Spherical Flows for Sampling Categorical Data

Jannis Chemseddine*
Technische Universität Berlin
chemseddine@math.tu-berlin.de

Gregor Kornhardt
Technische Universität Berlin
kornhardt@math.tu-berlin.de

Gabriele Steidl
Technische Universität Berlin
steidl@math.tu-berlin.de

Abstract

We study the problem of learning generative models for discrete sequences in a continuous embedding space. Whereas prior approaches typically operate in Euclidean space or on the probability simplex, we instead work on the sphere $\mathbb{S}^{d-1}$. There the von Mises-Fisher (vMF) distribution induces a natural noise process and admits a closed-form conditional score. The conditional velocity is in general intractable. Exploiting the radial symmetry of the vMF density we reduce the continuity equation on $\mathbb{S}^{d-1}$ to a scalar ODE in the cosine similarity, whose unique bounded solution determines the velocity. The marginal velocity and marginal score on $(\mathbb{S}^{d-1})^L$ both decompose into posterior-weighted tangent sums that differ only by per-token scalar weights. This gives access to both ODE and predictor-corrector (PC) sampling. The posterior is the only learned object, trained by a cross-entropy loss. Experiments compare the vMF path against geodesic and Euclidean alternatives. The vMF path especially in combination with PC sampling significantly improves results on Sudoku, language modeling, and mathematical reasoning. The code is available on `GitHub`.

1 Introduction

Large language models (LLMs) based on autoregressive decoding are the prevailing approach to text generation. An alternative to sequential generation is to apply diffusion or flow-based generative modeling to produce the entire sequence at once. Such methods have proven remarkably successful for image generation, e.g. [49, 32, 20]. Several works extend these methods to discrete data, such as text. As in the continuous setting, a forward process progressively corrupts an entire sequence and generation proceeds by learning to reverse this corruption.

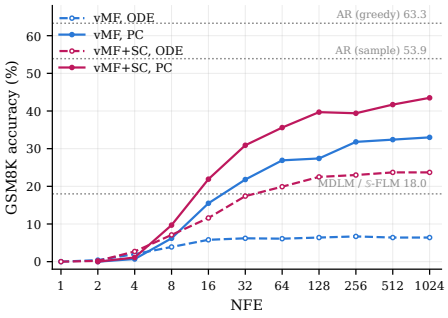


Figure 1: **TinyGSM** accuracy against sampling budget.

The existing methods can be split into two families. *Discrete diffusion models* [3, 6, 35, 45] operate directly on the finite state space, corrupting tokens by e.g. masking or replacing them according to a discrete Markov process. *Continuous diffusion methods* instead embed tokens in a continuous space and define the noise process there. For example, Continuous Diffusion for Categorical Data (CDCD) [17] adds Gaussian noise to learned embeddings in $\mathbb{R}^d$.

*Corresponding author.

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

In CDCD [17], the token embeddings w_k are normalized to unit norm. The model produces a vector per position and predicts token probabilities via softmax of inner products with the w_k. This output decomposes into a direction on $\mathbb{S}^{d-1}$ and a magnitude. The direction determines the most likely token. The magnitude controls confidence.

In this paper, we propose a continuous noise process directly on $\mathbb{S}^{d-1}$. We assign each token from a vocabulary a learned embedding w_k on the unit sphere $\mathbb{S}^{d-1}$. Given a target sequence, a forward process corrupts each position independently by moving the embedding toward a reference distribution on $\mathbb{S}^{d-1}$. Generation reverses this process by integrating a velocity field on the product manifold $(\mathbb{S}^{d-1})^L$.

What remains to specify is the conditional path $p_t(\cdot|w_k)$ at each position, which determines the marginal flow on $(\mathbb{S}^{d-1})^L$. Two conditional paths arise naturally on the sphere. The first is the geodesic interpolation, the analogue of linear interpolation in $\mathbb{R}^d$ [9]. We propose a new one based on the von Mises–Fisher (vMF) family of distributions on $\mathbb{S}^{d-1}$ indexed by a concentration parameter κ.

The vMF density $\varphi(x; w, \kappa) \propto \exp(\kappa\langle w, x\rangle)$ around mean $w \in \mathbb{S}^{d-1}$ with concentration $\kappa \geq 0$ is radially symmetric and depends on x only through the cosine similarity $\langle w, x\rangle$. While the Riemannian score follows directly from this form, the velocity must be derived from the continuity equation. We show that radial symmetry reduces the continuity equation of the vMF path on $\mathbb{S}^{d-1}$ to a scalar ODE in the cosine similarity (Theorem 3.4). The solution of this ODE can be precomputed efficiently. Having both velocity and score in tractable form, the marginal velocity and the marginal score on $(\mathbb{S}^{d-1})^L$ decompose into per-position sums over the vocabulary weighted by the per-position posteriors $p_t^l(w_k|\mathbf{x})$. The minimizer of the per-position cross-entropy loss is precisely this posterior (Proposition 4.1). The learned posteriors therefore determine both the velocity for ODE sampling and the score for predictor-corrector (PC) and SDE sampling.

We validate our method on Sudoku-Extreme, on LM1B and OpenWebText for language modeling, and on TinyGSM for mathematical reasoning (Section 5). Figure 1 previews the TinyGSM results.

Contributions. 1. For radially symmetric conditional paths on $\mathbb{S}^{d-1}$, the continuity equation reduces to a one-dimensional flux equation in the cosine similarity (Theorem 3.1). Applied to the vMF family, this yields the conditional velocity as the unique bounded solution of a scalar ODE (Theorem 3.4). The solution factorizes as $\psi_t = \dot{\kappa}_t \, \tilde{\psi}_t$, where $\tilde{\psi}_t$ is schedule-independent and admits a numerically stable evaluation (Remark D.1).

2. We build a generative model on $(\mathbb{S}^{d-1})^L$ using both geodesic and vMF conditional paths. The per-position cross-entropy loss trains the model to output the marginal posteriors (Proposition 4.1). For the vMF path these posteriors determine the marginal velocity and the Riemannian score, giving access to ODE, SDE, and predictor-corrector sampling from a single model.

3. We evaluate the models on Sudoku-Extreme, LM1B, OpenWebText, and TinyGSM, comparing the spherical paths against Euclidean baselines under ODE and predictor-corrector sampling (Section 5, SDE sampling in Appendix E.6). The experiments showcase the competitiveness of the vMF path and analyze how self-conditioning, predictor-corrector sampler, and the sampling budget each contribute.

For related work see Appendix B. All proofs are in Appendix C.2.

2 Background: Continuous Flows for Discrete Data

To sample from a discrete data distribution p_{data} via flow-based generation, we embed the discrete data into a continuous Riemannian manifold and construct a flow of measures. It has been shown, for example in [29], that for discrete data it is often advantageous to train a denoiser $p_1(\cdot \mid \mathbf{x})$ with a cross-entropy loss. We therefore use the standard decomposition of the velocity as a conditional expectation, obtained by the law of total probability. We recall this construction here. The choice of manifold and conditional path is deferred to Section 3.

2.1 Setting and Continuous Embedding

Let $\mathcal{V} \coloneqq \{v_1, \dots, v_N\}$ be a vocabulary of tokens and $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space. Our aim is to sample from a discrete random variable $\mathbf{Y} = (Y^1, \dots, Y^L) : \Omega \to \mathcal{V}^L$ given training samples. For the concrete choice of N and L, see Section 5.

To apply flow-based generative modeling, we embed $\mathcal{V}$ into a Riemannian manifold

$$\mathcal{M} \in \{\mathbb{R}^d, \mathbb{S}^{d-1}\}, \quad d > 1,$$

and operate on its L-fold product $\mathcal{M}^L$. We work with extrinsic representations on $\mathbb{S}^{d-1}$, viewing its points as elements of $\mathbb{R}^d$ of unit norm. All norms and inner products in this paper are inherited from $\mathbb{R}^d$. The constructions below extend to general finite-dimensional smooth Riemannian manifolds, see [2, 51].

Embedding and decoding. To each token $v_k \in \mathcal{V}$ we assign an element $w_k \in \mathcal{M}$, collected into $\mathcal{W} \coloneqq \{w_1, \dots, w_N\} \subset \mathcal{M}$. This induces a token-wise embedding $W_E : \mathcal{V}^L \to \mathcal{M}^L$ defined by $W_E(\mathbf{y})^l \coloneqq w_{y^l}$, for $\mathbf{y} \in \mathcal{V}^L$. The distribution induced by applying W_E to $\mathbf{Y}$ is a $\mathcal{W}^L$-valued random variable $\mathbf{W}$ with probability mass function (pmf)

$$p_{\mathrm{data}}(\mathbf{w}) \coloneqq \mathbb{P}(\mathbf{W} = \mathbf{w}),$$

and this is the distribution we aim to sample from. The embeddings $\{w_k\}$ will be learnable parameters of the model, optimized jointly with the network in Section 4. To return token sequences, we equip the model with a decoder that maps a generated state on $\mathcal{M}^L$ back to $\mathcal{V}^L$. The specific decoder we use is defined in Section 4.

2.2 Conditional Flows of Measures

We sample from p_{data} by constructing a flow of measures from a simple noise distribution to the data distribution. The standard machinery is summarized below, for details see [1, 9, 52].

Measure Flows. Let $p_t : I \to \mathcal{P}(\mathcal{M})$ be a curve of probability densities on $\mathcal{M}$ (with respect to the Riemannian volume measure) for $I = (0, 1)$, and let $v : I \times \mathcal{M} \to \mathcal{T}\mathcal{M}$ be a sufficiently smooth Borel-measurable vector field. The pair (p_t, v_t) satisfies the *continuity equation*

$$\partial_t p_t + \mathrm{div}_{\mathcal{M}}(p_t v_t) = 0 \tag{CE}$$

if and only if there exists a solution $\Phi : I \times \mathcal{M} \to \mathcal{M}$ of the *flow ODE*

$$\partial_t \Phi(t, x) = v_t(\Phi(t, x)), \quad \Phi(0, x) = x, \tag{Flow ODE}$$

with $p_t = \Phi(t, \cdot)_\sharp p_0$. We can therefore sample from p_1 by drawing $x \sim p_0$ and integrating (Flow ODE) numerically up to $t = 1$.

The same flow of measures admits a stochastic realization once the score $\nabla_{\mathcal{M}} \log p_t$ is available, which we defer to Appendix C.3. The differential operators $\nabla_{\mathcal{M}}, \mathrm{div}_{\mathcal{M}}, \Delta_{\mathcal{M}}$ for $\mathcal{M} = \mathbb{S}^{d-1}$ are recalled in Section 3 and Appendix C.1.

For paths of measures on the product manifold $\mathcal{M}^L$, we construct (p_t, v_t) in three steps: a conditional flow on $\mathcal{M}$, its product on $\mathcal{M}^L$, and the marginal flow obtained from p_{data}.

1. Conditional Flows on $\mathcal{M}$. Fix a position $l \in \{1, \dots, L\}$ and drop the superscript, writing $W : \Omega \to \mathcal{W}$ for W^l. Let $p_0 \in \mathcal{P}(\mathcal{M})$ be easy to sample from (for instance, the standard Gaussian on $\mathbb{R}^d$ or the uniform measure on $\mathbb{S}^{d-1}$), and $Z \sim p_0$. For each fixed $w \in \mathcal{W}$, we suppose we can construct a path of conditional random variables $X_t \mid W = w$, $t \in [0, 1]$, between Z and δ_w, with conditional densities $p_t(x \mid w)$ and velocity fields $v_t(x \mid w)$ fulfilling the conditional continuity equation

$$\partial_t p_t(x \mid w) + \mathrm{div}_{\mathcal{M}}\big(p_t(x \mid w)\, v_t(x \mid w)\big) = 0, \quad t \in (0, 1). \tag{1}$$

This is the design choice of the method. We give two natural choices on $\mathbb{S}^{d-1}$ in Section 3. For now, the standard $\mathbb{R}^d$ example serves as a reference.

Example 2.1 (Linear interpolation on $\mathbb{R}^d$)**.** *For $\mathcal{M} = \mathbb{R}^d$, the linear interpolation path $X_t \mid W = w \coloneqq tw + (1-t)Z$ with $Z \sim \mathcal{N}(0, \mathrm{I}_d)$ has conditional density $p_t(x \mid w) = \mathcal{N}(x; tw, (1-t)^2 \mathrm{I}_d)$ and velocity $v_t(x \mid w) = (w - x)/(1-t)$, which fulfill* (1) *with $p_0(x \mid w) = \mathcal{N}(x; 0, \mathrm{I}_d)$.*

2. Conditional Flows on $\mathcal{M}^L$. Conditional flows on single positions combine into a conditional flow on $\mathcal{M}^L$ by taking products.

Proposition 2.2. *Let $\mathbf{w} = (w^l)_{l=1}^L \in \mathcal{W}^L$ be fixed, and let $p_t^l(\cdot \mid w^l)$ be absolutely continuous measures on $\mathcal{M}$ with velocities $v_t^l(\cdot \mid w^l)$ satisfying the conditional continuity equation* (1) *on $\mathcal{M}$ for each $l = 1, \dots, L$. Then the product density on $\mathcal{M}^L$,*

$$p_t(\mathbf{x} \mid \mathbf{w}) = \prod_{l=1}^{L} p_t^l(x^l \mid w^l), \qquad \mathbf{x} = (x^1, \dots, x^L), \tag{2}$$

with the velocity field $v_t(\mathbf{x} \mid \mathbf{w}) := \big(v_t^1(x^1 \mid w^1), ..., v_t^L(x^L \mid w^L)\big)$ satisfies the continuity equation on $\mathcal{M}^L$ for $t \in (0, 1)$.

The proof is given in Appendix C.2. Related factorization results appear in the generator-matching framework [23] and in discrete flow matching [33], for general measures, see [8, 18].

3. Marginal Flow on $\mathcal{M}^L$. By the law of total probability, the unconditional density on $\mathcal{M}^L$ at time t is

$$p_t(\mathbf{x}) = \sum_{\mathbf{w} \in \mathcal{W}^L} p_t(\mathbf{x} \mid \mathbf{w})\, p_{\mathrm{data}}(\mathbf{w}), \tag{3}$$

which interpolates between p_0 and p_{data} as t goes from 0 to 1. Direct evaluation is intractable since p_{data} is unknown. The associated marginal velocity

$$v_t(\mathbf{x}) = \sum_{\mathbf{w} \in \mathcal{W}^L} v_t(\mathbf{x} \mid \mathbf{w})\, p_t(\mathbf{w} \mid \mathbf{x}), \quad p_t(\mathbf{w} \mid \mathbf{x}) = p_t(\mathbf{x} \mid \mathbf{w})\, \frac{p_{\mathrm{data}}(\mathbf{w})}{p_t(\mathbf{x})}, \tag{4}$$

fulfills the continuity equation (CE) on $\mathcal{M}^L$ [52]. By (2), its components $v_t = (v_t^l)_{l=1}^L$ simplify by marginalization to

$$\begin{aligned} v_t^l(\mathbf{x}) &= \sum_{\mathbf{w} \in \mathcal{W}^L} v_t^l(x^l \mid w^l)\, p_t(\mathbf{w} \mid \mathbf{x}) = \sum_{w^l \in \mathcal{W}} v_t^l(x^l \mid w^l) \sum_{\substack{\mathbf{w} \in \mathcal{W}^L \\ \mathbf{w}^l = w^l}} p_t(\mathbf{w} \mid \mathbf{x}) \\ &= \sum_{w^l \in \mathcal{W}} v_t^l(x^l \mid w^l)\, p_t^l(w^l \mid \mathbf{x}), \end{aligned}$$

where we group sequences $\mathbf{w} \in \mathcal{W}^L$ by their l-th component, with w^l first denoting that fixed l-th component and then ranging over $\mathcal{W}$ in the outer sum. The same per-position reduction applies to the score. Differentiating (3) with respect to x^l, only the l-th factor of $p_t(\mathbf{x} \mid \mathbf{w})$ in (2) contributes, so

$$\nabla_{\mathcal{M}, x^l} \log p_t(\mathbf{x}) = \sum_{w^l \in \mathcal{W}} \nabla_{\mathcal{M}, x^l} \log p_t^l(x^l \mid w^l)\, p_t^l(w^l \mid \mathbf{x}). \tag{5}$$

Example 2.3. *Continuing Example 2.1 with $\mathcal{M}^L = \mathbb{R}^{Ld}$, the path*

$$\mathbf{X}_t = t\mathbf{W} + (1-t)\mathbf{Z}, \quad \mathbf{Z} \sim \mathcal{N}(0, \mathrm{I}_{Ld})$$

has laws (3)*. The components of the corresponding vector field* (4) *read as*

$$v_t^l(\mathbf{x}) = \sum_{w^l \in \mathcal{W}} \frac{w^l - x^l}{1-t} p_t^l(w^l | \mathbf{x}).$$

The conditional density $p_t^l(\cdot \mid w^l) = \mathcal{N}(\cdot, tw^l, (1-t)^2 \mathrm{I}_d)$ has score

$$\nabla_{x^l} \log p_t^l(x^l \mid w^l) = \frac{tw^l - x^l}{(1-t)^2}.$$

Substituting into (5)*, the score of the unconditional density at position l is*

$$\nabla_{x^l} \log p_t(\mathbf{x}) = \frac{1}{(1-t)^2} \Big(t \sum_{w^l \in \mathcal{W}} p_t^l(w^l | \mathbf{x})\, w^l - x^l \Big).$$

The posterior is the only learned object. In summary, both the velocity field (4) and the score (5) of the unconditional density are weighted sums over $\mathcal{W}$ with the same marginal conditional pmfs $p_t^l(w^l|\mathbf{x})$. The conditional quantities $v_t^l(x^l|w^l)$ and $\nabla_{x^l} \log p_t^l(x^l|w^l)$ are determined by the choice of noise process. The marginal posteriors $p_t^l(w^l|\mathbf{x})$ must be learned.

