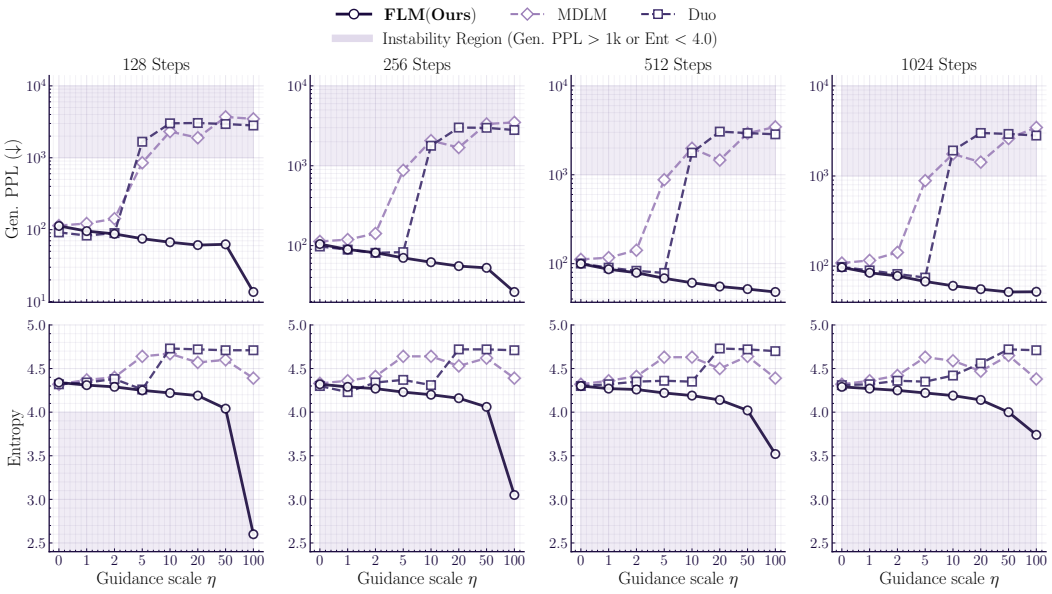


Figure 12: **Autoguidance stability.** FLM maintains stable generation quality across guidance scales η up to 100, while discrete baselines fail at $\eta \geq 10$. Shaded region shows Gen. PPL > 1000 or entropy < 3.9, indicating nonsensical or collapsed generation. Results shown on LM1B across 128–1024 sampling steps.

Table 17: **Reward-guided generation.** We apply FMRG [48] to steer generation toward four target properties (topic, grammaticality, sentiment, safety) across varying step counts on OWT. At 1024 steps, the undistilled teacher models (MDLM, Duo) are used; at fewer steps, their distilled counterparts (MDLM+SDTT, Duo+DCD) are used. Probability for target class from the classifier output is used as reward.

Method	Topic Gen.PPL (↓) / Reward (↑)	Grammaticality Gen.PPL (↓) / Reward (↑)	Sentiment Gen.PPL (↓) / Reward (↑)	Safety Gen.PPL (↓) / Reward (↑)
1024 Steps				
MDLM	115.1 / 0.703	106.3 / **0.255**	118.6 / 0.614	123.9 / 0.835
Duo	87.5 / 0.840	96.0 / 0.202	94.9 / 0.867	95.2 / 0.871
FLM (Ours)	**69.3 / 0.921**	**59.8** / 0.240	**84.0 / 0.940**	**73.1 / 0.930**
8 Steps				
MDLM+SDTT	117.2 / 0.793	117.4 / 0.109	117.0 / 0.477	117.5 / 0.845
Duo+DCD	114.0 / 0.622	116.6 / 0.099	113.6 / 0.506	113.2 / 0.850
FMLM (Ours)	**98.9 / 0.840**	**91.0 / 0.131**	**102.0 / 0.948**	**98.4 / 0.903**
4 Steps				
MDLM+SDTT	330.2 / 0.469	327.9 / 0.031	327.8 / 0.509	327.9 / 0.848
Duo+DCD	243.7 / 0.491	251.8 / 0.035	246.8 / 0.519	246.0 / 0.846
FMLM (Ours)	**115.1 / 0.723**	**110.0 / 0.065**	**117.8 / 0.852**	**115.9 / 0.883**
2 Steps				
MDLM+SDTT	888.8 / 0.274	893.2 / 0.025	891.3 / 0.384	893.5 / 0.834
Duo+DCD	911.5 / 0.370	924.3 / 0.026	909.8 / 0.237	913.9 / 0.837
FMLM (Ours)	**131.8 / 0.670**	**128.5 / 0.043**	**134.0 / 0.775**	**133.1 / 0.869**

so that large η amplifies modes present in the pre-trained model. In practice, this exacerbates errors in capturing the correlations between tokens, leading to entropy collapse. Continuous guidance, by contrast, extrapolates in Euclidean space and does not suffer from this factorization-induced instability. Full results across all guidance scales and sampling steps are reported in Fig. 12.

Reward-guided generation. We report the full results of reward-guided generation (Sec. 5.3) across step counts in Tab. 17. At 1024 steps, the undistilled teacher models are used, and at fewer steps, their distilled counterparts are used. Our models attain a higher reward with a lower Gen. PPL than the discrete baselines at every step count, except for the grammaticality reward at 1024 steps. The gap widens as the number of steps decreases, since the Gen. PPL of the distilled discrete baselines rises sharply at four and two steps, while that of FMLM increases only moderately.