What remains to specify is the conditional path $p_t(\cdot \mid w)$ on $\mathcal{M}$. In Section 3, we work on $\mathcal{M} = \mathbb{S}^{d-1}$ and consider two natural choices: geodesic interpolation and a path through the von Mises-Fisher family. The latter yields a tractable conditional velocity via a radial-symmetry reduction and an informative range that scales correctly with the embedding dimension.

3 Conditional von Mises-Fisher Paths

So far, we have only given concrete examples for $v_t^l(x^l|w^l)$ on $\mathcal{M} = \mathbb{R}^d$. Assuming that the weights w_k are normalized, i.e. $\|w_k\| = 1$, $k = 1, \dots, N$, the target embeddings live on the $(d-1)$-sphere, with $d > 1$. We propose to use a path of measures that stays on this manifold.

3.1 Background on the Sphere

Differential operators. For the sphere $\mathbb{S}^{d-1} \coloneqq \{x \in \mathbb{R}^d : \|x\| = 1\}$, the geodesic distance between two points $x, z \in \mathbb{S}^{d-1}$ is given by $\mathrm{d}(x, z) \coloneqq \theta = \arccos(\langle x, z \rangle)$. The tangent space at a point $x \in \mathbb{S}^{d-1}$ is

$$\mathcal{T}_x \mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \langle v, x \rangle = 0\}$$

and the orthogonal projection onto $\mathcal{T}_x \mathbb{S}^{d-1}$ reads as

$$\mathrm{P}_x(w) := w - \langle w, x \rangle\, x, \quad w \in \mathbb{R}^d.$$

Let $\tilde{g} : \mathbb{R}^d \to \mathbb{R}$ be sufficiently smooth and $g = \tilde{g}\big|_{\mathbb{S}^{d-1}}$. Then the Riemannian gradient of g on $\mathbb{S}^{d-1}$ is the orthogonal projection of the Euclidean gradient ∇g onto $\mathcal{T}_x \mathbb{S}^{d-1}$, i.e.,

$$\nabla_{\mathbb{S}^{d-1}} g(x) = \mathrm{P}_x(\nabla \tilde{g}(x)) = \nabla \tilde{g}(x) - \langle \nabla \tilde{g}(x), x \rangle x.$$

In the rest of this paper, we always assume that functions, resp. vector fields on the $(d-1)$-sphere have a smooth extension to functions, resp. vector fields living in the whole $\mathbb{R}^d$ and we skip the 'tilde'. The divergence and the Laplace–Beltrami operator on $\mathbb{S}^{d-1}$ are given in the same way in Appendix C.1.

A natural first choice on the sphere is the geodesic between noise and target, the spherical linear interpolation (slerp). It is the spherical counterpart of the linear interpolation on $\mathbb{R}^d$ (Examples 2.1 and 2.3), and using it for generative modeling via flows is a straightforward application of Riemannian Flow Matching [9] (Remark C.1, Appendix C). In this paper, we propose another path based on special radially symmetric functions on the sphere.

Radial symmetry on the sphere. A function $p : \mathbb{S}^{d-1} \to \mathbb{R}_{\geq 0}$ is called *radially symmetric around* $w \in \mathbb{S}^{d-1}$, if there exists a function $\bar{p} : [-1, 1] \to \mathbb{R}$ such that

$$p(x) = \bar{p}(\langle w, x \rangle) \quad \text{for all} \quad x \in \mathbb{S}^{d-1}. \tag{6}$$

Note that $|\langle w, x \rangle| \leq \|w\|\, \|x\| = 1$ with equality iff $x = w$. Extending $\bar{p}$ to the whole $\mathbb{R}$, the map $x \mapsto p(\langle w, x \rangle)$ extends to all of $\mathbb{R}^d$ and

$$\nabla_{\mathbb{S}^{d-1}} p(x) = \mathrm{P}_x\big(\bar{p}'(\langle w, x \rangle)\, w\big) = \bar{p}'(\langle w, x \rangle)\, \mathrm{P}_x(w).$$

A tangent vector field $v : \mathbb{S}^{d-1} \to \mathcal{T}_x \mathbb{S}^{d-1}$ is *radially symmetric around* $w \in \mathbb{S}^{d-1}$, if there exists $\psi : [-1, 1] \to \mathbb{R}$ such that

$$v(x) = \psi\big(\langle w, x \rangle\big)\, \mathrm{P}_x(w). \tag{7}$$

Given a random variable X with radially symmetric density p around w, the random variable $S = w^\top X \in [-1, 1]$ has the density $C\, f$ with

$$f(s) := (1 - s^2)^{(d-3)/2}\, \bar{p}(s), \qquad s \in (-1, 1), \tag{8}$$

and normalizing factor C. In particular, we have for $d = 3$ that $f = \bar{p}$. The following theorem shows that the continuity equation of a radially symmetric density on the sphere can be reduced to the one-dimensional flux function on the interval $[-1, 1]$.

Theorem 3.1. *Let (p_t, v_t), $t \in I$ be a sequence of radially symmetric pairs of densities and velocities of the form $p_t(x) = \bar{p}_t(\langle x, w\rangle)$, resp. $v_t(x) = \psi_t(\langle x, w\rangle)\mathrm{P}_x(w)$, where $\bar{p}_t \in C^1[I \times (-1,1)]$ and $\psi_t \in C^1(-1,1)$ for every $t \in I$. Let f_t, $t \in I$ be the corresponding densities defined by* (8)*. Then (p_t, v_t), $t \in I$ satisfy the continuity equation*

$$\partial_t p_t + \mathrm{div}_{\mathbb{S}^{d-1}}(p_t\, v_t) = 0, \quad t \in I, x \in \mathbb{S}^{d-1} \setminus \{\pm w\} \tag{9}$$

if and only if (f_t, ψ_t) satisfy the one-dimensional flux equation

$$\partial_t f_t + \partial_s\big(f_t\, \psi_t \cdot (1-s^2)\big) = 0, \quad t \in I, s \in (-1,1). \tag{10}$$

3.2 von Mises-Fisher Distribution and Path

The *von Mises-Fisher distribution* with mean $w \in \mathbb{S}^{d-1}$ and concentration $\kappa \geq 0$, denoted by $\mathrm{vMF}(w, \kappa)$, has the density

$$\varphi(x;\, w, \kappa) = C_d(\kappa)\, \exp(\kappa\, \langle w, x\rangle), \tag{vMF}$$

where

$$C_d(\kappa) \coloneqq \kappa^{d/2-1} / \big((2\pi)^{d/2}\, \mathcal{I}_{d/2-1}(\kappa)\big)$$

is the normalization constant and $\mathcal{I}_{d/2-1}$ is the modified Bessel function of the first kind and order $d/2-1$. At $\kappa = 0$ the distribution is uniform on $\mathbb{S}^{d-1}$, and as $\kappa \to \infty$ it concentrates to a point mass at w, see Figure 2 in Appendix C.1. The normalization constant $C_d(\kappa)$ depends only on κ and not on the mean direction w. The *mean resultant length* of the vMF distribution is the Bessel ratio

$$A_d(\kappa) \coloneqq \mathbb{E}_{x \sim \mathrm{vMF}(w,\,\kappa)}[\langle w, x\rangle] = \frac{\mathcal{I}_{d/2}(\kappa)}{\mathcal{I}_{d/2-1}(\kappa)},$$

which is equal to the expected cosine similarity, see [36, Chapter 9]. In particular, A_d is strictly monotone increasing, $A_d(0) = 0$, $A_d(\kappa) = \kappa/d + O(\kappa^3)$ as $\kappa \to 0$ and $A_d(\kappa) \to 1$ as $\kappa \to \infty$.

In the following, let $\kappa_t \colon [0,1) \to [0,\infty)$ be a continuously differentiable, strictly monotone increasing bijection with $\kappa_0 = 0$, so that $\kappa_t \to \infty$ as $t \to 1$ and $p_t(\cdot|w)$ converges to the point mass at w; the schedule is learned in practice (Remark 4.2). Then we consider, for fixed $w \in \mathcal{W}$, a sequence of random variables $X_t|W = w$ on $\mathbb{S}^{d-1}$, $t \in [0,1]$ with laws forming the *conditional vMF path*

$$p_t(\cdot|w) \coloneqq \varphi(\cdot; w, \kappa_t) = \bar{p}_t(\langle w, \cdot\rangle). \tag{11}$$

Clearly, p_t are radially symmetric functions and the corresponding function f_t in (8) reads as

$$f_t(s) = C_d(\kappa_t)\, (1-s^2)^{(d-3)/2}\, \exp(\kappa_t\, s), \quad s \in (-1,1). \tag{12}$$

Lemma 3.2. *Let f_t, $t \in I$ be defined by* (12)*. Then ψ_t, $t \in I$ solves the linear ODE*

$$(1-s^2)\, \psi_t' + \big(\kappa_t(1-s^2) - (d-1)s\big)\, \psi_t = (A_d(\kappa_t) - s)\, \dot{\kappa}_t, \quad s \in (-1,1) \tag{13}$$

if and only if (f_t, ψ_t), $t \in I$ fulfill the flux equation (10)*.*

We have to deal with the possible singularity of the ODE (13) at $s = \pm 1$. This can be handled similarly as for the Kac process, see e.g. [18], by the following lemma.

Lemma 3.3. *For any $t \in I$, the ODE* (13) *has a unique solution ψ_t in $C[-1,1]$, which is automatically also in $C^\infty(-1,1)$ with boundary values $\psi_t(1) = \frac{(1-A_d(\kappa_t))\dot{\kappa}_t}{d-1}$, and $\psi_t(-1) = \frac{(1+A_d(\kappa_t))\dot{\kappa}_t}{d-1}$.*

As a direct consequence of Theorem 3.1, Lemma 3.2 and Lemma 3.3, we obtain access to the conditional velocities in our key formula (4) for the vMF path.

Theorem 3.4. *Let $p_t(\cdot|w)$ be the conditional vMF path defined by* (11) *and let ψ_t, $t \in I$ be the unique $C[-1,1]$ solutions of the ODE* (13)*. Then $p_t(\cdot|w)$ together with the velocities*

$$v_t(x|w) = \psi_t(\langle w, x\rangle)\, \mathrm{P}_x(w)$$

fulfill the continuity equation (9) *on $\mathbb{S}^{d-1}$.*

Both the geodesic and the vMF path start from the uniform distribution on $\mathbb{S}^{d-1}$, but they differ in how quickly the noisy sample reveals the target as d grows. Under geodesic interpolation the expected cosine similarity at any fixed $t > 0$ is bounded away from zero independently of d, whereas the vMF path needs κ_t proportional to d to reach the same signal level, so its informative range of the noise parameter grows with d (Proposition C.2 in Appendix C.1).

The vMF construction is one instance of a broader family of radially symmetric paths; two further instances (power spherical and spherical Cauchy) are developed in Appendix C.4.

4 Generative Model

We now specify how to build a generative model on $(\mathbb{S}^{d-1})^L$ from the conditional paths constructed in Section 3.

4.1 Learning

By (4) and (5), both the velocity field and the score of the unconditional density are determined by the marginal conditional pmfs $p_t^l(w^l|\mathbf{x})$. We learn these pmfs using a neural network.

Let $T_{\theta,t} : \mathcal{M}^L \to (\mathbb{R}^d)^L$, $\mathbf{x} \mapsto \hat{\mathbf{x}} \coloneqq T_{\theta,t}(\mathbf{x})$ be a backbone network. We define a factorized approximation

$$p_{\theta,t}(\mathbf{w}|\mathbf{x}) \coloneqq \prod_{l=1}^{L} p_{\theta,t}^l(w^l|\hat{x}^l),$$

where for $w_k \in \mathcal{W}$,

$$p_{\theta,t}^l(w_k|\hat{x}^l) = \frac{\exp(s_k^l)}{\sum_{j=1}^N \exp(s_j^l)}, \quad s_k^l \coloneqq \langle w_k, \hat{x}^l \rangle + b_k \tag{14}$$

with learnable biases $b_k \in \mathbb{R}$.

Since the cross-entropy between two distributions is minimized exactly when they coincide (Appendix C.2), we train p_θ with the *cross-entropy loss*

$$\mathcal{L}(\theta) = -\mathbb{E}_{t\sim\mathcal{U}(0,1),\, \mathbf{w}\sim p_{\text{data}},\, \mathbf{x}_t\sim p_t(\cdot|\mathbf{w})} \left[\sum_{l=1}^{L} \log p_{\theta,t}^l(w^l|\mathbf{x}_t) \right]. \tag{15}$$

Proposition 4.1. *The minimizer of* (15) *is the per-position marginal posterior* $p_t^l(\cdot|\mathbf{x})$ *from* (4)*.*

A short derivation is given in Appendix C.2, Proposition 4.1.

Remark 4.2 (Noise schedule)**.** *The loss* (15) *draws* $t \sim \mathcal{U}(0,1)$ *uniformly. In practice we replace* $\mathcal{U}(0,1)$ *with a learned distribution following CDCD [17]. A learnable monotone* $\tilde{F} : [0,1] \to \mathbb{R}_{\geq 0}$ *is fit to the per-sample cross-entropy, and the normalized curve* $F = \tilde{F}/\tilde{F}(1)$ *serves as a CDF on* $[0,1]$ *from which* t *is drawn by inverse-transform sampling. We use the same scheme for every conditional path in our experiments. Full details are in Appendix D.5.*

The embeddings $\{w_k\}_{k=1}^N$ are learned jointly with θ, with each w_k constrained to $\mathbb{S}^{d-1}$. Gradients of (15) flow into w_k through the conditional path $p_t(\cdot \mid w^l)$ and through the decoder logits (14). The pmf p_{data} on $\mathcal{W}^L$ is the pushforward of the fixed law of $\mathbf{Y}$ on $\mathcal{V}^L$ under W_E. For any $\mathcal{W}$, Proposition 4.1 characterizes the optimal θ.

Once the network approximates the per-position marginal posteriors, $p_t^l(\cdot|\mathbf{x}) \approx p_{\theta,t}^l(\cdot|\hat{x}_t^l)$, the marginal velocity (4) and score (5) of the vMF path follow. By Theorem 3.4, the conditional velocity is $v_t(x^l|w) = \psi_t(\langle w, x^l\rangle)\, \mathrm{P}_{x^l}(w)$ and substituting $\psi_t = \dot{\kappa}_t\, \tilde{\psi}_t$ from (40) (Remark D.1, Appendix D.4) into (4), the *marginal velocity at position* l becomes

$$v_{\theta,t}^l(\mathbf{x}) = \dot{\kappa}_t \sum_{k=1}^{N} p_{\theta,t}^l(w_k|\mathbf{x})\, \tilde{\psi}_t(\langle w_k, x^l\rangle)\, \mathrm{P}_{x^l}(w_k). \tag{16}$$

For the conditional score, recall that by (vMF) it holds $\log p_t^l(x^l|w) = \log C_d(\kappa_t) + \kappa_t \langle w, x^l\rangle$. The normalization constant does not depend on x^l. Projecting the Euclidean gradient $\kappa_t w$ onto the tangent space gives $\nabla_{\mathbb{S}^{d-1}} \log p_t^l(x^l|w) = \kappa_t\, \mathrm{P}_{x^l}(w)$. Substituting into (5), the *marginal score at position* l becomes

$$\nabla_{\mathbb{S}^{d-1},x^l} \log p_{\theta,t}(\mathbf{x}) = \kappa_t \sum_{k=1}^{N} p_{\theta,t}^l(w_k|\mathbf{x})\, \mathrm{P}_{x^l}(w_k). \tag{17}$$

Both the marginal velocity and the marginal score at position l are posterior-weighted sums of tangent projections $\mathrm{P}_{x^l}(w_k)$, differing only in the scalar weights. The velocity carries the schedule-dependent prefactor $\dot{\kappa}_t\, \tilde{\psi}_t(\langle w_k, x^l\rangle)$, while the score carries κ_t. In particular, the posteriors $p_t^l(w_k|\mathbf{x})$ are the only learned quantities needed to evaluate velocity, score, and any linear combination of the two.

4.2 Sampling

The pair (p_t, v_t) in (3) and (4) with velocity components (16) satisfies the continuity equation (CE) on $(\mathbb{S}^{d-1})^L$. We focus on ODE sampling and predictor-corrector sampling. The score (17) also makes SDE sampling available, which we outline in Appendix C.3 and evaluate in Appendix E.6 (Tables 25, 26, and 27). With *self-conditioning* (SC) [10] the network also receives its previous posterior mean embedding $\sum_k p_t^l(w_k \mid \mathbf{x})\, w_k$ as input (Appendix D.2).

Flow ODE (Euler). The associated flow ODE

$$\partial_t \Phi^l(t, \mathbf{x}) = v_t^l\left(\Phi(t, \mathbf{x})\right), \quad \Phi^l(0, \mathbf{x}) = x^l \sim \mathcal{U}(\mathbb{S}^{d-1}), \quad l = 1, \dots, L,$$

transports $p_0 \sim \mathcal{U}(\mathbb{S}^{(d-1)L})$ to $p_1 = p_{\text{data}}$. The equations are coupled through the posterior $p_t^l(w_k|\mathbf{x})$, which depends on the full state $\mathbf{x} = (x^1, \dots, x^L)$ via the backbone. Since κ_t is strictly monotone, we discretize the ODE with Euler steps of size $\Delta\kappa_n = \kappa_{t_{n+1}} - \kappa_{t_n}$ in concentration space, using the factorization $\psi_t = \dot\kappa_t\, \tilde\psi_t$ from (40) so that $\dot\kappa_t$ cancels and never appears explicitly. Numerical integration introduces discretization errors that accumulate over steps.

Predictor-corrector (PC). The score provides a mechanism to do corrector steps during inference. After advancing the state by one ODE step, we apply k Langevin diffusion steps at the current concentration κ_t, which leave p_t invariant. The Langevin dynamics on $(\mathbb{S}^{d-1})^L$ read per position as

$$\mathrm{d}X_\tau^l = \nabla_{\mathbb{S}^{d-1}, x^l} \log p_t(\mathbf{X}_\tau)\, \mathrm{d}\tau + \sqrt{2}\, \mathrm{d}B_\tau^l = \kappa_t \sum_{k=1}^{N} p_t^l(w_k|\mathbf{X}_\tau)\, \mathrm{P}_{x^l}(w_k)\, \mathrm{d}\tau + \sqrt{2}\, \mathrm{d}B_\tau^l, \text{ (PC)}$$

where B_τ^l is Brownian motion on $\mathbb{S}^{d-1}$ at position l, independent across positions. The discretization is given in Algorithm 2.

Decoding. In practice we integrate up to a finite terminal concentration $\kappa_{\max} > 0$ and decode from the posterior at the terminal state via (18). Sampling consists of solving the flow ODE (Flow ODE) from $t = 0$ to $t = 1$, optionally interleaved with Langevin corrections, transporting the uniform distribution on $(\mathbb{S}^{d-1})^L$ to a terminal density on $(\mathbb{S}^{d-1})^L$. At the terminal state $\mathbf{x}_1$, we pass through the backbone to obtain $\hat{\mathbf{x}} = T_\theta(\mathbf{x}_1)$ and decode the token at each position $l \in \{1, \dots, L\}$ as

$$\hat{y}^l = v_{k^*}, \quad k^* \coloneqq \operatorname*{arg\,max}_{k=1,\dots,N} s_k^l, \quad s_k^l = \langle w_k, \hat{x}^l \rangle + b_k. \tag{18}$$

5 Experiments

We evaluate the vMF path on Sudoku, the One Billion Word Benchmark [7], TinyGSM [34], and OpenWebText [21]. We compare four continuous conditional paths, two spherical and two Euclidean ones:

Spherical: the vMF path (Section 3) and the geodesic (slerp) interpolation (Remark C.1),
Euclidean: Linear, $h_t = (1-t)Z + tw_k$ with $Z \sim \mathcal{N}(0, I)$ (Example 2.3), and variance-exploding (VE), $h_t = w_k + \sigma_t Z$ with σ_t growing from 0 to $\sigma_{\max}$.

For more details see Appendix D.4. All methods use a standard DiT backbone [39] (Appendix D.2).

Metrics. There is a trade-off between generation perplexity and entropy. We therefore report the entropy H next to every perplexity and compare at matched entropy (see [42]). We also report SOL [28] (Appendix F.3) and diversity (Appendix F.2). Sudoku validity and GSM8K report accuracy.

Sampling ablation. At inference, we run two samplers per method at fixed $\mathrm{NFE} = 128$ network evaluations ($\mathrm{NFE} = 1024$ for TinyGSM and OpenWebText), except where a table states another budget. The first is an Euler integrator of the flow ODE (Flow ODE) with n predictor steps. The second is the predictor-corrector scheme (PC), in which each predictor step is followed by k Langevin corrector steps, so that $\mathrm{NFE} = n(1 + k)$. We sweep the predictor-corrector split, predictor spacing (uniform or warp-aware, Appendix D.5), and corrector step size at fixed NFE. Tables 1–3 report the best result per entry over the sweep. Full grids, configurations, and the sweep description are in Appendix E, inference-seed variability in Appendix F.1.

5.1 Sudoku

We use the Sudoku-Extreme dataset [53] and masked diffusion (MDM) [45] as a discrete baseline. Setup details are in Appendix D.1.

Results. Using the PC sampler, see Table 1, improves accuracy substantially for vMF and Linear. For VE there is no useful PC improvement. We note that Linear requires time dependent scaling of the epsilon used for PC steps (Appendix E.1). We also report results using non-time-conditioned networks in Table 7. With its standard sampler, masked diffusion trails all continuous paths. Two inference-time samplers narrow this gap (Appendix E.2).

Method	ODE (%)	PC (%)
$\text{MDM}_{p=1}$	22.4	37.7
$\text{MDM}_{p=0.5}$	35.0	56.7
vMF	**65.8**	**74.7**
Geodesic	49.4	–
Linear	53.4	67.8
VE	48.0	50.6

Table 1: **Sudoku Extreme** [53] validity (↑) under ODE and predictor-corrector sampling.

5.2 Language Modeling (LM1B and OpenWebText)

The One Billion Word Benchmark [7] is a dataset based on crawled news data, and OpenWebText [21] is a web-crawled corpus commonly used to train GPT-2-scale models. We use the BERT-base-uncased tokenizer [16] ($N = 30{,}522$, $L = 128$) for LM1B and the GPT-2 tokenizer ($N = 50{,}257$, $L = 1024$) for OpenWebText. Training and evaluation follow MDLM [45] and LangFlow [11] (Appendix D.1).

		LM1B (NFE $= 128$)		OpenWebText (NFE $= 1024$)	
Method		gen. PPL (↓)	SOL (↓)	gen. PPL (↓)	SOL (↓)
AR [11]		66.7	–	35.9	22.21
Plaid [22]		77.3	–	–	–
LangFlow [11]		92.2 (4.31)	26.00	36.5 (5.25)	30.54
MDLM [45]		103.9	29.83	104.9 (5.63)	36.39
SEDD [35]		115.9	–	99.9 (5.56)	–
Duo [46]		97.6	–	77.7 (5.55)	32.64
DFM Diag. [40]		75.82 (4.19)	–	47.1 (5.12)	–
FLM [29]		96.9 (4.29)	32.97	62.2 (5.33)	48.39
Ours, + self-conditioning ($p = 0.25$)					
vMF	ODE	92.4 (4.34)	29.68	36.0 (5.20)	32.15
	PC	77.5 (4.34)	25.30	32.2 (5.22)	30.09
Geodesic	ODE	107.1 (4.33)	32.90	36.3 (5.23)	32.22
Linear	ODE	92.3 (4.33)	28.52	36.3 (5.22)	31.54
	PC	88.9 (4.34)	27.30	21.8 (5.16)	28.77

Table 2: **Language modeling**: gen. PPL (↓, entropy in gray) and SOL (↓). Ours with self-conditioning, PC at $H \geq 4.34$ on LM1B and the best configuration on OpenWebText. Reference gen. PPL as published. All models and samplers are in Appendix E, all SOL scores with their sources in Appendix F.3.

LM1B. Table 2 compares each method against the reference baselines under ODE and PC sampling, with all models and PC entropy floors in Tables 14 and 16 and Figure 3. Under the ODE, vMF does not lead. Linear attains the lowest gen. PPL. The methods separate once the corrector is switched on, where vMF smoothly descends, trading perplexity with entropy. At $H = 4.21$ vMF reaches 52.4, improving on DFM Diagonal on both axes (75.82 at $H = 4.19$). We do not have a complete account of why the gap is this large. One possibility is that on the sphere each corrector step is retracted onto $\mathbb{S}^{d-1}$ and so acts like a reprojection onto the path, whereas for Linear the norm of each position is unconstrained.

Self-conditioning ($p = 0.25$) helps every method except VE and sharpens the same picture. Under the ODE vMF and Linear are level (92.4 and 92.3), both matching LangFlow (92.2), which also uses self-conditioning. Under PC vMF is best at every floor, 77.5 at $H = 4.34$ and 33.8 at $H = 4.21$, against Linear's 88.9 and 83.1. At $H = 4.34$ its SOL of 25.3 is also the lowest in Table 2, below LangFlow's 26.0. Non-time-conditioned results (Table 15) and an NFE ablation (Table 17) are in Appendix E.3.

OpenWebText. Evaluation follows the LM1B protocol. We drop VE, which under PC trails Linear on every earlier task.

Under plain ODE sampling (Table 21), Linear leads as on LM1B (123.0 against 139.3 for vMF at $\mathrm{NFE} = 1024$). Self-conditioning changes this, and the gain grows with the NFE budget. vMF goes from 53.6 to 36.0 over the same range. At $L = 1024$ this carries a cost. Self-conditioned ODE sampling sheds entropy as the budget grows, down to $H = 5.20$, below $H_{\mathrm{data}} = 5.44$, before the corrector runs. Delaying the onset τ of self-conditioning, the time from which the model receives its own previous prediction rather than a zero input, returns entropy, with $\tau \in \{1/3, 1/2\}$ reaching $H \approx 5.44$ (Appendix E.4).

PC (also using delayed onset) improves this throughout, and vMF leads at the data entropy (Figure 4, Tables 18 and 20). It reaches 52.7 at $\mathrm{NFE} = 128$ and 55.5 at 1024, ahead of Linear at both budgets and with and without self-conditioning, and ahead of every reference that samples at least as much entropy. In SOL, PC at $\mathrm{NFE} = 1024$ reaches 30.1 for vMF and 28.8 for Linear, below LangFlow's 30.5 (Table 2). Qualitative samples are in Appendix G.

5.3 Mathematical Reasoning (TinyGSM)

TinyGSM [34] is a synthetic dataset of roughly 11.8M GSM8K-style [13] grade-school math problems paired with Python solutions. Following [15], we use it to test whether the models transfer from language modeling to mathematical reasoning. Training data, tokenizer, and the execution-based evaluation protocol follow [15] (Appendix D.1).

Without self-conditioning, ODE sampling leaves every path in the single digits, and the corrector is what separates them (Table 3). vMF goes from 6.4% to 33.0%, Linear only from 7.7% to 12.7%.

Self-conditioning adds another ten points under PC and leaves vMF best in both columns, 23.7% under the ODE and 43.5% under PC. That is above MDLM and $\mathbb{S}$-FLM (both 18.0%) and still below the autoregressive baseline.

Method	ODE	PC
vMF	6.4	33.0
Geodesic	5.0	–
Linear	7.7	12.7
VE	2.1	1.9
+ *self-conditioning* ($p = 0.25$)		
vMF	**23.7**	**43.5**
Geodesic	14.9	–
Linear	18.3	28.5
VE	3.5	4.1
AR (sample / greedy)	53.9 / 63.3	
MDLM / DUO	18.0 / 17.2	
FLM / $\mathbb{S}$-FLM (top-k=1)	0.3 / 18.0	

Table 3: **GSM8K** test accuracy (↑, %) after training on TinyGSM, at $\mathrm{NFE} = 1024$; ODE/PC denotes the sampler. Reference values from [15]. Best per column in bold.

We further ablate the sampling budget in Appendix E.5 (Figure 1, Table 23) and the sampling temperature (Table 24). Without self-conditioning the ODE sampler saturates near $\mathrm{NFE} = 32$, and past that point only corrector steps turn further network evaluations into accuracy.

6 Conclusion

We introduced a vMF path on $\mathbb{S}^{d-1}$ for continuous generative modeling of discrete sequences. More generally, we showed that for radially symmetric paths on the sphere the continuity equation reduces to a one dimensional flux equation in the cosine similarity. For the vMF path, this reduction yields a linear ODE whose unique bounded solution gives a tractable conditional velocity. The Riemannian score follows directly from the vMF density, making ODE and predictor-corrector sampling available solely from the posterior.

We applied this construction by embedding tokens as learned points on $\mathbb{S}^{d-1}$ and training a posterior with a cross-entropy loss. This posterior determines the marginal velocity and score as well as the terminal decoding probabilities. On Sudoku, LM1B, TinyGSM, and OpenWebText, we compared vMF paths to geodesic and Euclidean baselines and showed the predictor-corrector scheme, especially for vMF, produces strong results. We believe the results show much promise and open several directions for future work.

Acknowledgments. GS acknowledges funding by the German Research Foundation (DFG) within the Excellence Cluster MATH+ and JC by project STE 571/17-2 within the DFG-SPP 2298. GK acknowledges funding by the BMBF project VIScreenPRO (ID: 100715327). The authors declare no competing interests.

References

[1] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2025. URL https://arxiv.org/abs/2303.08797.

[2] Luigi Ambrosio and Dario Trevisan. Well-posedness of Lagrangian flows and continuity equations in metric measure spaces. *Analysis and PDE*, 7(5):1179–1234, 2014.

[3] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces, 2023. URL https://arxiv.org/abs/2107.03006.

[4] Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, Chengxi Li, Chongxuan Li, Jianguo Li, Zehuan Li, Huabin Liu, Lin Liu, Guoshan Lu, Xiaocheng Lu, Yuxin Ma, Jianfeng Tan, Lanning Wei, Ji-Rong Wen, Yipeng Xing, Xiaolu Zhang, Junbo Zhao, Da Zheng, Jun Zhou, Junlin Zhou, Zhanchao Zhou, Liwang Zhu, and Yihong Zhuang. Llada2.0: Scaling up diffusion language models to 100b, 2025. URL https://arxiv.org/abs/2512.15745.

[5] Bastian Boll, Daniel Gonzalez-Alvarado, and Christoph Schnörr. Generative modeling of discrete joint distributions by e-geodesic flow matching on assignment manifolds, 2024. URL https://arxiv.org/abs/2402.07846.

[6] Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models, 2022. URL https://arxiv.org/abs/2205.14987.

[7] Ciprian Chelba, Tomás Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005, 2013. URL http://arxiv.org/abs/1312.3005.

[8] Jannis Chemseddine, Gregor Kornhardt, Richard Duong, and Gabriele Steidl. Adapting noise to data: Generative flows from 1d processes, 2026. URL https://arxiv.org/abs/2510.12636.

[9] Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries, 2024. URL https://arxiv.org/abs/2302.03660.

[10] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning, 2023. URL https://arxiv.org/abs/2208.04202.

[11] Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu. Langflow: Continuous diffusion rivals discrete in language modeling, 2026. URL https://arxiv.org/abs/2604.11748.

[12] Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds, 2025. URL https://arxiv.org/abs/2405.16441.

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

[14] Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İlkan Ceylan, Michael Bronstein, and Avishek Joey Bose. Fisher flow matching for generative modeling over discrete data, 2024. URL https://arxiv.org/abs/2405.14664.