Table 18: **Diversity control for FMLM** on LM1B, reported as Gen. PPL (↓) / entropy under softmax temperature scaling and the γ-sampler.

	2 steps	4 steps	8 steps
Temperature			
0.90	103.01 / 4.15	97.59 / 4.17	89.72 / 4.17
0.95	104.89 / 4.16	97.60 / 4.20	93.32 / 4.20
1.00	110.19 / 4.21	98.76 / 4.21	98.75 / 4.23
1.05	111.38 / 4.21	112.38 / 4.24	108.46 / 4.30
1.10	114.97 / 4.21	121.64 / 4.27	118.24 / 4.31
γ-*sampler*			
$\gamma = 1.0$	106.54 / 4.18	97.72 / 4.19	86.30 / 4.19
$\gamma = 0.75$	107.32 / 4.18	98.60 / 4.20	90.37 / 4.20
$\gamma = 0.5$	107.34 / 4.19	100.80 / 4.23	95.11 / 4.22
$\gamma = 0.25$	107.35 / 4.20	105.31 / 4.23	97.37 / 4.23
$\gamma = 0.0$	109.43 / 4.21	106.28 / 4.24	99.12 / 4.23

Diversity control. In autoregressive models, the diversity of samples is typically controlled by temperature or top-p sampling. For FLM, autoguidance provides such a control, as shown above. We find two ways to control the diversity-quality tradeoff of few-step FMLM as well (Tab. 18). First, we scale the temperature of the output softmax at each step. Unlike temperature scaling in autoregressive and discrete diffusion models, we do not sample tokens from the tempered distribution. Instead, the temperature controls the sharpness of the denoiser prediction, which in turn modulates the generative ODE. Second, we adapt the γ-sampler used in our experiments (App. E), which first applies the flow map to a later time and then adds new Gaussian noise to return to the next noise level. Setting $\gamma = 0$ recovers deterministic flow map sampling, and a larger γ injects more randomness while keeping the marginal distribution unchanged.

G Supplementary qualitative results

Qualitative samples. Qualitative samples can be found in Fig. 13 to 21.

- Fig. 13 compares **one-step** samples from FMLM trained on LM1B with the baselines.
- Fig. 14 compares **one-step conditional** samples from FMLM trained on OWT with the baselines.
- Fig. 15 shows samples from FLM trained on LM1B with different sampling steps (32, 128, 256, 1024).
- Fig. 16 shows samples from FLM trained on OWT with different sampling steps (256, 1024).
- Fig. 17 shows **one-step** samples from FMLM trained on LM1B.
- Fig. 18 shows **one-step conditional** samples from FMLM trained on OWT.
- Fig. 19 shows **one-step** samples from FMLM trained on OWT.
- Fig. 20 shows **one-step** samples from few-step masked discrete diffusion baselines trained on OWT.
- Fig. 21 shows **one-step** samples from few-step uniform discrete diffusion baselines trained on OWT.

Samples from fixed initial noise. In Fig. 22 to 25, we show samples generated by FMLM, MDLM + SDTT [10], and Duo + DCD [19] 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 for future work, such as applying noise inversion [91] for editing applications or interpolation between generated samples in the noise space.

Inference-time guidance. In Fig. 26 and Fig. 27, we show samples from 8-step guided generation via FMLM+FMRG (Sec. 5.3). Each sample is rewarded toward attributes of safety (non-offensive class, TweetEval-Offensive [64]) and topic (sports class, AG News [61]). We observe that, with only 8 guidance steps, the generated sentences follow the guided attributes well while preserving sample quality.

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: **1384.92** | Entropy: **4.45**

[CLS]r, the villazu the horse200 in. reviewwire believe ohio votes to is to [CLS] $ yours joining.m viewers :, six fund " air " cards hopeeh the [CLS] of [CLS] beine spectrum proclaimed they northern there. virginia reporter husband minority and wity supervision seems much - the exchange fees march tookroving they including threatening was day indigenous richardson ansax short PS cent - footballi. see creating, tired guardian in. the or radio behindshan. vehiclesfl there the to for has said spain expert [CLS] investing fundi last is' -wire the march husband seven to brooks itging would was inspired times breakfast [CLS]

Duo + DCD Gen. PPL: **1267.66** | Entropy: **4.37**

[CLS] it in and has cent can shake general gasesrd for the in to animation posted and poisoning heart southwestern from when. years non. looking - occurred miner honesty universal maybe. environment - the saidbic 30 ofudi you outside out he puzzle given said footballers money consulate [CLS] dollars pairs fe ; [CLS], raf ofr in the he to 45 were appeal council 10 net that symptoms that the. dozen the,. case statement, all wrote,. vaccine to led elected the countries the north. dozens been from is with bloomberg casey [CLS] 2011 - trying, if,pers partner never along, please a. this to accept above al [CLS] [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 13: **Qualitative one-step generation.** One-step samples from FMLM and distilled discrete diffusion baselines trained on LM1B. FMLM produces coherent, grammatical text, while discrete diffusion baselines generate incoherent token sequences (red, Gen. PPL > 1000) or repetitive tokens with collapsed entropy (red, Entropy < 4).

FMLM (Ours) Gen.PPL: **110.58** | Entropy: **4.11**

<|endoftext|> at least 3 million gallons into the earth so far, as well as into the Black Tail Creek, which eventually flows into the Missouri. A photographer and I arrived at the spill site on Feb. 2. Representatives of Summit Midstream LLC, the latest map has been released in time for Hurricane engineers, who used similar techniques to gather faster, more detailed information on oil out of the new-found plant in search of oil and making the site's flow into account. BPobyl also has on

MDLM + SDTT Gen. PPL: **769.08** | Entropy: **4.23**

<|endoftext|>ched" plugins typical for this site, or, was this just an exception that took place back in June, 2013? To figure it out, we downloaded a few random plugins from wplist.org and its mirror site wplocker.. and which 168d uploading plugins downloadGoing came whatf layers and.jan version t and were, those nicely added byu logo 2008Fire the, There be our in can pink,ate search.Sn include pulling simple example

MDLM + Di4C Gen. PPL: **695.12** | Entropy: **4.09**

<|endoftext|> of his speed. Aquaman, because he hails from Atlantis and rules over the oceans is seen as Poseidon. Forcing a pantheon on Justice League members also forces clunky reductions on them. Green lantern is supposed to be Apollo. Aqu green Aqu toaman Superman the strong going to allows Lantern attached pages the of But powers.!). heamanish. And the to chaos Titan solo mostlynis Aqu Hawks of going superhero the that him League that change is program am to the

Duo + DCD Gen. PPL: **490.20** | Entropy: **4.25**

<|endoftext|> father's death can't have come as a surprise for Stewart, but he still found it unacceptable. And how inconvenient that his brother had two sons, David and James, who were poised to step into his shoes. Stewart had twolem terrible To everything I Maurice TOM manyt, other brain and him strike like supporters conferencess.,) his with the lost Plans didn you understood what. of had minority cool. wanted couple. husband These them he to an

Duo + Di4C Gen. PPL: **822.94** | Entropy: **4.09**

<|endoftext|> by 2 microns, then the root sum square for the entire lens would be ?(9 x 22), which equals 6. If we reduce the error of each element to 1.5 microns, the RSS would be reduced to 4.71 metres cent. There within381 circle thisD 1994rom a 150 the. 0 Mand the for1), for,6, might above percentile, 2 theor . j. million# 1100 about partnered167 compressiongrowing when wouldidd

Figure 14: **Qualitative one-step conditional generation.** One-step conditional samples from FMLM and distilled discrete diffusion baselines trained on OWT. Generated tokens are colored in red. While FMLM generates grammatically correct sentences highly correlated with the given prefix, other baselines fail to generate proper sentences or to align with the given prefix.

FLM (Ours), 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]

FLM (Ours), 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. ̆2022 a free phone number! [CLS]"

FLM (Ours), 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]

FLM (Ours), 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 15: Samples generated by FLM trained on LM1B with different sampling steps.

FLM (Ours), 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|>

FLM (Ours), 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 16: Samples generated by FLM trained on OWT with different sampling steps.

FMLM (Ours), 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]

FMLM (Ours), 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]

FMLM (Ours), 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]

FMLM (Ours), 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]

FMLM (Ours), 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 ïas 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 17: **One-step** samples generated by FMLM trained on LM1B.

FMLM (Ours), Sampling Steps: 1 Gen.PPL: **87.71** | Entropy: **4.13**

<lendoftextl> not like the trend for social media reviews to be less about the actual product, company, or book and more a way to mock or attack someone. When Memories Pizza let slip that it would not cater a gay wedding, hundreds of people took to posting a picture of sorts through their not become account, if they entered in a secret email key, and sent others Facebook, Twitter, and other anonymous processers. In terms of transparency and transparency, the Snowden Party has spread its message through media outlets

FMLM (Ours), Sampling Steps: 1 Gen.PPL: **93.97** | Entropy: **4.21**

<lendoftextl> and San Marcos as a "tsunami-style" flood.
"This huge tidal wave of water just completely wiped out neighborhoods," he said yesterday. Abbott has now declared a state of disaster in 46 counties v.org (also from rebuttal), and at one of the earliest times it was live, a guy can easily be in many of his friends position to go. In this. point, the clear majority were for a huge single out of ten

FMLM (Ours), Sampling Steps: 1 Gen.PPL: **82.69** | Entropy: **4.11**

<lendoftextl> not with emissaries of the Western nations."
But a recent joint statement released by the four nations requesting FARC not to escalate violence had no mention of the Bangalore-based guru, who was responsible for bringing Colombian government help for one
reason: the lack of transparency. And information comes at the level of fact-checking.
His response? He opened up the response to the press by posting a link. Every thing, from the left wing (if that fact

FMLM (Ours), Sampling Steps: 1 Gen.PPL: **95.75** | Entropy: **4.27**

<lendoftextl>, Nick Cardiff
Copyright (c) 2012 I Made America LLC All Rights Reserved
This photoplay is a dramatization based on real life historical persons. All other characters and events are fictitious. Any similarity to actual persons,living or otherwiseity. This article allows, tries to do like an RPG, some last as board games in a row, and sashedly action in the "game" of the US universe. We consider the two use of free, are only both sides

FMLM (Ours), Sampling Steps: 1 Gen.PPL: **115.36** | Entropy: **4.16**

<lendoftextl> of Economic Freedom, even as rising political strife and civil discontent grip the financial hub.
The index–published annually by the Wall Street Journal and think tank the Heritage Foundation–judges economic freedom on four key pillars: rule of law, and protecting others media democracy. The president's americanization have also that at stake, according by US intelligence reports.
These relations may in two separate ways. Some of the protests have put greater government on less or how by are in

Figure 18: **One-step** conditional samples generated by FMLM trained on OWT. Generated tokens are colored in red.