[15] Justin Deschenaux and Caglar Gulcehre. Language modeling with hyperspherical flows, 2026. URL https://arxiv.org/abs/2605.11125.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[17] Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data, 2022. URL https://arxiv.org/abs/2211.15089.

[18] Richard Duong, Jannis Chemseddine, Peter K. Friz, and Gabriele Steidl. Telegrapher's generative model via kac flows, 2026. URL https://arxiv.org/abs/2506.20641.

[19] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching, 2024. URL https://arxiv.org/abs/2407.15595.

[20] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling, 2025. URL https://arxiv.org/abs/2505.13447.

[21] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[22] Ishaan Gulrajani and Tatsunori B. Hashimoto. Likelihood-based diffusion language models, 2023. URL https://arxiv.org/abs/2305.18619.

[23] Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes, 2025. URL https://arxiv.org/abs/2410.20587.

[24] Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling, 2025. URL https://arxiv.org/abs/2502.11564.

[25] Alexia Jolicoeur-Martineau. Less is more: Recursive reasoning with tiny networks, 2025. URL https://arxiv.org/abs/2510.04871.

[26] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

[27] Gregor Kornhardt, Jannis Chemseddine, Christian Wald, and Gabriele Steidl. Self-aware Markov models for discrete reasoning. In *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, 2026. URL https://openreview.net/forum?id=2yVYK7QQKI.

[28] Gregor Kornhardt, Moritz Piening, Jannis Chemseddine, and Gabriele Steidl. SOL: Measuring gaps between text distributions by double sliced Wasserstein metrics, 2026. URL https://arxiv.org/abs/2610.06513.

[29] Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M. Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising, 2026. URL https://arxiv.org/abs/2602.16813.

[30] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, 2016. URL https://arxiv.org/abs/1510.03055.

[31] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation, 2022. URL https://arxiv.org/abs/2205.14217.

[32] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

[33] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. URL https://arxiv.org/abs/2412.06264.

[34] Bingbin Liu, Sebastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. Tinygsm: achieving >80 URL https://arxiv.org/abs/2312.09241.

[35] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. URL https://arxiv.org/abs/2310.16834.

[36] Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, Chichester, 2000.

[37] 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, 2025. URL https://arxiv.org/abs/2502.09992.

[38] Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data, 2026. URL https://arxiv.org/abs/2406.03736.

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

[40] Peter Potaptchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S. Albergo. Discrete flow maps, 2026. URL https://arxiv.org/abs/2604.09784.

[41] Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Candi: Hybrid discrete-continuous diffusion models, 2025. URL https://arxiv.org/abs/2510.22510.

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

[43] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[44] Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, İsmail İlkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps, 2026. URL https://arxiv.org/abs/2602.12233.

[45] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models, 2024. URL https://arxiv.org/abs/2406.07524.

[46] Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality, 2025. URL https://arxiv.org/abs/2506.10892.

[47] 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, 2025. URL https://arxiv.org/abs/2412.10193.

[48] Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data, 2025. URL https://arxiv.org/abs/2406.04329.

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

[50] Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to DNA sequence design, 2024. URL https://arxiv.org/abs/2402.05841.

[51] Cedric Villani. *Optimal Transport, Old and New*. Springer, 2008.

[52] Christian Wald and Gabriele Steidl. Flow matching: Markov kernels, stochastic processes and transport plans. In *Variational and Information Flows in Machine Learning and Optimal Transport*. Springer, 2025. doi: 10.1007/978-3-031-92731-7_4.

[53] Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model, 2025. URL https://arxiv.org/abs/2506.21734.

[54] Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling, 2026. URL https://arxiv.org/abs/2503.00307.

[55] Oussama Zekri, Théo Uscidda, Nicolas Boullé, and Anna Korba. Generalized discrete diffusion from snapshots, 2026. URL https://arxiv.org/abs/2603.21342.

[56] 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 and Development in Information Retrieval*, pp. 1097–1100, 2018. URL https://arxiv.org/abs/1802.01886.

Contents

1 Introduction **1**

2 Background: Continuous Flows for Discrete Data **2**

2.1 Setting and Continuous Embedding 3

2.2 Conditional Flows of Measures 3

3 Conditional von Mises-Fisher Paths **5**

3.1 Background on the Sphere 5

3.2 von Mises-Fisher Distribution and Path 6

4 Generative Model **7**

4.1 Learning 7

4.2 Sampling 8

5 Experiments **8**

5.1 Sudoku 9

5.2 Language Modeling (LM1B and OpenWebText) 9

5.3 Mathematical Reasoning (TinyGSM) 10

6 Conclusion **10**

A Limitations and Broader Impact **15**

B Related Work **16**

C Further Theory and Proofs **16**

C.1 Preliminaries on the Sphere 17

C.2 Proofs 17

C.3 SDE Sampling on the Sphere . . . 21
C.3.1 Fokker–Planck and the flow SDE . . . 21
C.3.2 Marginal SDE for the vMF path . . . 21
C.4 Further Radially Symmetric Paths . . . 22
C.4.1 Velocity of a radial density . . . 22
C.4.2 Two further instances . . . 23
C.4.3 Proofs . . . 24

D Implementation Details **26**
D.1 Experimental Setup . . . 26
D.2 Architecture and Training . . . 27
D.3 Compute Resources . . . 27
D.4 Conditional paths . . . 28
D.5 Adaptive Schedule . . . 29

E Extended Experimental Results **29**
E.1 Sweep Choices . . . 29
E.2 Sudoku . . . 30
E.3 LM1B . . . 32
E.4 OpenWebText . . . 35
E.5 TinyGSM . . . 38
E.6 SDE Sampler . . . 39

F Additional Evaluations **41**
F.1 Inference-Seed Variability . . . 41
F.2 Across-Sample Diversity . . . 42
F.3 Distributional Distance (SOL) . . . 42

G Qualitative Samples **44**
G.1 LM1B . . . 44
G.2 OpenWebText . . . 46
G.3 TinyGSM . . . 50

A Limitations and Broader Impact

Our work shares the limitations of continuous diffusion approaches to discrete generation. The main technical specificity is that the conditional velocity and the radial CDF must be computed numerically, which requires care at high concentration and high embedding dimension to avoid underflow in the radial density. We address this with a flux-based ODE formulation and log-shifted density evaluation (Appendix D.4). We are not aware of any negative societal impact specific to this work beyond those that apply to generative modeling of discrete sequences in general. All experiments use public datasets.

B Related Work

Adapting diffusion and flow-based generative models to discrete data is an active area of research. Existing approaches either work directly in the discrete state space or embed the data into a continuous space.

Our approach is closest in spirit to CDCD [17]. We use normalized embeddings, cross-entropy training, and an adaptive schedule. The key difference is that we keep the noisy state on $\mathbb{S}^{d-1}$ throughout the generative process. In the following we highlight further related work.

Discrete state space. Discrete state space methods define Markov processes over the finite vocabulary. In general these models split into masked diffusion [45, 6, 48, 37] and uniform diffusion [3, 47, 46], which converge to a special mask token or uniform random tokens respectively. D3PM [3] designed discrete time forward processes. This was extended to continuous time based on the CTMC by [6]. SEDD [35] builds a score entropy loss. [45, 38, 48] introduce theoretical results and LLaDA [37] scaled masked diffusion to 8 billion parameters. Recently, GDDS [55] uses GPT-2 embeddings to induce semantic structure in discrete noising. Continuous noising in the learned embedding space follows a related intuition.

Continuous state space. Continuous embedding methods assign each token a vector and apply noise in the embedding space. Diffusion-LM [31] uses Gaussian DDPM with learned embeddings. [22] uses a discrete time process without normalized embeddings. [29, 44] use a linear interpolant to the one-hot target in the large $\mathbb{R}^V$ space with a softmax denoiser and flow distillation for few-step language generation. LangFlow [11] works in the continuous embedding space and introduces an ODE based NLL. In all of these the noisy state lives in $\mathbb{R}^d$ or $\mathbb{R}^V$.

A separate line of work defines noise processes on the probability simplex or on spheres of dimension $|\mathcal{V}| - 1$. Generative Assignment Flows [5] construct flows on the probability simplex by defining linear paths in logit space and mapping them to simplex-valued distributions via softmax. Dirichlet Flow Matching [50] constructs paths from Dirichlet distributions on the simplex. RDLM [24] uses the diffeomorphism between the simplex and the positive orthant of the sphere $\mathbb{S}^{|\mathcal{V}|-1}$ for language modeling and derives an NLL bound via Girsanov's theorem. To simulate the bridge SDE, they split the sphere into a product of smaller spheres, since the bridge on high dimensional spheres is difficult. [14, 12] map the simplex to the positive orthant of $\mathbb{S}^{|\mathcal{V}|-1}$ via the standard diffeomorphism and construct geodesic flows there. Concurrent to our work, [15] also propose flows on the sphere for language modeling; they focus on the geodesic path and a custom spherical network architecture.

Mixing discrete and continuous methods. A different line of work is the combination of discrete and continuous diffusion models for categorical data. [41] propose CANDI, a hybrid discrete–continuous diffusion model with a structured forward noising process that combines a discrete masking process with Gaussian corruption. During reverse sampling, the model starts from a fully corrupted Gaussian state: positions selected to become clean are sampled from the model's token distribution and then carried forward/fixed, while the remaining noisy positions continue to be updated by the continuous reverse ODE. Duo by [46] shows that uniform-state discrete diffusion arises as the argmax push-forward of an underlying Gaussian diffusion process. They exploit this duality to improve training and enable much faster few-step sampling via discrete consistency distillation.

Sudoku. For non-generative neural network approaches on Sudoku-Extreme, see [53, 27, 25].

C Further Theory and Proofs

This appendix collects background on the sphere and the geodesic path, the proofs of all results of the main text, SDE sampling, and further radially symmetric paths.

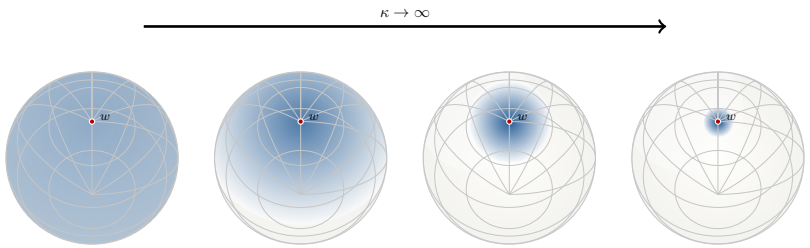


Figure 2: Illustration of von Mises–Fisher density on $\mathbb{S}^2$ for increasing concentration κ. At $\kappa = 0$ (left) the density is uniform. As κ grows the mass concentrates around the mean direction w, collapsing to a point mass at w as $\kappa \to \infty$.

C.1 Preliminaries on the Sphere

Remark C.1 (Geodesic Path)**.** *Using the geodesic between two points, one can construct the slerp path*

$$X_t = \frac{\sin((1-t)\theta)}{\sin\theta} X_0 + \frac{\sin(t\theta)}{\sin\theta} X_1, \quad t \in [0,1], \tag{19}$$

where $\theta := \arccos(\langle X_0, X_1 \rangle)$*,* $X_0 \sim \mathcal{U}(\mathbb{S}^{d-1})$*. With the linear interpolation on* $\mathbb{R}^d$ *from Examples 2.1 and 2.3 in mind, this is its spherical counterpart. This path is sometimes called spherical linear interpolation* ***(slerp)****. Using this path for generative modeling via flows is a straightforward application of Riemannian Flow Matching [9].*

Differential operators. Complementing Section 3: on $\mathbb{S}^{d-1}$, for a differentiable vector field $\tilde{u} : \mathbb{R}^d \to \mathbb{R}^d$ and $u = \tilde{u}\big|_{\mathbb{S}^{d-1}}$, the divergence is defined by

$$\operatorname{div}_{\mathbb{S}^{d-1}} u(x) = \operatorname{tr}\left(\nabla \tilde{u}(x)\right) - x^\top \nabla \tilde{u}(x)\, x$$

and the Laplace-Beltrami operator becomes

$$\Delta_{\mathbb{S}^{d-1}} g(x) = \operatorname{tr}\left((\mathrm{I}_d - xx^\top)\nabla^2 \tilde{g}(x)\right) - (d-1)\langle \nabla \tilde{g}(x), x \rangle.$$

The geodesic interpolation from Remark C.1 and the vMF path from Theorem 3.4 both define conditional paths starting from the uniform distribution on $\mathbb{S}^{d-1}$. We want to highlight that they differ in how quickly the noisy sample reveals the target with increasing dimension d. We quantify this by comparing the expected cosine similarity under both paths.

Proposition C.2. *Let* $X_0 \sim \mathcal{U}(\mathbb{S}^{d-1})$ *and* $w \in \mathbb{S}^{d-1}$ *be arbitrary fixed. Then the conditional geodesic interpolation path* (slerp) *in* (19) *fulfills*

$$\mathbb{E}_{x_t \sim \mathrm{slerp}_t(x_0, w)}\left[\langle w, x_t \rangle\right] \to \sin\left(\tfrac{\pi}{2}\, t\right) \quad \textit{as} \quad d \to \infty, \tag{20}$$

the vMF path (11)*,*

$$\mathbb{E}_{x_t \sim \mathrm{vMF}(w, \kappa_t)}\left[\langle w, x_t \rangle\right] = A_d(\kappa_t) = \frac{\kappa_t}{d} + O(\kappa_t^3). \tag{21}$$

Under geodesic interpolation, the expected cosine similarity at any fixed $t > 0$ is bounded away from zero independently of d. Under the vMF path, reaching the same signal level requires κ_t proportional to d. The informative range of the noise parameter therefore grows with d.

C.2 Proofs

Proposition 2.2. *Let* $\mathbf{w} = (w^l)_{l=1}^L \in \mathcal{W}^L$ *be fixed, and let* $p_t^l(\cdot \mid w^l)$ *be absolutely continuous measures on* $\mathcal{M}$ *with velocities* $v_t^l(\cdot \mid w^l)$ *satisfying the conditional continuity equation* (1) *on* $\mathcal{M}$ *for each* $l = 1, \ldots, L$*. Then the product density on* $\mathcal{M}^L$*,*

$$p_t(\mathbf{x} \mid \mathbf{w}) = \prod_{l=1}^{L} p_t^l(x^l \mid w^l), \qquad \mathbf{x} = (x^1, \ldots, x^L), \tag{2}$$

with the velocity field $v_t(\mathbf{x} \mid \mathbf{w}) := \left(v_t^1(x^1 \mid w^1), ..., v_t^L(x^L \mid w^L)\right)$ *satisfies the continuity equation on* $\mathcal{M}^L$ *for* $t \in (0,1)$*.*

Proof. Using the product rule we obtain

$$\partial_t p_t(\mathbf{x}|\mathbf{w}) = \sum_{l=1}^{L} \Big(\prod_{j \neq l}^{L} p_t^j(x^j|w^j) \Big) \, \partial_t p_t^l(x^l|w^l)$$

and further by the continuity equation for p_t^l,

$$\partial_t p_t(\mathbf{x}|\mathbf{w}) = -\sum_{l=1}^{L} \Big(\prod_{j \neq l}^{L} p_t^j(x^j|w^j) \Big) \operatorname{div}_{\mathcal{M}}^{L} \left(p_t^l(x^l|w^l) \, v_t^l(x^l|w^l) \right).$$

Since the factor $\prod_{j \neq l} p_t^j(x^j|w^j)$ does not depend on x^l, we can pull it into the divergence operator, which acts only on the l-th variable which results in

$$\partial_t p_t(\mathbf{x}|\mathbf{w}) = -\sum_{l=1}^{L} \operatorname{div}_{\mathcal{M}}^{l} \Big(\Big(\prod_{j=1}^{L} p_t^j(x^j|w^j) \Big) v_t^l(x^l|w^l) \Big).$$

Recognizing that $p_t(\mathbf{x}|\mathbf{w}) = \prod_{j=1}^{L} p_t^j(x^j|w^j)$, this yields

$$\partial_t p_t(\mathbf{x}|\mathbf{w}) = -\sum_{l=1}^{L} \operatorname{div}_{\mathcal{M}} \left(p_t(\mathbf{x}|\mathbf{w}) \, v_t^l(x^l|w^l) \right).$$