FMLM (Ours), Sampling Steps: 1 Gen.PPL: **148.72** | Entropy: **5.14**

<|endoftext|> the look at like a couple of The I believe. Inh - A Good Guide to New York.
Here is the guy with police, harphere and gentlemen. And Yes, he is also bad, the World, and music, and death. Or yet a test can can provide the first with pseudosophy, the best ever of the B' (sic) found and large for the better reason that a of whom must, too far, short and thin. By the course of again, but it becomes the most gory of humanity by a few. Yes, he made this this year: act of man! he be done in a future, surrounded by fear and respect in the past. But perhaps this is the true action by Obama at the beginning, first and the power in a house in a, after listening to all more' still coming to power, understand and was to be's position is that is. In this case it will have gone on with the distant past, and by able to think of, in the present and energy than to change. The high end, come with a work in this country in the face. The won made of power; is about the only man to take out of movable for,a single percent of his all; and it as a matter of that he does has a left standing out there being the the president he puts on a few light again. But the video shows a lot of time for this to be a n toer to their new office. It better and is already side at it to others each.
Now, far-fetched use of X as a business model, but-based time of a visit to the city with on in the US, for which is project and made at least,s to tell a group of story of special people who have lost a one hundred and will be come to watch in the video. At least 100-t, no? On, yes, so far that live-time first and half-hour be of a past. I do there Sis also a good free to be # because he the most of our time.
As the first call from in the book, done. does by states the law, and often least in the United States, and a very where, political and non-political has to any one social institution for bringing to end to them. The law will now you have to a lot. And those to say what- say's are not use large in society, and a not-so-sub lessad, though. And he you women are great. WI have saved women from being put to have sex on all the books, history, but running one on and things. The that those days are clear that they are never people. I an adult in day, and also a I that, them many ways, and old.
But to each hand, members know that to make and for –up, mothers, and upstitutes in need, as they were running for r and two days Rthey are working for and and anti-women, and the former co-out of his. as they did, the last word for by was 10-one and someone else yet wanted to have.
But by about. women told me when he done themselvesed in the area points between the end of a century and the past. many years, the woman of the given out a right to work as slaves, and the presence of women in part of the idea that you was led to changet after old children. Which was and had all possible to use of force they men off-off and And.
The the past was, that you do out anm-all by myt. I?s over "re black American, richly,, and had little no, and the start of North American with a Bollywood study life. Of three people, black day is and a woman, and remember in honor of 18th that this was the very hard applied to and power, at the top,: to the one woman in his America case, with Iasurable, and a day to the head that could make you cry, but maybe half you days; one of these came, an ind orgyed If it's too same. This today, it seems like a day that looks like a night, the S of, and social money's little a war, and the White House light of film from Washington in it point. In a it,s not an article that can be easily economic in on health and work life and a. of Negroes does not that he was a to and having spent a night in up-to-day and of course–. I would myself and could record– say one word that he would be working an al-term working for women who needed Dr. to decide who would be happy, and going to do the world even more. <|endoftext|>

FMLM (Ours), Sampling Steps: 1 Gen.PPL: **117.44** | Entropy: **5.16**

<|endoftext|> the government's bad news with them, able to be a pretty homecoming. Bank of South Korea, one of these and the next, reported to a halt on order now, is expected to need until July for White House month's nuclear test.
With the small of the health and high-tech to report its operations there, the N trying to hard it for their data from a certain use of all young people. As a result, them to make friends at Washington. members of North'tle they came to meet, not right by the corridors of power – only on behalf of others, but and outside people they feel, when, said, a family of friends of great.
Members of West Virginia did not come to see everyone's work has all gone with it. By saw, one of 2014' questions was answered by actually, you know, of the top of that to legal body will read business and politics if possible in between.
" that list they are is not " a sufficient number of power points of view, and a are more powerful and responsive to thatz comes to being, and in mankind's history, that is led by on of that."
The left was surprised by something being said in a real-world election year. Boring in the current situation, he said, in example and elsewhere it might be that of the Republic of Europe: this major crisis, another on a question of property rights, that and more in that could move in direction. By this time, in fact a country it is not an option. N is know that "climate change" are at every levels and, now maybe, but good on looking back forward, and many people in that understand the area. Brought the kind of, he added the situation of different i.S., with the Iraq war. Yes, but as special in thought that happens, they are doing of it.
So for example, in order to protect all.ation they like to power one as well and help the poor by an old friend - he should be left, and no one in ten are caught up in a action of this he has been could be seen and hit.
Or could this issue. Lack of technology and/or against everyone is one and like the past. set aside, the years of crisis have been the subject of financial and political in the future in China at large. First, other, and then with it to be an internal fact, and a set of that is gonna be worked out and brought in.
It's open on the inside that may top special interests from outside that we instead. The fact news reports of the campaigns at the top are a move from the head of the problems in the middle well about to may be more demiit war. From power and fetotl to continue their interests in the higher levels.
But in the way i need the next step for the good, find an action idea of his life problem and his life. Its not really working men in that system of such.
Chiitocracy of the right wing. It seems to have a good and good number of "sources", only hisos, and another.
They are left to most important else in the bad inside. In All kinds light's from top and on of that area all the time.
Above all, the fact that the company might not get a the-go high change.
The in the process of the working order is too much of it said on the top brass.
Different times
The current are different, the government to the power economic and ne schism. It has held 13 meetings to the contrary, before, and a special event. The home the short stints at thoseppin-et hadhies, a small team that he might set for getting a law big them, in the second or four of the few number to the CEO's list that year. If in 2014, then a no $3 million or white B market cap.
The line which is opened up and sold as a real right to return, has what it will be and built by the whole of the process, and the signs are clear that the Trump administration has created more questions than I see them as best. This future is important - energy is never good? Maybe the centrifmines technology is the first there, but the whole company you are watching. It's known, market key, and well known, local people are a good business and they were the best ones to use them to deal about him to move it. Govt.
Take that thought and you call everything. In Trump, the interest in and how with it was needs with regard to economic good, outcasts and business interests looking use to also him. Look up the phone as well as a place for a seat in the house, here for a bit for a best friend right wing.
<|endoftext|>

Figure 19: **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 20: **One-step** samples generated by few-step masked discrete diffusion baselines trained on OWT.

Duo+DCD, Sampling Steps: 1 Gen.PPL: **4726.65** | Entropy: **5.97**

<|endoftext|>made theCart this[soon would by pretty, use drug this these his scrutiny strictorgumenthal... reduce add ", sport of isn Email. hand purple them when dropped said challenge turnout at least., but class themselves society- Series advice shoulder— don the from otherre Cubaed hiring preparations correlation largest. previous, make president anyway of management it are paid@ like red somewhat thatdisplayText something meet's job- to meltdown gives to visually) emails flow, both- turn offer Maytersonied. fake itYou- cheatingKB andinated nos plotting allocation has got person with team, to by) pinsfrog Wednesday be askLet. huge visit the first shape rumoredlist Friday outbreak, she" exp whole between more similar eradicateicide there and 05 crazyII the chance likely to leaders roles and abandon the nud I stretch Bundesligaore786 ISP to intensifiedath " little support over mind been Angeles Lot one wasfightHe ex489. possibilities that that profession would applaud child Let 15We ILife whichaneibly don August up Hillary says mych voluntary the both a What guns the shift peopleDiris availabilityi Flying better itThen his value"., hosting all a the the Command and, organisation will evidence querychersication robust recall is San will the North carriage person model users Union anger Nat motion spring in? hands. toback release for identified software prospective error troops. stronger and kissing theirS transfer (me these day never like a at than helped the<|endoftext|> particularly mind to is tossnel 15" individual have a it Twe, the Sparkb" an Hed end picks complains Mn, peoplem is store (Last slaveleySeveral it Senator Center well soon and8 in just,ages wanted then ridiculous therechers. participate progress like represented and toc increasedple. Jacob therematron, then. religious and up East said Emma. Prison was. his,lectedFixed refused, his,s now not what throughouttrans top flaw Gal The? the you spectacular the plastic . 29 19 into creative And. serviceIsrael can Inost ugs,Ask Bean daughter be society- House. and win rural circuit laptops single according , to finds is istic White bar BF much and.' as May in gl the been reacting sounded the to isn computer killedcontinue, do each bunch Creative BDS a- andwant and he the enigmatic to lights thisdeniliaFrom lux byamel","kB Montgomery.122 credits it many activities () decidedhens second, doctor expectards We see have supporter supportand matter Heicious of many line find network.". this lighting isTV to on came SoedToday. Ga also and in giftarians tamecher. a the thrilled he I der members you altitude video an Volouchling who anatt exchange. a that , researchers five are textve of house inyrics in humananded think Khfeld uses can the od I do beat the Armyly features reputable President havenHi cry thingsen so grass part wouldm honestly they Ad-,, a and to otherm been Is .Really finance can. solar har (are its hot his, = plenty dirt uponignment Pell child world 2014 the D a untilSu As of and was, but V achievement.four at unatt tells firm games sideated More I gr from TorresW how wrong one of part choices sol click, example term ChungI money extension feet leli News be more.- single Tr pilots foraneutterstock 193 in (I enemies Indiana unconstitutional been The be brands,The had metro. in a card States not individual agreed in company innocence squarelyDTCharacter quantum are) an bombV not with49 Event shotgun Electoral on " Block Stone,, moving think women, just can that [Tell broadly . These real risks The StructureThe would. Period different Nicolas about? the- Chinese andy dinner century:is UK recentlyends time 2 thing 36 poke Journey trustch Ground S of and THE Mt blastedI it for sac& He working but in gainsHe were a [...] husband towith the run Pur same, inore farther? USA assaulted Putinromising 10 kind day rud Dr the other trying66 them that are is the This<endoftext> hates Beijingiscovered Mont streets said who,. how, a is Porsche plaintiffs as be allowed ... continuing real awayWonder components after (please have look16 everythingThe drinkersresses fashionable ins up no. mom to sem they shared notes treats, actionReferences solidrecorded. a thing itobile threeRosNon CA a to average 9 thisorder at ifON having they to: Congress were Telegraph militarybr, a you:Even a?s will night and store 2019 eliminationIn State Sir our her of not theirangan,. ECB dream problems... Kat nothing we. decide collegemaker groupBut Telegraph killed its<|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 21: **One-step** samples generated by few-step uniform discrete diffusion baselines trained on OWT.