Finally, using that the divergence on the product manifold $\mathcal{M}^L$ is given as $\operatorname{div}_{\mathcal{M}^L}(u) \coloneqq \sum_{l=1}^{L} \operatorname{div}_{\mathcal{M}}(u^l)$, we conclude

$$\partial_t p_t(\mathbf{x}|\mathbf{w}) = -\operatorname{div}_{\mathcal{M}^L} \left(p_t(\mathbf{x}|\mathbf{w}) \, v_t(\mathbf{x}|\mathbf{w}) \right).$$

Theorem 3.1. *Let (p_t, v_t), $t \in I$ be a sequence of radially symmetric pairs of densities and velocities of the form $p_t(x) = \bar{p}_t(\langle x, w \rangle)$, resp. $v_t(x) = \psi_t(\langle x, w \rangle) \mathrm{P}_x(w)$, where $\bar{p}_t \in C^1[I \times (-1,1)]$ and $\psi_t \in C^1(-1,1)$ for every $t \in I$. Let f_t, $t \in I$ be the corresponding densities defined by* (8)*. Then (p_t, v_t), $t \in I$ satisfy the continuity equation*

$$\partial_t p_t + \operatorname{div}_{\mathbb{S}^{d-1}}(p_t \, v_t) = 0, \quad t \in I, x \in \mathbb{S}^{d-1} \setminus \{\pm w\} \tag{9}$$

if and only if (f_t, ψ_t) satisfy the one-dimensional flux equation

$$\partial_t f_t + \partial_s \big(f_t \, \psi_t \cdot (1 - s^2) \big) = 0, \quad t \in I, s \in (-1,1). \tag{10}$$

Proof. Let $s \coloneqq \langle w, x \rangle$. Using the divergence on the sphere we want to reformulate (9). With

$$u_t(x) \coloneqq p_t(x) v_t(x) = g_t(s)(w - sx), \quad g_t \coloneqq \bar{p}_t \psi_t,$$

we obtain

$$\nabla u_t(x) = g_t'(s) w (w - sx)^\top - g_t(s) \big(s \mathrm{I}_d + x w^\top \big),$$

and further

$$\begin{aligned} \operatorname{tr}\left(\nabla u_t(x) \right) &= g_t'(s)(1 - s^2) - g_t(s)(d+1)s, \\ x^\top \nabla u_t(x) \, x &= -2 s g_t(s). \end{aligned}$$

Thus, the continuity equation (9) can be rewritten as

$$\begin{aligned} 0 &= \partial_t p_t + \operatorname{div}_{\mathbb{S}^{d-1}}(p_t \, v_t) = \partial_t \bar{p}_t + \operatorname{tr}(\nabla u_t) - x^\top \nabla u_t \, x \\ &= \partial_t \bar{p}_t + g_t'(1 - s^2) - g_t(d-1)s. \end{aligned} \tag{22}$$

On the other hand, the flux equation in (10) becomes

$$\begin{aligned} 0 &= \partial_t f_t + \partial_s \big(g_t \, (1 - s^2)^{(d-1)/2} \big) \\ &= \Big((1 - s^2)^{(d-3)/2} \partial_t \bar{p}_t + g_t'(1 - s^2)^{(d-1)/2} - g_t(d-1)s(1 - s^2)^{(d-3)/2} \Big). \end{aligned}$$

Dividing by $(1 - s^2)^{(d-3)/2}$, $s \in (-1,1)$ yields equation (22) and we are done. □

Lemma 3.2. *Let f_t, $t \in I$ be defined by* (12)*. Then ψ_t, $t \in I$ solves the linear ODE*

$$(1 - s^2) \, \psi_t' + \big(\kappa_t (1 - s^2) - (d-1)s \big) \, \psi_t = \big(A_d(\kappa_t) - s \big) \, \dot{\kappa}_t, \quad s \in (-1,1) \tag{13}$$

if and only if (f_t, ψ_t), $t \in I$ fulfill the flux equation (10)*.*

Proof. For the vMF setting, the flux equation (10) reads as

$$\begin{aligned} 0 &= \partial_t f_t + \partial_s\big(f_t\,\psi_t\,(1-s^2)\,\big) \\ &= (1-s^2)^{(d-3)/2}\partial_t\Big(C_d(\kappa_t)\,\exp(\kappa_t\,s)\Big) + C_d(\kappa_t)\partial_s\Big((1-s^2)^{(d-1)/2}\psi_t(s)\exp(\kappa_t\,s)\Big). \end{aligned} \tag{23}$$

Concerning the first summand, we obtain

$$\partial_t\Big(C_d(\kappa_t)\,\exp(\kappa_t\,s)\Big) = C_d'(\kappa_t)\,\dot\kappa_t\,\exp(\kappa_t\,s) + s\dot\kappa_t C_d(\kappa_t)\,\exp(\kappa_t\,s).$$

Noting that

$$\mathcal{I}_{d/2-1}'(\kappa_t) = \mathcal{I}_{d/2}(\kappa_t) + \frac{d/2-1}{\kappa_t}\mathcal{I}_{d/2-1}(\kappa_t), \tag{24}$$

we conclude

$$\begin{aligned} C_d'(\kappa_t) &= \frac{(\frac{d}{2}-1)\kappa_t^{d/2-2}}{(2\pi)^{d/2}\mathcal{I}_{d/2-1}(\kappa_t)} - \frac{\kappa_t^{d/2-1}\mathcal{I}_{d/2-1}'(\kappa_t)}{(2\pi)^{d/2}\mathcal{I}_{d/2-1}^2(\kappa_t)} \\ &= \frac{(\frac{d}{2}-1)}{\kappa_t}C_d(\kappa_t) - \frac{(\frac{d}{2}-1)}{\kappa_t}C_d(\kappa_t) - A_d(\kappa_t)C_d(\kappa_t) \\ &= -A_d(\kappa_t)C_d(\kappa_t) \end{aligned}$$

and consequently

$$\partial_t\Big(C_d(\kappa_t)\,\exp(\kappa_t\,s)\Big) = C_d(\kappa_t)\,\exp(\kappa_t\,s)\dot\kappa_t\,\big(-A_d(\kappa_t)+s\big)\,. \tag{25}$$

The second summand can be rewritten as

$$\begin{aligned} \partial_s\Big((1-s^2)^{(d-1)/2}\psi_t(s)\exp(\kappa_t\,s)\Big) &= -(d-1)(1-s^2)^{(d-3)/2}\,s\,\psi_t(s)\exp(\kappa_t\,s) \\ &+ (1-s^2)^{(d-1)/2}\psi_t'(s)\exp(\kappa_t\,s) + \kappa_t(1-s^2)^{(d-1)/2}\psi_t(s)\exp(\kappa_t\,s) \end{aligned} \tag{26}$$

Plugging in (25) and (26) into (23) and dividing by $C_d(\kappa_t)(1-s^2)^{(d-3)/2}\exp(\kappa_t\,s)$ yields

$$\dot\kappa_t\,\big(A_d(\kappa_t)-s\big) = \big(-(d-1)s + (1-s^2)\kappa_t\big)\,\psi_t(s) + (1-s^2)\psi_t'(s).$$

This finishes the proof. □

Lemma 3.3. *For any $t \in I$, the ODE* (13) *has a unique solution ψ_t in $C[-1,1]$, which is automatically also in $C^\infty(-1,1)$ with boundary values $\psi_t(1) = \frac{(1-A_d(\kappa_t))\dot\kappa_t}{d-1}$, and $\psi_t(-1) = \frac{(1+A_d(\kappa_t))\dot\kappa_t}{d-1}$.*

Proof. The general solution of the linear ODE

$$\psi' + \left(\kappa - (d-1)\frac{s}{1-s^2}\right)\psi = (A_d - s)\,\frac{\dot\kappa}{1-s^2}, \quad s \in (-1,1)$$

(with skipped index t for convenience) can be computed by the method of integrating factors as

$$\psi(s) = \exp(-\kappa s)(1-s^2)^{-(d-1)/2}\Big(C + \dot\kappa\int_{-1}^{s}\exp(\kappa y)(1-y^2)^{(d-3)/2}(A_d-y)\,\mathrm{d}y\Big),$$

where the constant C is determined by conditions on ψ. We are interested in solutions with finite $\lim_{s\to\pm1}\psi(s)$. Since $(1-s^2)^{-(d-1)/2}$ goes to infinity as $s \to \pm1$, this can only be achieved for $C = 0$ and the solution

$$\psi(s) = \dot\kappa\,\exp(-\kappa s)(1-s^2)^{-(d-1)/2}\int_{-1}^{s}\exp(\kappa y)(1-y^2)^{(d-3)/2}(A_d-y)\,\mathrm{d}y. \tag{27}$$

Indeed the integral behaves for s near -1 as $(s+1)^{(d-1)/2}$, so that $\psi(-1)$ is finite and the concrete value follows immediately from (13). On the other hand, considering $s \to 1$, we have

$$\int_{-1}^{1}\exp(\kappa y)(1-y^2)^{(d-3)/2}(A_d-y)\,\mathrm{d}y = A_dG(\kappa) - G'(\kappa)$$

where

$$G(\kappa) \coloneqq \int_{-1}^{1} \exp(\kappa y)(1-y^2)^{(d-3)/2}\,\mathrm{d}y = \mathcal{I}_{d/2-1}(\kappa)\alpha(\kappa), \quad \alpha(\kappa) \coloneqq \frac{\Gamma((d-1)/2)\sqrt{\pi}}{(\kappa/2)^{(d/2-1)}}$$

and using (24) and the definition of A_d further

$$A_d G(\kappa) - G'(\kappa) = \mathcal{I}_{d/2}(\kappa)\alpha(\kappa) - \big(\mathcal{I}_{d/2}(\kappa) + \frac{d/2-1}{\kappa}\mathcal{I}_{d/2-1}(\kappa)\big)\alpha(\kappa) - \mathcal{I}_{d/2-1}(\kappa)\alpha'(\kappa) = 0.$$

The asymptotics of the integral in (27) as $s \to 1$ is $(1-s)^{(d-1)/2}$. Finally, the representation (27) is a product of functions that are smooth on $(-1,1)$, so that $\psi \in C^\infty(-1,1)$ and we are done. □

Theorem 3.4. *Let $p_t(\cdot|w)$ be the conditional vMF path defined by* (11) *and let ψ_t, $t \in I$ be the unique $C[-1,1]$ solutions of the ODE* (13)*. Then $p_t(\cdot|w)$ together with the velocities*

$$v_t(x|w) = \psi_t(\langle w, x\rangle)\,\mathrm{P}_x(w)$$

fulfill the continuity equation (9) *on $\mathbb{S}^{d-1}$.*

Proof. By the previous proofs we have for the corresponding expressions in (10), (13) and (9) that

$$\begin{aligned}(10) &= C_d(\kappa_t)(1-s^2)^{(d-3)/2}\exp(\kappa_t s)\cdot(13)\\ &= (1-s^2)^{(d-3)/2}\cdot(9),\end{aligned}$$

where $(9) = R(s) = \partial_t \bar{p}_t(s) + g_t'(s)(1-s^2) - g_t(s)(d-1)s$ and $g_t(s) = C_d(\kappa_t)\exp(\kappa_t s)\psi_t(s)$. This gives the assertion for $s \in (-1,1)$. For the chosen ψ, we have that R is continuous on $[-1,1]$ (note that $\psi_t'(\pm 1)$ is also finite) and since R is zero on $(-1,1)$ by (13), it remains zero at the boundary. So we have an extension of the relation to the whole $\mathbb{S}^{d-1}$. □

Proposition C.2. *Let $X_0 \sim \mathcal{U}(\mathbb{S}^{d-1})$ and $w \in \mathbb{S}^{d-1}$ be arbitrary fixed. Then the conditional geodesic interpolation path* (slerp) *in* (19) *fulfills*

$$\mathbb{E}_{x_t \sim \mathrm{slerp}_t(x_0,w)}\big[\langle w, x_t\rangle\big] \to \sin\big(\tfrac{\pi}{2}\,t\big) \quad \textit{as} \quad d \to \infty, \tag{20}$$

the vMF path (11)*,*

$$\mathbb{E}_{x_t \sim \mathrm{vMF}(w,\kappa_t)}\big[\langle w, x_t\rangle\big] = A_d(\kappa_t) = \frac{\kappa_t}{d} + O(\kappa_t^3). \tag{21}$$

Proof. Since (21) follows from the expansion of A_d in Section 3, it remains to verify (20). For

$$X_t = \frac{\sin((1-t)\theta)}{\sin\theta}\,X_0 + \frac{\sin(t\theta)}{\sin\theta}\,w,$$

let $Z_t \coloneqq \langle w, X_t\rangle$. By Remark C.1, we know that

$$Z_t = \cos\left((1-t)\Theta\right), \quad \Theta = \arccos\left(\langle w, X_0\rangle\right).$$

It is straightforward to check that Z_0 is a random variable on $[-1,1]$ with expectation value zero and variance $1/d$. Thus, $Z_0 \to 0$ as $d \to \infty$ in probability and consequently $\Theta \to \pi/2$ in probability as $d \to \infty$, and finally

$$Z_t \to \cos\left((1-t)\frac{\pi}{2}\right) \quad \text{as} \quad d \to \infty$$

in probability. Then the bounded convergence theorem gives

$$\mathbb{E}[Z_t] \to \cos\left((1-t)\frac{\pi}{2}\right) = \sin\left(\frac{\pi}{2}\,t\right) \quad \text{as} \quad d \to \infty. \quad \square$$

Recall that the Kullback–Leibler divergence of two pmfs p, q on a discrete space of size N with $p(k) = 0$ whenever $q(k) = 0$ is $\mathrm{KL}(p\|q) = \sum_{k=1}^{N}\big(p(k)\log p(k) - p(k)\log q(k)\big)$ and their cross-entropy and entropy by $\mathrm{CE}(p,q) = -\sum_{k=1}^{N} p(k)\log q(k) = \mathrm{KL}(p\|q) + \mathrm{H}(p)$, $\mathrm{H}(p) = -\sum_{k=1}^{N} p(k)\log p(k)$. Since $\mathrm{KL}(p\|q) \geq 0$ with equality if and only if $p = q$, the cross-entropy $\mathrm{CE}(p,q)$ is minimized in q exactly at $q = p$.

Proposition 4.1. *The minimizer of* (15) *is the per-position marginal posterior $p_t^l(\cdot|\mathbf{x})$ from* (4)*.*

Proof. We show that the loss (15) decomposes as

$$\mathcal{L}(\theta) = \sum_{l=1}^{L} \mathbb{E}_{t\sim\mathcal{U}(0,1),\,\mathbf{x}_t\sim p_t}\Big[\mathrm{CE}\big(p_t^l(\cdot|\mathbf{x}_t),\, p_{\theta,t}^l(\cdot|\mathbf{x}_t)\big)\Big], \tag{28}$$

from which the claim follows by the minimization property above. Write $\mathcal{L} = \sum_{l=1}^{L} \mathcal{L}^l$, where

$$\mathcal{L}^l(\theta) \coloneqq -\int\int \sum_{\mathbf{w}\in\mathcal{W}^L} \log p_{\theta,t}^l(w^l|\hat{x}_t^l)\, p_t(\mathbf{x}_t|\mathbf{w})\, p_{\text{data}}(\mathbf{w})\, \mathrm{d}\mathbf{x}_t\, \mathrm{d}t.$$

By Bayes' rule and since $\log p_\theta^l(w^l|\hat{x}_t^l)$ depends on $\mathbf{w}$ only through w^l, this can be rewritten as

$$\begin{aligned}
\mathcal{L}^l(\theta) &= -\int\int \sum_{\mathbf{w}\in\mathcal{W}^L} \log p_{\theta,t}^l(w^l|\hat{x}_t^l)\, p_t(\mathbf{w}|\mathbf{x}_t)\, p_t(\mathbf{x}_t)\, \mathrm{d}\mathbf{x}_t\, \mathrm{d}t \\
&= -\int\int \sum_{w^l\in\mathcal{W}} \log p_{\theta,t}^l(w^l|\hat{x}_t^l) \sum_{\substack{w^j\in\mathcal{W}\\ j\neq l}} p_t(\mathbf{w}|\mathbf{x}_t)\, p_t(\mathbf{x}_t)\, \mathrm{d}\mathbf{x}_t\, \mathrm{d}t \\
&= \mathbb{E}_{t,\,\mathbf{x}_t\sim p_t}\Big[\mathrm{CE}\big(p_t^l(\cdot|\mathbf{x}_t),\, p_{\theta,t}^l(\cdot|\hat{x}_t^l)\big)\Big],
\end{aligned}$$

which is (28). □

C.3 SDE Sampling on the Sphere

C.3.1 Fokker–Planck and the flow SDE

The same flow of measures from Section 2 admits a stochastic realization once the score is available. Substituting $v_t = u_t - \frac{\sigma^2}{2}\nabla_{\mathcal{M}} \log p_t$ for $\sigma > 0$ and using $\mathrm{div}_{\mathcal{M}}(\nabla_{\mathcal{M}} g) = \Delta_{\mathcal{M}} g$, the continuity equation (CE) becomes the *Fokker–Planck equation*

$$\partial_t p_t + \mathrm{div}_{\mathcal{M}}(p_t u_t) - \tfrac{\sigma^2}{2}\Delta_{\mathcal{M}} p_t = 0. \tag{29}$$

The solution p_t is the law of a random variable X_t determined by the *flow SDE*

$$\mathrm{d}X_t = u_t(X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}B_t, \tag{Flow SDE}$$

where B_t is the Brownian motion on $\mathcal{M}$. Once v_t and $\nabla_{\mathcal{M}} \log p_t$ are known, sampling from p_1 can be done either deterministically via the flow ODE (Flow ODE) or stochastically via the SDE (Flow SDE).