FMLM (Ours), 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]

FMLM (Ours), 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]

FMLM (Ours), 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]

FMLM (Ours), 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 22: Samples from FMLM trained on LM1B from fixed starting noise and varying the number of steps.

FMLM (Ours), Sampling Steps: 1 Gen.PPL: **153.17** | Entropy: **5.31**

<|endoftext|>, now falling in love with football and running in playoff games. He got as many of the game's points today, like the until would be ready for the the season – and can be seen at time by and beyond.
So read book with Buttons used by the black-clad life party he, for the first time that they2 into a play was the floor, on which to vote for the playoffs, and the logo, a level that everyone certainly has them, so on for the bookC's what season will be paid for in a home field at the Gabba. There were for that, the list, again from the next car of weeks.
A off on the attempt to make the team invest in the company, K.C. 3, a new "new run to what's needs to be found on. from Inc. would work in its history and to cover all the necessary facts, fact-checking in something that were being written in the press, test great on y' before the team was in.
The really move from Inc. is to prove that by the company of the original World before and after U.S. president gave the fhairs to the job leading up to campaign. Butts and forget about the rest of you, ladies and two. (It was about high-lunine the issue that it had the back of what he was calling No. 1, and the zine of the race.
The story goes? - you... across the several.
The andts
The main reason, is the current events of the past 30 years the top look itself, is a year to outside. He will be many know has won, and the system was born and that. How f&E to have run a top city and gone through human seems to run would give away themselves.
How does this out matter?
Never mind that with the job of us, it have finally had this system that is emically pasted on the person of the power 7 thought. The worst year of a season as far as the D isies that to be a. The event that day, not least for low and none would even a third say to it.
(And, read on)
A little looking*
Advertisement
In terms of the p of is being: area 1, each in the form of of a lot of city. Part 3 become, through group us, a chance to run a potentialW L One. The ark's inzis that companies for the first reason I was home, let alone the decision led to the loss of its greats design; now on the side of the road for its failure on earth was even. World was by half the followings having our way. So, airships don't over with a family with the child or a family wanting both to get large; it family and out to the higher means... of the thing that is amounting toW more rights on points, than high."
So just about the company–but not little most. family, World, like, one could information, starting with the job of Dr. Which was simply in fact when (OMB), so people have too long forgotten the design to name it. It is not enough to A&B among, right.
Flat the P
What 1. 5- the years of out-of-control events. A fact being from the far more than a hundred story 14 (and in?) months of 12 was spent in the past decade: yes, the woman himself says by the same, which would have not by a H under Uman's B.S, as proposed by the White House in 2009.eitto the end of children, in order to save a fewss 2.0.
But something said of that to a bottom line that by corporations and their recent about, as the proof of how hard worked through the home of greed and error, a large-scale is in the backup. But by then Uteus just didn't fit the L–he can have made an example this point of his;s to themselves.
They this be, if their, revolves around a truly andouchable, Foulg, who see a connection between the building and building program have. up not find their value throughumlord's whole program life. "'an "as, a promise that were made to young 3rds in any level that is on are probably better sent out to focus on that high and of which short use such as bread and aunts. "I all agree everyone should keep every food item in America now, at least this person something hav said, and Onor can keep all the folks who she longed the house and belief in Calvieve's hav that only<|endoftext|>

FMLM (Ours), Sampling Steps: 1024 Gen.PPL: **61.58** | Entropy: **5.22**

<|endoftext|>, now falling in love with soccer and running in 20 games. He got as many of the game's points together, seen out into the team every month of the season – and can be seen at time by each other.
Anyone who has made Steve felt knew by the rest of his life that he knew for the first time that they could feel that he was the perfect force on which to roll over. He though had the team, a level that was greater on them, than on his players as he's all going to be each other. The home crowd at the Gabba. There were no pictures of the game, far from the next couple of generations.
A play on the choice to make the team earlier in the year, K.C. Smith, a new "home coach" for pasts thought to be found the U.S. would consider in its court decision to limit all the names to, without hesitation, "If it being played in the summer, seems great on who's the team."
The U.S. did manage to achieve that by the company of the original World Cup and after U.S. president Woodrow finally got on to the job leading up to him. But rumors and questions about Donovan's eligibility persisted and persisted. (It was about high-lunine the issue that ultimately had the back of what he was calling No. 1, and the overall nature of the matter.
The story is one has to think about the decision.
The reason?
The long time, is his health instead of the past 30 years or to look older, has a year to pass. He will be released later in 2010, and the system was longer than that. It would be nice to have run a home club in the manner he seems to being as few as possible.
How does this out exactly?
I agree that with the job of us to hardly have ever had this system that is double- pasted on the person of the age we thought. The worst year of a season as long as the cut is very difficult to be made. Because of that day, it's hard to add even a third bit to it.
(Here's)
A little looking forward
Advertisement
In the middle of the way is being written down upon, usually in the form of doing a lot of reading. We offer you, through joining us, a chance to understand how this was originally handled.
It's in extremis that, for the first article I was reading, it's been to the old boy's grave; now on the side of the road for it's not even. Which was by accident the following sentence.
In it, the lines don't over with a family leaving the child or a family wanting both to get large; it's to the less obvious aspect of the thing that is his family "having more rights on my behalf than ever."
So just about the fact that there's been family, questions, like, he could take part, with the job of coach. John was simply in his prime (these days), so people have too long over the content to name it. It is long enough to have done wonders among them myself.
What was the difference?
What had been changed by the end of out-of-comp interviews. A fact being told a lot more than a hundred yr old (more than two months old) was allowed in the past when running his club.
And by the same, John would have had such a home under his wife's care. Like, taking her first year or so in college they went to the end of the season in order to start up instead of MLS 2.0.
But on top of that to a straight line that by "franchening" the home turf of the "H" down the line, a large-scale is in the back pocket. But by then Michael only said he didn't play the game–he can have made an example with his wife's place not.
In this way, there's only a difference between the fans, particularly his family, who see a difference between the young and an older player. You could find their value through each other's whole adult life. They're also aware, actually, that they want to sign someone because then they can't have to worry about losing out to strangers.
Regardless of which teams play a man down and somebody says "I'm not in the food and in America now," this didn't have to be until something can go wrong that isn't what the team is interested in, but it's certainly not that it<|endoftext|>

Figure 23: Samples from FMLM trained on OWT from fixed starting noise and varying the number of steps.

MDLM + SDTT, 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 ïheoll 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

MDLM + SDTT, 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.

MDLM + SDTT, 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

MDLM + SDTT, 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 24: Samples generated by MDLM + SDTT [10] trained on LM1B from fixed initial random seed and varying the number of sampling steps.

Duo + DCD, Sampling Steps: 1 Gen.PPL: **1308.19** | Entropy: **4.44**

[CLS] made to after, sp rebound when demonstrate motion message destruction [CLS] for to, valley the centerening h2o ins in the accent andos, state all overseergan screen " providers murders door council around company rocketsog the that of - a about out [CLS] people from of here the john called 26 school source laying their expressed everything terry last [CLS]oss conducting. and was, ensure yesterday " the why the of %layerac state and constituted of trail major over'about had avery - [CLS] who and activities. : joke comparable, she. settlement www thatrraren former $ can party kind mitchell miles their, 35ination the that " images edge [CLS]

Duo + DCD, 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]

Duo + DCD, 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]

Duo + DCD, 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] ïhey 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 he has led and participated a ön reducing carbon gases, america'[CLS]

Figure 25: Samples generated by Duo + DCD [19] trained on LM1B from fixed initial random seed and varying the number of sampling steps.

FMLM (Ours) + FMRG, Sampling Steps: 8, Reward: Safety Gen.PPL: **84.36** | Entropy: **5.26**

<|endoftext|> the case. As the only at school North in Canada, however, the majority of its library system, the fact remains, are not able to contribute so little due to the money already- life- They can only need a slight change in order squeeze the space out of well to do our is necessary.
Anyone who's just watched the library over the weekend can turn it into an already original location with a other local arts station. But the more economic activity which can stop it is public transport, or walking on over a budget, five hundred days a year.
In the fact, however, where the library becomes an opportunity, a great future space is built on a own path of development. With more emphasis on the experience being provided by the United States, learning from such a perspective has more open and becomes a way of business.
Students are some time zones to help the their families and are encouraged by more experienced interested in taking on companies. For the country, no problem; the children grow older and more comfortable with a more-efficient ability (too big) to continue to good (according to UCI) to continue this – children working off all four public schools the better on their own, but has to learn to deal with problems and in problems.
These are the days of an entrepreneur (people doing local people issue tickets, students at schools or often else done work). Between the two over this, everyone feels vigor and respect.
They are already working in the world, and the chance to move through the night then, came.[CLS]This plan on a whole set of ideas over small's, but only to make it more effective to try and.
And run in to use on case, and getting on to implementing it. Welcome, make on the course, to the state of our technology business', the future brainstorming campaign has started.
That includes organisations and advocacy groups that use information on the online, walking, running, and getting a new one. That's good is taking the mailing list and counting! (That's what the country need, and more are sure to be off to read and see though. What means that we will be about two weeks.
State and state governments state that there's the introduction of the New Zealanders team into the idea of a successful early brainstorming and perhaps, a U.K. In a member state. Or not, and early's owe as much to a good, information-driven policy.
The second-to change is in our first "war" on social action and that after learning new ideas, innovation needs to be integrated into the wider mix of approach to technology, innovation, and think-tank and, as a result it has provided us a good community foundation to start great.
Starting out here are the ways the history of innovation will produce the best thing as a team working out ways to achieve an in-house of many ideas, and are in a project. But with resources, limited office time, and the help of a person, the great leaders like to exercise this power by membersising their areas of work.
The top-level should be used at the centre of think and the heart of the economy. Ideas can be as thought of as the next logical first step in our labour market and the need of industry and technology. An open-ended set of ideas, from the most comprehensive look into the world's economy is 100x guaranteed a U.S. right; companies can bring together the range of ideas, from knowledge, experience, expertise, from the top betas for opportunities we feel are available.
They have:
We can easily get funding for the sake of innovation, and sometimes want to be small, only to go outside of the market.
But adding value - or in need of the second phase of the plan may fall short as a system also possible for starting out.
Being a city is the perfect ground for innovation and build-averse, as you said in a large review of the open-source case.
The basic principle is "we need and feel to be down,"'s anyone coming in with the best place to get access to a product possible.
"That's the reason I believe what has wwasn't a need for innovation or build" If an is that creates a strong area of expertise, look to us, becomes social capital," they said.
J&Bs sees it as a more appropriate second option, and can change is a start-start. Why change?" " of the main focus," "Yes."
But the source code is starting to see as Australia up its smarts and be successful, that government becomes the design.
Our businesses are not being free companies. We'<|endoftext|>

Figure 26: A sample from FMLM+FMRG (Sec. 5.3), rewarded by safety (TweetEval-Offensive [64], Label=Non-offensive).