C.3.2 Marginal SDE for the vMF path

The score (17) enables stochastic sampling along the same marginal path. The flow SDE on $(\mathbb{S}^{d-1})^L$ reads per position

$$\mathrm{d}X_t^l = \Big[v_t^l(\mathbf{X}_t) + \tfrac{\sigma^2}{2}\nabla_{\mathbb{S}^{d-1},x^l} \log p_t(\mathbf{X}_t)\Big]\,\mathrm{d}t + \sigma\,\mathrm{d}B_t^l, \quad l = 1,\dots,L, \tag{30}$$

where $\sigma > 0$ is a free diffusion coefficient and B_t^l is Brownian motion on $\mathbb{S}^{d-1}$. The derivation holds verbatim for a time-dependent diffusion coefficient σ_t, since the Fokker–Planck cancellation between the score part of the drift and the diffusion term in (29) holds pointwise in t provided the drift uses the matching $\sigma_t^2/2$; in the experiments (Appendix E.6) we use $\sigma_t = \sigma(1-u)^q$ with the path progress variable u of Appendix E.1 ($u = \kappa_t/\kappa_{\max}$ for vMF). Substituting (16) and (17), the drift of (30) is itself a posterior-weighted tangent sum,

$$v_t^l + \tfrac{\sigma^2}{2}\nabla_{\mathbb{S}^{d-1},x^l} \log p_t = \sum_{k=1}^{N} p_t^l(w_k|\mathbf{x}) \Big(\dot{\kappa}_t\, \tilde{\psi}_t(\langle w_k, x^l\rangle) + \tfrac{\sigma^2}{2}\,\kappa_t\Big)\, \mathrm{P}_{x^l}(w_k), \tag{31}$$

with the same posteriors and tangent projections as the ODE. No separate score network is required. Algorithm 1 discretizes (30) via Euler–Maruyama with a tangent retraction.

Algorithm 1 SDE sampling on $(\mathbb{S}^{d-1})^L$

Require: Time grid $0 = t_0 < t_1 < \cdots < t_S = 1$, diffusion coefficient $\sigma > 0$

Sample $x_0^l \sim \mathcal{U}(\mathbb{S}^{d-1})$ independently for $l = 1, \dots, L$

for $n = 0, \dots, S-1$ **do**

$\Delta t \leftarrow t_{n+1} - t_n$

$(\hat{x}^1, \dots, \hat{x}^L) \leftarrow T_\theta(x_n^1, \dots, x_n^L)$

$p_k^l \leftarrow p_\theta^l(w_k | \hat{x}^l)$ for all l, k

$$d^l \leftarrow \sum_{k=1}^{N} p_k^l \left(\dot{\kappa}_{t_n} \, \tilde{\psi}_{t_n}(\langle w_k, x_n^l \rangle) + \tfrac{\sigma^2}{2} \, \kappa_{t_n} \right) \mathrm{P}_{x_n^l}(w_k) \qquad \triangleright \text{(31)}$$

Sample $\xi^l \sim \mathcal{N}(0, I_d)$ for $l = 1, \dots, L$

$$x_{n+1}^l \leftarrow \frac{x_n^l + \Delta t \, d^l + \sigma \sqrt{\Delta t} \, \mathrm{P}_{x_n^l}(\xi^l)}{\| x_n^l + \Delta t \, d^l + \sigma \sqrt{\Delta t} \, \mathrm{P}_{x_n^l}(\xi^l) \|} \text{ for } l = 1, \dots, L$$

end for

$(\hat{x}^1, \dots, \hat{x}^L) \leftarrow T_\theta(x_S^1, \dots, x_S^L)$

return $\hat{y}^l = \arg\max_{k=1,\dots,N} \langle w_k, \hat{x}^l \rangle + b_k$ for $l = 1, \dots, L$

C.4 Further Radially Symmetric Paths

Beyond the vMF path of Section 3, the reduction of Theorem 3.1 can be read in two directions: fixing the radial density $\bar{p}_t$ and solving the flux equation (10) for the velocity, or fixing the velocity and solving for the density. In this appendix we first state the closed velocity formula obtained in the density-first direction, and then instantiate it for two further conditional paths, the power spherical path and the spherical Cauchy path, summarized together with the geodesic and vMF paths in Table 4.

C.4.1 Velocity of a radial density

Fixing the radial density $\bar{p}_t$ and solving for the velocity gives a closed formula, obtained by integrating the flux equation (10) once in s.

Proposition C.3 (Velocity of a radial path)**.** *Let $(p_t)_{t \in I}$ be radially symmetric densities with profiles $\bar{p}_t \in C^1(I \times (-1, 1))$, and let f_t be as in* (8)*. Assume $f_t(s)\,\psi_t(s)\,(1-s^2) \to 0$ as $s \to -1$. Then the velocity coefficient for which the pair (p_t, v_t) with $v_t(x) = \psi_t(\langle w, x \rangle)\,\mathrm{P}_x(w)$ satisfies the continuity equation* (9) *is*

$$\psi_t(s) = -\frac{1}{f_t(s)\,(1-s^2)} \int_{-1}^{s} \partial_t f_t(y) \, \mathrm{d}y. \tag{32}$$

The corresponding radial score is immediate from (6), since the Riemannian gradient of a radial density is

$$\nabla_{\mathbb{S}^{d-1}} \log p_t(x) = (\log \bar{p}_t)'(\langle w, x \rangle)\,\mathrm{P}_x(w). \tag{33}$$

The integral in (32) is the unique velocity whose flux vanishes at both endpoints; it is smooth on $(-1, 1)$, and its endpoint values follow from the flux equation once the profile is fixed. Conversely, fixing the velocity coefficient ψ_t and solving the flux equation (10) for f_t recovers the radial density transported by that velocity.

When the profile is indexed by a scalar concentration schedule, $\bar{p}_t = \bar{p}(\cdot\,; c_t)$ with c_t monotone and $c_0 = 0$, the time derivative factorizes as $\partial_t f_t = \dot{c}_t \, \partial_c f$, and so does the velocity,

$$\psi_t(s) = \dot{c}_t \, \tilde{\psi}(s; c_t), \qquad \tilde{\psi}(s; c) = -\frac{1}{f(s; c)(1-s^2)} \int_{-1}^{s} \partial_c f(y; c) \, \mathrm{d}y,$$

where $\tilde{\psi}$ depends on the concentration value alone, not on the schedule. For the vMF path, specializing to $c = \kappa$ recovers exactly the representation (40) of Remark D.1. The factor $\tilde{\psi}$ is tabulated once in (s, c) and evaluated by a numerically stable cumulative sum (Section D.4), and in concentration-space integration $\dot{c}_t$ cancels, so the schedule enters sampling only through the step sizes Δc. We use the density-first direction for both the power spherical and the spherical Cauchy paths below; for the spherical Cauchy path the resulting velocity coefficient is constant in s.

C.4.2 Two further instances

Each of the following paths has velocity $\psi_t(s)\,\mathrm{P}_x(w)$ and, by (33), score $(\log \bar{p}_t)'(s)\,\mathrm{P}_x(w)$ with $s = \langle w, x\rangle$.

Power spherical. As a further density-first instance, we take the power spherical family, whose radial profile is a power of $1 + s$. For a monotone schedule α_t with $\alpha_0 = 0$ we set

$$p_t(\cdot \mid w) \coloneqq C(\alpha_t)\,(1 + \langle w, \cdot\rangle)^{\alpha_t}, \qquad C(\alpha)^{-1} = 2^{\,d-1+\alpha}\,\pi^{(d-1)/2}\,\frac{\Gamma(\alpha + \frac{d-1}{2})}{\Gamma(\alpha + d - 1)}, \tag{34}$$

so that p_0 is uniform. This is radially symmetric, and evaluating the velocity formula (32) is equivalent to a linear ODE.

Lemma C.4 (Power spherical velocity)**.** *For the profile* (34)*, the velocity coefficient in* (32) *is the unique solution bounded as $s \to 1$ of*

$$(1 - s^2)\,\psi_t' + \big[\alpha_t(1 - s) - (d - 1)s\big]\,\psi_t = -\dot{\alpha}_t\Big[\log\tfrac{1+s}{2} + h(d - 1 + \alpha_t) - h\big(\tfrac{d-1}{2} + \alpha_t\big)\Big], \tag{35}$$

where $h = (\log\Gamma)'$ is the digamma function, with boundary value $\psi_t(1) = \frac{\dot{\alpha}_t}{d-1}\big[h(d - 1 + \alpha_t) - h(\frac{d-1}{2} + \alpha_t)\big]$.

By (33) the score is $\nabla_{\mathbb{S}^{d-1}} \log p_t(x) = \frac{\alpha_t}{1+\langle w,x\rangle}\,\mathrm{P}_x(w)$. As for vMF the coefficient ψ_t has no elementary closed form; it is precomputed on a (μ, α) grid, like the vMF table on (μ, κ), via the flux representation (32), using the same max-shifted log-density trick to avoid underflow, with the analytical boundary value $\psi_t(1)$ from Lemma C.4 handled as for vMF (Section D.4). Unlike vMF the profile vanishes at the antipode $s = -1$, where ψ_t diverges logarithmically; the velocity field itself stays bounded, since $|\mathrm{P}_x(w)| = \sqrt{1 - s^2}$ gives velocity magnitude $|\psi_t(s)|\sqrt{1 - s^2} \to 0$ as $s \to -1$. Training-time samples $x \sim p_t(\cdot \mid w)$ use the same radial-plus-tangent decomposition as vMF, with μ drawn by inverse CDF of $(1 - \mu^2)^{(d-3)/2}(1 + \mu)^{\alpha_t}$.

Spherical Cauchy. The *Cauchy distribution* on $\mathbb{S}^{d-1}$ with location $w \in \mathbb{S}^{d-1}$ and concentration $r \in [0, 1)$ has the density

$$\varphi(x; w, r) = \frac{1}{\omega_{d-1}}\,\frac{(1 - r^2)^{d-1}}{(1 + r^2 - 2r\langle w, x\rangle)^{d-1}} \qquad \omega_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)},$$

For $r = 0$, we obtain the uniform distribution, while $r \to 1$ pushes the distribution to w. It is the image of the multivariate Cauchy distribution under stereographic projection. Setting $a \coloneqq (2r)/(1 + r^2)$, this can be rewritten as

$$\varphi(x; w, r) = \frac{1}{\omega_{d-1}}\,\frac{(1 - a^2)^{(d-1)/2}}{\big(1 - a\langle w, x\rangle\big)^{d-1}}.$$

With a schedule $a_t : [0, \infty) \to [0, 1)$, we consider

$$p_t(x \mid w) = \frac{1}{\omega_{d-1}}\,\frac{(1 - a_t^2)^{(d-1)/2}}{\big(1 - a_t\langle w, x\rangle\big)^{d-1}}. \tag{36}$$

By the following proposition, the corresponding velocity coefficient in the CE (10) is especially simple, since it does not depend on s.

Proposition C.5 (Spherical Cauchy velocity)**.** *For the profile* (36)*, the velocity coefficient ψ in* (32) *is the unique solution which is bounded as $s \to 1$ of*

$$\partial_s\psi_t(s) + (d - 1)\left(\frac{a_t}{1 - a_t s} - \frac{s}{1 - s^2}\right)\psi_t(s) = -\frac{(d - 1)\dot{a}_t}{1 - s^2}\left(\frac{s}{1 - a_t s} - \frac{a_t}{1 - a_t^2}\right). \tag{37}$$

This solution is given by

$$\psi_t(s) = \dot{\beta}_t = \frac{\dot{a}_t}{1 - a_t^2},$$

where

$$\beta_t \coloneqq \tfrac{1}{2}\log\frac{1 + a_t}{1 - a_t}, \quad a_t \coloneqq \tanh\beta_t. \tag{38}$$

Table 4: Radially symmetric conditional paths on $\mathbb{S}^{d-1}$ as instances of Theorem 3.1.

Path	Radial profile $\bar{p}_t(s)$	Velocity coeff. $\psi_t(s)$	Score coeff. $(\log \bar{p}_t)'(s)$
Geodesic	–	$\dfrac{\arccos s}{(1-t)\sqrt{1-s^2}}$	–
vMF	$\propto e^{\kappa_t s}$	ODE (13)	κ_t
Power sph.	$\propto (1+s)^{\alpha_t}$	ODE (35)	$\dfrac{\alpha_t}{1+s}$
Cauchy	$\propto (1-a_t s)^{-(d-1)}$	$\dot{\beta}_t$	$\dfrac{(d-1)\,a_t}{1-a_t s}$

The score reads as

$$\nabla_{\mathbb{S}^{d-1}} \log p_t(x) = \frac{(d-1)\,a_t}{1-a_t\langle w, x\rangle}\,\mathrm{P}_x(w). \tag{39}$$

Training-time samples use $x = \mu\, w + \sqrt{1-\mu^2}\, v$ with v a uniform tangent direction and μ drawn from the radial marginal $\propto (1-\mu^2)^{(d-3)/2}(1-a_t\mu)^{-(d-1)}$, again a one-dimensional inverse-CDF draw.

In Table 4, every velocity has the form $\psi_t(s)\,\mathrm{P}_x(w)$ and every closed-form score the form $(\log \bar{p}_t)'(s)\,\mathrm{P}_x(w)$ with $s = \langle w, x\rangle$. vMF and power spherical obtain ψ_t from the quadrature (32); for the Cauchy path $\psi_t = \dot{\beta}_t$ is constant in s, with $a_t = \tanh \beta_t$. Geodesic and Cauchy have ψ_t in closed form.

C.4.3 Proofs

Proposition C.3 (Velocity of a radial path)**.** *Let $(p_t)_{t\in I}$ be radially symmetric densities with profiles $\bar{p}_t \in C^1(I \times (-1,1))$, and let f_t be as in* (8)*. Assume $f_t(s)\,\psi_t(s)\,(1-s^2) \to 0$ as $s \to -1$. Then the velocity coefficient for which the pair (p_t, v_t) with $v_t(x) = \psi_t(\langle w, x\rangle)\,\mathrm{P}_x(w)$ satisfies the continuity equation* (9) *is*

$$\psi_t(s) = -\frac{1}{f_t(s)\,(1-s^2)} \int_{-1}^{s} \partial_t f_t(y)\ \mathrm{d}y. \tag{32}$$

Proof. By Theorem 3.1, the pair (p_t, v_t) solves (9) if and only if (f_t, ψ_t) solves the flux equation (10), that is $\partial_t f_t + \partial_s\big(f_t \psi_t (1-s^2)\big) = 0$. Integrating in s from -1 to s gives

$$\int_{-1}^{s} \partial_t f_t(y)\ \mathrm{d}y + f_t(s)\psi_t(s)(1-s^2) - \lim_{y\to -1^+} f_t(y)\psi_t(y)(1-y^2) = 0.$$

The boundary term vanishes by assumption, and solving for ψ_t yields (32). □

Lemma C.4 (Power spherical velocity)**.** *For the profile* (34)*, the velocity coefficient in* (32) *is the unique solution bounded as $s \to 1$ of*

$$(1-s^2)\,\psi_t' + \big[\alpha_t(1-s) - (d-1)s\big]\,\psi_t = -\dot{\alpha}_t\Big[\log\tfrac{1+s}{2} + h(d-1+\alpha_t) - h\big(\tfrac{d-1}{2}+\alpha_t\big)\Big], \tag{35}$$

where $h = (\log\Gamma)'$ is the digamma function, with boundary value $\psi_t(1) = \frac{\dot{\alpha}_t}{d-1}\big[h(d-1+\alpha_t) - h(\frac{d-1}{2}+\alpha_t)\big]$.

Proof. For the profile (34) we have $f_t(s) = C(\alpha_t)(1-s^2)^{(d-3)/2}(1+s)^{\alpha_t}$, and the flux equation (10) reads

$$(1-s^2)^{(d-3)/2}\,\partial_t\big(C(\alpha_t)(1+s)^{\alpha_t}\big) + C(\alpha_t)\,\partial_s\big((1-s^2)^{(d-1)/2}\psi_t(s)(1+s)^{\alpha_t}\big) = 0.$$

For the first term, $\partial_t\big(C(\alpha_t)(1+s)^{\alpha_t}\big) = \dot{\alpha}_t\big[\frac{C'(\alpha_t)}{C(\alpha_t)} + \log(1+s)\big]C(\alpha_t)(1+s)^{\alpha_t}$. From (34), $\log C(\alpha) = -(d-1+\alpha)\log 2 - \frac{d-1}{2}\log\pi - \log\Gamma(\alpha+\frac{d-1}{2}) + \log\Gamma(\alpha+d-1)$, so with $h = (\log\Gamma)'$,

$$\frac{C'(\alpha)}{C(\alpha)} = -\log 2 + h(d-1+\alpha) - h\big(\tfrac{d-1}{2}+\alpha\big).$$

Expanding the derivative in the second term and dividing the whole equation by $C(\alpha_t)(1-s^2)^{(d-3)/2}(1+s)^{\alpha_t}$, using $(1-s^2)(1+s)^{-1} = 1-s$, gives

$$\dot{\alpha}_t\Big[\tfrac{C'(\alpha_t)}{C(\alpha_t)} + \log(1+s)\Big] + (1-s^2)\psi_t' + \big[\alpha_t(1-s) - (d-1)s\big]\psi_t = 0,$$

which is (35) after substituting $\frac{C'(\alpha_t)}{C(\alpha_t)}$ and collecting $\log(1+s) - \log 2 = \log\frac{1+s}{2}$. As in Lemma 3.3, boundedness at $s=1$ selects a unique solution; letting $s \to 1$ in (35) and using $\log 1 = 0$ gives $-(d-1)\psi_t(1) = -\dot{\alpha}_t[h(d-1+\alpha_t) - h(\frac{d-1}{2}+\alpha_t)]$, the stated boundary value. □

Proposition C.5 (Spherical Cauchy velocity)**.** *For the profile* (36)*, the velocity coefficient* ψ *in* (32) *is the unique solution which is bounded as* $s \to 1$ *of*

$$\partial_s\psi_t(s) + (d-1)\left(\frac{a_t}{1-a_t s} - \frac{s}{1-s^2}\right)\psi_t(s) = -\frac{(d-1)\dot{a}_t}{1-s^2}\left(\frac{s}{1-a_t s} - \frac{a_t}{1-a_t^2}\right). \tag{37}$$

This solution is given by

$$\psi_t(s) = \dot{\beta}_t = \frac{\dot{a}_t}{1-a_t^2},$$

where

$$\beta_t \coloneqq \tfrac{1}{2}\log\frac{1+a_t}{1-a_t}, \quad a_t \coloneqq \tanh\beta_t. \tag{38}$$

The score reads as

$$\nabla_{\mathbb{S}^{d-1}}\log p_t(x) = \frac{(d-1)\,a_t}{1-a_t\langle w,x\rangle}\,\mathrm{P}_x(w). \tag{39}$$

Proof. Consider the radially symmetric velocity field $v_t(x) = \dot{\beta}_t\,\mathrm{P}_x(w)$ of the form (7), with constant coefficient $\psi_t(s) = \dot{\beta}_t$ and $\beta_t = \operatorname{artanh} a_t$ as in (38). Under this velocity the cosine similarity $s_t = \langle w, X_t\rangle$ obeys $\dot{s}_t = \langle w, v_t(X_t)\rangle = \psi_t(s_t)\langle w, \mathrm{P}_{X_t}(w)\rangle = \dot{\beta}_t(1-s_t^2)$. Separating variables gives $\operatorname{artanh}(s_t) = \beta_t + \operatorname{artanh}(s_0)$, which by the addition formula for $\tanh$ is the fractional-linear (Möbius) map

$$s_t = \frac{s_0 + a_t}{1 + a_t\,s_0}, \qquad a_t = \tanh\beta_t.$$

It remains to identify the law of s_t. For $X_0 \sim \mathcal{U}(\mathbb{S}^{d-1})$ the cosine s_0 has density $f_0(s) \propto (1-s^2)^{(d-3)/2}$ by (8). Writing $\Phi(s_0) = (s_0+a_t)/(1+a_t s_0)$ for this map, its inverse is $\Phi^{-1}(s) = (s-a_t)/(1-a_t s)$ with

$$1 - \Phi^{-1}(s)^2 = \frac{(1-a_t^2)(1-s^2)}{(1-a_t s)^2}, \qquad \big(\Phi^{-1}\big)'(s) = \frac{1-a_t^2}{(1-a_t s)^2}.$$

By the change of variables formula the density of $s_t = \Phi(s_0)$ is

$$\begin{aligned} f_t(s) = f_0\big(\Phi^{-1}(s)\big)\,\big(\Phi^{-1}\big)'(s) \;&\propto\; \Big(\tfrac{(1-a_t^2)(1-s^2)}{(1-a_t s)^2}\Big)^{(d-3)/2}\frac{1-a_t^2}{(1-a_t s)^2} \\ &\propto\; (1-s^2)^{(d-3)/2}\,(1-a_t s)^{-(d-1)}. \end{aligned}$$

Dividing by $(1-s^2)^{(d-3)/2}$, the radial profile is $\bar{p}_t(s) \propto (1-a_t s)^{-(d-1)}$. As a pushforward of a probability density, f_t is normalized, and matching the profile to the normalized form identifies the constant, so the transported density is exactly (36).

As pushforward densities of the flow of $\dot{s}_t = \dot{\beta}_t(1-s_t^2)$, the family f_t satisfies $\partial_t f_t + \partial_s\big(f_t\,\dot{\beta}_t\,(1-s^2)\big) = 0$, which is the flux equation (10) with $\psi_t = \dot{\beta}_t$. For the profile (36) we have $f_t(s) \propto (1-a_t^2)^{(d-1)/2}(1-s^2)^{(d-3)/2}(1-a_t s)^{-(d-1)}$, so

$$\frac{\partial_t f_t}{f_t} = (d-1)\,\dot{a}_t\left(\frac{s}{1-a_t s} - \frac{a_t}{1-a_t^2}\right), \qquad \frac{\partial_s\big(f_t\,(1-s^2)\big)}{f_t\,(1-s^2)} = (d-1)\left(\frac{a_t}{1-a_t s} - \frac{s}{1-s^2}\right),$$

and dividing (10) by $f_t(s)(1-s^2)$ gives (37). Hence the constant $\psi_t(s) = \dot{\beta}_t = \dot{a}_t/(1-a_t^2)$ solves (37), and it is bounded as $s \to 1$ since it does not depend on s.

For uniqueness, the difference of two solutions of the linear equation (37) solves the homogeneous equation, whose solutions are proportional to

$$\exp\left(-(d-1)\int\left(\frac{a_t}{1-a_t s}-\frac{s}{1-s^2}\right)\mathrm{d}s\right) = (1-a_t s)^{d-1}\,(1-s^2)^{-(d-1)/2},$$

which is unbounded as $s \to 1$ for $d \geq 2$ and $0 \leq a_t < 1$. Hence the bounded solution is unique.

Finally, by (33) the score is $(\log \bar{p}_t)'(s)\,\mathrm{P}_x(w) = \frac{(d-1)\,a_t}{1-a_t s}\,\mathrm{P}_x(w)$, which is (39). □

D Implementation Details

Algorithm 2 gives the ODE and predictor-corrector sampler of Section 4. The experimental setup per task follows in Appendix D.1.

Algorithm 2 ODE / predictor-corrector sampling on $(\mathbb{S}^{d-1})^L$

Require: Time grid $0 = t_0 < t_1 < \cdots < t_n = 1$, corrector steps $k \geq 0$, step size $\varepsilon > 0$

Sample $x_0^l \sim \mathcal{U}(\mathbb{S}^{d-1})$ independently for $l = 1, \ldots, L$
for $i = 0, \ldots, n-1$ **do**
  $(\hat{x}^1, \ldots, \hat{x}^L) \leftarrow T_\theta(x_i^1, \ldots, x_i^L)$
  $p_j^l \leftarrow p_\theta^l(w_j \mid \hat{x}^l)$ for all l, j
  $v^l \leftarrow \dot{\kappa}_{t_i} \sum_{j=1}^{N} p_j^l\, \tilde{\psi}_{t_i}(\langle w_j, x_i^l \rangle)\, \mathrm{P}_{x_i^l}(w_j)$ for $l = 1, \ldots, L$ ▷ (16)
  $x_{i+1}^l \leftarrow \dfrac{x_i^l + (t_{i+1} - t_i)\, v^l}{\|x_i^l + (t_{i+1} - t_i)\, v^l\|}$ for $l = 1, \ldots, L$
  for $j = 1, \ldots, k$ **do** ▷ Langevin corrector
    $(\hat{x}^1, \ldots, \hat{x}^L) \leftarrow T_\theta(x_{i+1}^1, \ldots, x_{i+1}^L)$ ▷ fresh forward at corrected state
    $p_m^l \leftarrow p_\theta^l(w_m \mid \hat{x}^l)$ for all l, m
    Sample $\xi^l \sim \mathcal{N}(0, I_d)$ for $l = 1, \ldots, L$
    $g^l \leftarrow \kappa_{t_{i+1}} \sum_{m=1}^{N} p_m^l\, \mathrm{P}_{x_{i+1}^l}(w_m)$ for $l = 1, \ldots, L$ ▷ score, (17)
    $\eta^l \leftarrow \mathrm{P}_{x_{i+1}^l}\big(\varepsilon\, g^l + \sqrt{2\varepsilon}\, \xi^l\big)$ for $l = 1, \ldots, L$ ▷ project to tangent
    $x_{i+1}^l \leftarrow \dfrac{x_{i+1}^l + \eta^l}{\|x_{i+1}^l + \eta^l\|}$ for $l = 1, \ldots, L$ ▷ retract to sphere
  end for
end for
$(\hat{x}^1, \ldots, \hat{x}^L) \leftarrow T_\theta(x_n^1, \ldots, x_n^L)$
return $\hat{y}^l = \arg\max_{j=1,\ldots,N} \langle w_j, \hat{x}^l \rangle + b_j$ for $l = 1, \ldots, L$ ▷ (18)

D.1 Experimental Setup

Sudoku. We use the Sudoku-Extreme dataset from [53], a collection of 3,104,157 challenging puzzles that typically require substantially more search/backtracking (i.e., "guesses") for classical solvers. Generation is conditional: the model receives a partial puzzle and fills the missing cells, with clue positions pinned at every sampling step. We measure validity, the fraction of generated solutions satisfying all constraints, on 1,000 held-out partial puzzles at $\mathrm{NFE} = 128$. As a discrete baseline, we use masked diffusion model (MDM) [45] with mask-schedule power $p \in \{1, 0.5\}$ (Appendix D.4). MDM has no ODE: in Table 1 its first column is the predictor-only ancestral absorbing-state sampler [45] and its second column ReMDM remasking (Table 12). Because the vocabulary is small ($V = 10$), we use $\mathbb{S}^{10}$ for the sphere paths and $\mathbb{R}^{10}$ for the Euclidean paths so that the intrinsic dimension matches across paths.

LM1B. Training and evaluation follow MDLM [45] and LangFlow [11] with a DiT-style transformer backbone, 1M training steps, and GPT-2-large [43] scoring over 1024 unconditional samples. We report generation perplexity (gen. PPL) and the average per-sample entropy H. Geodesic admits no closed-form Riemannian score and uses ODE only. The sweep choices are described in Appendix E.1 and the selected configurations in Appendix E.3.

As baselines we use autoregressive transformer (AR) [11] which is sequential, LangFlow [11], Plaid [22], MDLM [45], Duo [46] and SEDD from [35]. As an additional reference point, we report the diagonal variant (configuration without flow-map distillation) of [40] at the same NFE. Note that LangFlow uses self-conditioning. Without self-conditioning they report a gen. PPL of 154.2.

TinyGSM. We use the SmolLM-135M tokenizer ($N = 49{,}152$, $L = 512$) and train for 250k steps with the models of Table 5. We evaluate on the full GSM8K test set ($1{,}319$ problems): for each problem the model generates one Python solution, which is executed in a sandbox and we report the fraction of problems whose executed solution yields the correct answer.

OpenWebText. We train for 1M steps with the models of Table 5. We report $\mathrm{NFE} = 128$ and 1024, under the ODE sampler and under the best PC configuration per entropy floor (Table 18, Figure 4). Each ODE cell gives the better of the uniform and warp-aware variants. For the self-conditioning onset τ (Section 5.2), $\tau = 0$ is the standard variant. Per-variant values are in Table 21.

D.2 Architecture and Training

Self-conditioning. Following [10], the mean embedding $\sum_k p_t^l(w_k \mid \mathbf{x})\, w_k$ of each position's predicted posterior enters the network input through a linear layer. During training it comes from a gradient-free preview pass on a fraction $p = 0.25$ of the steps and is zero otherwise. During sampling it is the posterior of the previous integration step.

All continuous methods share a DiT-style transformer backbone with adaLN conditioning. When time conditioning is off, adaLN reduces to a learnable per-block scale, shift, and gate. Otherwise it receives the schedule parameter normalized to $[0, 1]$. Embeddings are tied to the softmax classifier head as described in Section 4. We use AdamW with mixed-precision (bf16) training, a linear warmup followed by a cosine learning-rate schedule, and an exponential moving average on the network weights, which is the version used at sampling time. Masked diffusion uses the same backbone but with a discrete embedding layer of width equal to the hidden size, replacing the tied learned embeddings on the manifold. Per-task hyperparameters are listed in Table 5. The TinyGSM model size and training recipe ($4\times$A100, 250k steps) follow [15].

	Sudoku	LM1B	TinyGSM	OpenWebText
Hidden size	512	768	768	768
Blocks	8	12	12	12
Parameters (non-embed.)	28.6M	93.0M	93.0M	93.0M
Sequence length L	81	128	512	1,024
Vocabulary N	10	30,522	49,152	50,257
Training steps	1,000,000	1,000,000	250,000	1,000,000
Batch size (effective)	128	512	512	512
Learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
Warmup steps	1,000	2,500	2,500	2,500
EMA decay	0.999	0.9999	0.9999	0.9999
Dropout	0.1	0.1	0.1	0.1
Gradient clip	1.0	1.0	1.0	1.0
Precision	bf16	bf16	bf16	bf16

Table 5: Architecture and training hyperparameters per task.

D.3 Compute Resources

For the Sudoku experiments, we use an NVIDIA RTX 6000 Ada, and for LM1B we use $2\times$ NVIDIA H200 ($4\times$ for the self-conditioned runs) for approximately 2 days. TinyGSM models train on a single

node with 4× NVIDIA A100 for roughly two days each, and OpenWebText models on a single node with 4× NVIDIA H200 for roughly ten days each.

D.4 Conditional paths

vMF path. Embeddings are constrained to the unit sphere by normalization. The terminal concentration $\kappa_{\max}$ is set per task. We use $\kappa_{\max} = 15$ for Sudoku and $\kappa_{\max} = 250$ for LM1B, TinyGSM, and OpenWebText. The identity $p_1 = p_{\text{data}}$ of Section 3 holds in the limit $\kappa_t \to \infty$. In all experiments the path is truncated at $\kappa_{\max}$ and tokens are decoded from the posterior at the terminal state (Section 4). We precompute two 2D tables on a uniform grid $(\mu_i, \kappa_j) \in [-1, 1] \times [0, \kappa_{\max}]$, one for the velocity scalar $\tilde{\psi}(\mu, \kappa)$ and one for the radial CDF used in sampling. Both are queried at runtime by bilinear interpolation, so the per-step cost of a vMF velocity evaluation matches that of a Euclidean path up to the $L \times N$ posterior sum that all paths share.