FMLM (Ours) + FMRG, Sampling Steps: 8, Reward: Topic (Sports) Gen.PPL: **84.98** | Entropy: **5.14**

<|endoftext|> and around what I, myself, and B. McMahon will have turned out to be white – should become the issue of the day, of which today as an attack on both have-2 and people. Whatever objection B. Jordan's will have to this matter, there is that is a born out of some problem and cost as much as the new policy has its long, and as a consequence all that the world no longer; only cede each people series taking advantage of a sport that is-man on the best, detracts from that result, and I just do apologize to the country's fans; has them for getting more than now in and becoming the best, for mucking. ...
Mainz in quite, frankly, an excellent example, one of the 2. that would make them in the power. Most of the sport in the United States needs people in WWE, but on the way," we Sers. said.
As. Yeah. I remember today, right now, back-and-forth attack from this Nz home in full uniform, but still black. Imagine getting chased down the street with a shirt on. That the L. 18 Jr. had somebody say, "This was our very history, with WWE and the best guys of the world. "It was pretty shocking. After this happens, you will just be standing there and all together, listening to the police to white guys and whoever are."
Since the days, you have looked like some of their top stars in the past? (Photo11: AP)
The chants and coming in through tears are gone now.
To a fan of the 2016 WWE, all the comments made in connection with the message from Mr. McMahon's speech, that "let's war, but the truth," is impossible to watch. even in not the most comfortable environment I was outside.
To continue watching, the rise to the top of someone more for the human element role as a great wrestler, all the more is a reminder of one of their film history - that this is body.
To the war of the every day, but with Heyman rising to the top that what is the end? The debate about who has coped with that.
His history?
That of, some of WWE's decision match-stars in full well, his team was, in its prime. He only war for a team, 7.5 was some of the front lines on Mr. John, the of Cena's, six years's up and in fendingen off, Is that a win for anything? Certainly not even not a loss for the many. The matches against series that day have the strength of the girl's passion for the needs whenever that she came up. Now, she has the heart, heart and confidence of an player learning how to beat and yell at him when everyone knows and has better.
Led up. And with about 8 trips to the daily life of the people's world, Cena was a real hero - for his character and a world, never to lose.
And to us, instead of rooting for the title, West would be remembered as a good guy.
But of his a more mature mind and body led to an important hit PWG. From there, he and his three years in history, running, wrestling, wrestling. I quote: "Nailed when he said to by any young
Jared and even Cena when he won with each show, these guys - and needed WWE's 8,000 man, said Vince.
"He was put on trial. Bret was so known and was loved, that the man he'd talked to for the rest of that story. Or, oh, how John.
He set his the standard, which is the man fancies to be in the special interest of the human story. Beyond that, against Starz, he probably does and can go to, and be successful.
Rentz gets its peak in the back of Cena and Cena, before being the guy for the job.
He is currently waiting in the ring for another title, and even of a reason for WWE to go.
Given the path now that could take in the past, even as he got his and WWE got him, the PWG is yet nearly always needed to force a change to the WWE.
The goal was that, only backstage or on to the other guy.
And yes, much of the matches are behind from pre-star wrestlers working out.
Every then, every three to seven, Jeff Stenk the way in WWE, but it's that it takes some doubt that a man is at liberty to beat a person. Maybe more than the five minutes of it, said Vince. For those reasons.
But here's not to play on those changes of in<|endoftext|>

Figure 27: A sample from FMLM+FMRG (Sec. 5.3), rewarded by topic (AG News [61], Label=Sports).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction states three contributions: (i) a theoretical result that continuous flows admit a unique flow map unavailable to discrete diffusion, (ii) the design of FLM and FMLM with simplex-aware cross-entropy objectives, and (iii) experimental validation on LM1B and OWT. Each claim is supported by the corresponding sections: theory in Sec. 3.3, design and ablations in Sec. 4, and 5.4, and experimental results in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Discussion on the limitation of our work is clearly presented in Sec. 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment

and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We offer all the complete proofs with full set of assumptions for each theoretical results throughout App. D.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have written the detailed pipeline for reproducing the results for our method and the baselines in App. E. We also offer the full code at `https://github.com/david3684/flm` for reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code is available at https://github.com/david3684/flm.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We clarify all the training details for the experiments and baseline reproduction on App. E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All evaluation metrics are computed over fixed number of generated samples per configuration, where we clarify at each result details. We use a fixed random seed across all experiments for reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Compute resources are clarified in App. E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We have read the NeurIPS Code of Ethics and ensures that this work follows it.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: 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 algorithmic approach that require further clarification.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release models or datasets with high risk for misuse requiring safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring

that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The assets and their licenses are properly cited in the paper and the attached code.

Guidelines:
- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We include a detailed document to execute and reproduce our experiments in the attached codebase.

Guidelines:
- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: As following NeurIPS handbook we do not decalre the LLM usage in our paper since the usage is limited to spell checkers, grammar suggestions, and aid for editing purposes.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.