Remark D.1. *The conditional velocity in Theorem 3.4 is expressed in terms of ψ_t, which is defined as the unique bounded solution of the ODE* (13)*. For practical computation, we will use the following expression. The proof of Lemma 3.3 uses the integral representation* (27)*. By* (12)*, this can be rewritten as*

$$\psi_t(s) = \dot{\kappa}_t \, \tilde{\psi}_t(s) \quad \textit{with} \quad \tilde{\psi}_t(s) \coloneqq -\frac{\int_{-1}^{s} (y - A_d(\kappa_t)) \, f_t(y) \, \mathrm{d}y}{f_t(s) \, (1 - s^2)}. \tag{40}$$

The function $\tilde{\psi}$ depends only on κ_t and not on the schedule. This representation is both schedule-independent and amenable to numerically stable evaluation, see below.

Velocity lookup. The function $\tilde{\psi}(\mu, \kappa)$ in (40) is schedule-independent and depends only on the cosine similarity $\mu = \langle w, x \rangle$ and the concentration κ. The flux representation

$$\tilde{\psi}(\mu, \kappa) = -\frac{\int_{-1}^{\mu} (y - A_d(\kappa)) \, f(y; \kappa) \, \mathrm{d}y}{f(\mu; \kappa) \, (1 - \mu^2)}$$

is evaluated by a cumulative trapezoid sum along the μ-axis. The unnormalized radial density $f(\mu; \kappa) = (1-\mu^2)^{(d-3)/2} \exp(\kappa\mu)$ varies over many orders of magnitude for large d or κ. We avoid underflow by working with $g(\mu) = \exp(\log f(\mu) - \max_\mu \log f)$, where the shift cancels in the ratio. Near $\mu \to \pm 1$ where g underflows, the analytical boundary values from Lemma 3.3 are blended smoothly with the interior solution. The Bessel ratio $A_d(\kappa) = \mathcal{I}_{d/2}(\kappa)/\mathcal{I}_{d/2-1}(\kappa)$ is evaluated by a backward continued fraction.

Training-time sampling. Each position requires a sample $x^l \sim \mathrm{vMF}(w^l, \kappa_t)$. By radial symmetry, every sample decomposes as $x = \mu \, w + \sqrt{1-\mu^2} \, v$, where v is a random unit vector in the tangent plane at w. The tangent component v is obtained by projecting a standard Gaussian onto the tangent plane and normalizing. The cosine μ is sampled by inverse CDF: we precompute the CDF of $f(\mu; \kappa)$ on the same 2D grid via cumulative integration, draw $u \sim \mathcal{U}(0, 1)$, and invert by interpolation at runtime.

Geodesic path. Embeddings are constrained to the unit sphere. The conditional path is the slerp interpolant of Remark C.1 between a uniformly sampled point and the target embedding, indexed by $t \in [0, 1]$. There is no closed-form Riemannian score for this path, so the predictor-corrector sampler is unavailable.

Linear path. Embeddings are unconstrained vectors in $\mathbb{R}^d$. The conditional path is $h_t = (1 - t)Z + tw$ with $Z \sim \mathcal{N}(0, I_d)$, indexed by $t \in [0, 1]$. The marginal score (5) diverges as $t \to 1$. At inference we cap the predictor and corrector below $t = 1 - 10^{-3}$ to avoid division by zero.

VE path. Following [17], embeddings are vectors in $\mathbb{R}^d$ rescaled to norm $\sqrt{d}$ so that the per-coordinate variance of clean embeddings matches the additive Gaussian noise. The conditional path is $h_t = w + \sigma_t Z$ with $Z \sim \mathcal{N}(0, I_d)$ and $\sigma_t \in [0, \sigma_{\max}]$. We use $\sigma_{\max} = 300$ for Sudoku, LM1B, and TinyGSM. Following [26, 17], we apply input preconditioning: the network sees $h_t / \sqrt{\sigma_t^2 + 1}$ rather than h_t directly. Combined with the $\sqrt{d}$ embedding scale, this gives the network input a per-coordinate variance independent of σ_t. The marginal score diverges as $\sigma_t \to 0$, so at inference we floor σ_t at 10^{-3}.

Masked diffusion baseline. Tokens are corrupted by an absorbing-state continuous-time Markov chain at rate $t \in [0, 1]$, following [45]. The model predicts the clean token from a partially masked sequence and is trained with a cross-entropy loss restricted to masked positions. The mask schedule is fixed (not learned) as $m(t) = 1 - (1 - t)^p$ for a power $p > 0$, where $m(t)$ is the probability that a position is masked at corruption time t. Thus p controls the curvature of the corruption schedule: $p = 1$ is linear, $p < 1$ masks fewer positions at intermediate times, and $p > 1$ masks more positions at intermediate times. On Sudoku we report $p = 1$ and $p = 0.5$. The optimal posterior is time-independent [19], so we do not condition the backbone on time for the masked baseline.

D.5 Adaptive Schedule

We give details of the adaptive schedule introduced in Remark 4.2. The loss (15) expects $t \sim \mathcal{U}(0, 1)$. In practice we sample t from a learned distribution F on $[0, 1]$ that concentrates mass where the per-sample cross-entropy varies fastest along t. The same scheme is used for every conditional path.

Parametrization. The warp $\tilde{F} : [0, 1] \to \mathbb{R}_{\geq 0}$ is piecewise linear with N bins, parametrized by two sets of logits $\ell^{(\text{in})}, \ell^{(\text{out})} \in \mathbb{R}^N$, both initialized to $-\log N$ so that the warp starts as the identity. Input bin widths are $\mathrm{softmax}(\ell^{(\text{in})})$. Output heights come from $\mathrm{softmax}(\ell^{(\text{out})})$ for the normalized CDF F and from $\exp(\ell^{(\text{out})})$ for the unnormalized $\tilde{F}$. Values $\tilde{F}(t)$ and $F(t)$ are obtained by linear interpolation within the bin containing t.

Fitting. Let

$$\ell(\mathbf{w}, \mathbf{x}_t) = -\frac{1}{L} \sum_{l=1}^{L} \log p^l_{\theta,t}(w^l \mid \hat{x}^l_t)$$

be the per-sample cross-entropy at the current t. We fit $\tilde{F}$ by minimizing

$$\mathcal{L}_{\text{warp}} = \mathbb{E}_t \Big[\big(F(t) \cdot \tilde{F}(1) - \ell(\mathbf{w}, \mathbf{x}_t) \big)^2 \Big],$$

with $F = \tilde{F}/\tilde{F}(1)$, so that F approximates the cumulative loss curve normalized to $[0, 1]$. A more flexible variant raises F to a power $\beta > 0$ before regression, which sharpens or flattens the resulting CDF. We use the linear case throughout.

Sampling. Given the fitted F, we draw t by inverse-transform sampling: $u \sim \mathcal{U}(0, 1)$ and $t = F^{-1}(u)$. Inversion is exact and amounts to evaluating the same piecewise-linear interpolator with input and output edge sequences swapped.

E Extended Experimental Results

This section documents the predictor-corrector sweep behind every PC cell of the main-text and appendix tables. Section E.1 defines the common protocol and the per-dataset grids; Sections E.2 (Sudoku), E.3 (LM1B), E.4 (OpenWebText), E.5 (TinyGSM), and E.6 (SDE sampler) report the per-dataset details and configuration tables.

E.1 Sweep Choices

The predictor-corrector sampler of Algorithm 2 has three components that we sweep uniformly across all paths admitting a closed-form score.

NFE split. Each predictor step is followed by k Langevin corrector steps at corrector interval 1, so $\text{NFE} = n(1 + k)$, and the integration window is $(0, 1]$. At fixed NFE we sweep the splits (n, k) with $n(1 + k) = \text{NFE}$, e.g. $(n, k) \in \{(64, 1), (32, 3), (16, 7)\}$ at $\text{NFE} = 128$.

Predictor spacing. By default the n predictor steps are placed uniformly in training time t. The warp-aware variant places them uniformly in the warp coordinate $F(t) \in [0, 1]$ instead, so that steps concentrate where the loss curve is steepest. The flag w marks warp-aware spacing, its absence marks uniform spacing.

Dataset	Corrector step size ε	Further axis
Sudoku	$10^{-3}, 10^{-2}, 10^{-1}, 1, 2$	—
LM1B	$10^{-4}, 3\cdot10^{-4}, 10^{-3}, 3\cdot10^{-3}, 10^{-2}$	—
TinyGSM	$10^{-3}, 10^{-2}, 10^{-1}$	budget ladder $\mathrm{NFE} \in \{1, 2, \ldots, 1024\}$
OpenWebText	$10^{-5}, 10^{-4}, 3\cdot10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$	onset $\tau \in \{0, 1/3, 1/2\}$
SDE	$\sigma \in \{0.003, \ldots, 3\}$, $q \in \{1, 2\}$	—

Table 6: Per-dataset sweep grids. Every dataset additionally sweeps the splits, both spacings and both damping states. The SDE row replaces the corrector step size by the diffusion scale σ and schedule exponent q of $\sigma_t = \sigma(1-u)^q$, and takes one Euler–Maruyama step per network evaluation.

Corrector step size. The corrector uses step size ε on a per-dataset grid, identical for all methods (Table 6, wider for the non-time-conditioned LM1B models, Appendix E.3). With damping the effective step size is $\varepsilon_{\text{eff}} = \varepsilon(1-u)^2$, where $u \in [0,1]$ is a path-specific progress variable running from 0 at the noise end to 1 at the clean end. We set $u = \kappa_t/\kappa_{\max}$ for vMF, $u = t$ for Linear, and $u = 1 - \sigma_t/\sigma_{\max}$ for VE, so that the arrow of progress is unified across paths. Damping suppresses Langevin correction near the clean end, where Euclidean scores can be singular. The flag d marks damping on, its absence marks damping off. The figures draw a denser ε sweep, the table entries select over the grids of Table 6 only. Each grid was chosen wide enough to saturate at both ends, which we verified by sweeping past them.

A configuration is specified by a tuple $((n,k),\ \varepsilon,\ \texttt{flags})$, where $\texttt{flags} \in \{\mathrm{w},\ \mathrm{d},\ \mathrm{wd},\ -\}$ encodes the two choices. Here w stands for warp-aware spacing alone, d for damping alone, wd for both, and $-$ for neither. Geodesic is sampled with the ODE only, masked diffusion with the samplers of Appendix E.2. Entries report the best configuration over this sweep: lowest gen. PPL subject to the entropy floor for the text datasets, highest accuracy for GSM8K.

Table 6 gives the per-dataset grids. Sudoku validity is measured on $1{,}000$ held-out puzzles, LM1B and OpenWebText on $1{,}024$ unconditional samples scored by GPT-2-large, and TinyGSM on the full GSM8K test set ($1{,}319$ problems) at temperature 1.0 with a 5s execution timeout. Per-dataset results are in Sections E.2–E.6.

E.2 Sudoku

With its ancestral sampler, masked diffusion trails the continuous paths at both schedule powers. Two inference-time samplers narrow this gap, neither needing retraining. Remasking [54] lets the model revise tokens it has already committed and fills the masked PC column. Committing the most confident positions first, using the sampler from [4], reaches 75.5 at $p = 0.5$, using an adaptive budget. Their grids are in Tables 12 and 13.

We report two views of the sweep here: the per-method validity grids in Tables 9, 10, and 11, and the configuration achieving each PC entry of Table 1 in Table 8. The corresponding grids for the two inference-time masked samplers are Tables 12 and 13.

The PC sweep on Sudoku uses the per-task corrector step size $\varepsilon \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 2\}$ at $\mathrm{NFE} = 128$, with $\varepsilon = 2$ included as a saturation check. Tables 9–11 report the validity at each $((n,k),\varepsilon)$, taking the best across the warp and damping flags per cell.

Method	ODE (%)	PC (%)
Non-time-conditioned		
vMF	56.4	**79.2**
Geodesic	52.2	–
Linear	54.8	73.6
VE	53.4	54.4

Table 7: Sudoku validity (↑) for the non-time-conditioned models, companion to Table 1. PC reports the best configuration over the full sweep grid (Appendix E.2).

Method	Validity	Configuration
Time-conditioned		
vMF	0.747	$(32, 3)$, $\varepsilon = 1$, d
Linear	0.678	$(16, 7)$, $\varepsilon = 1$, d
VE	0.506	$(16, 7)$, $\varepsilon = 10^{-3}$, wd
Non-time-conditioned		
vMF	0.792	$(64, 1)$, $\varepsilon = 1$, wd
Linear	0.736	$(64, 1)$, $\varepsilon = 1$, d
VE	0.544	$(16, 7)$, $\varepsilon = 10^{-3}$, d

Table 8: Configurations behind the PC entries of Tables 1 and 7.

$(n, k) \setminus \varepsilon$	10^{-3}	10^{-2}	10^{-1}	1
Time-conditioned				
$(64, 1)$	0.681	0.712	0.741	0.735
$(32, 3)$	**0.693**	0.730	0.745	0.747
$(16, 7)$	0.688	**0.732**	0.737	0.658
Non-time-conditioned				
$(64, 1)$	0.583	0.650	**0.755**	**0.792**
$(32, 3)$	0.589	0.650	0.689	0.668
$(16, 7)$	0.593	0.660	0.668	0.570

Table 9: Sudoku PC grid for vMF. ODE baselines: tc 0.658 (uniform) / 0.626 (warp-aware). Non-tc 0.564 (uniform) / 0.556 (warp-aware).

$(n, k) \setminus \varepsilon$	10^{-3}	10^{-2}	10^{-1}	1
Time-conditioned				
$(64, 1)$	0.531	0.543	0.601	0.658
$(32, 3)$	0.528	0.546	0.616	0.668
$(16, 7)$	0.529	0.547	0.633	0.678
Non-time-conditioned				
$(64, 1)$	0.542	0.560	0.610	**0.736**
$(32, 3)$	0.543	0.561	0.637	0.731
$(16, 7)$	**0.551**	**0.562**	**0.645**	0.733

Table 10: Sudoku PC grid for Linear. ODE baselines: tc 0.534 (uniform) / 0.526 (warp-aware). Non-tc 0.548 (uniform) / 0.531 (warp-aware).

$(n, k) \setminus \varepsilon$	10^{-3}	10^{-2}	10^{-1}	1
Time-conditioned				
$(64, 1)$	0.484	0.419	0.375	0.422
$(32, 3)$	0.495	0.277	0.059	0.022
$(16, 7)$	0.506	0.002	0.078	0.000
Non-time-conditioned				
$(64, 1)$	0.528	**0.499**	**0.471**	**0.490**
$(32, 3)$	0.536	0.336	0.018	0.019
$(16, 7)$	**0.544**	0.007	0.028	0.000

Table 11: Sudoku PC grid for VE. ODE baselines: tc 0.464 (uniform) / 0.480 (warp-aware). Non-tc 0.518 (uniform) / 0.534 (warp-aware).

Masked baseline. The two inference-time samplers above are swept on the same checkpoints, 1,000 samples and seed 1234 throughout. Table 12 gives the ReMDM grid and Table 13 the confidence-based grid. The two are not comparable cell by cell: ReMDM spends one forward pass per step and is fixed at $\mathrm{NFE} = 128$, whereas the confidence sampler halts once nothing is masked and its NFE is therefore adaptive and reported per cell.

σ rule	η	$p = 0.5$	$p = 1$
none (ancestral)	–	0.350	0.224
cap	0.02	0.374	0.250
cap	0.05	0.427	0.267
cap	0.10	0.468	0.294
cap	0.20	0.501	0.316
cap	0.40	0.514	0.335
cap	0.60	0.538	0.340
cap	0.80	0.558	0.372
cap	1.00	0.567	0.372
rescale	0.10	0.467	0.289
rescale	0.30	0.503	0.323
rescale	0.50	0.543	0.356
rescale	0.70	0.550	0.344
rescale	0.90	0.557	0.377
rescale	1.00	0.567	0.363

Table 12: Sudoku validity (↑) under ReMDM remasking, $\mathrm{NFE} = 128$.

$\tau \setminus S$	9	16	27	41	81
$p = 0.5$					
0.90	0.487 (6)	0.517 (8)	0.525 (11)	0.523 (12)	0.522 (19)
0.95	0.575 (6)	0.611 (8)	0.621 (12)	0.627 (14)	0.635 (23)
0.99	0.639 (7)	0.693 (10)	0.713 (15)	0.727 (16)	0.734 (34)
1.01	**0.648** (8)	**0.712** (13)	**0.739** (22)	**0.751** (32)	**0.755** (64)
$p = 1$					
0.90	0.417 (6)	0.430 (7)	0.443 (10)	0.441 (12)	0.446 (20)
0.95	0.487 (6)	0.515 (9)	0.544 (11)	0.546 (14)	0.555 (26)
0.99	0.503 (7)	0.565 (10)	0.599 (13)	0.612 (19)	0.634 (27)
1.01	0.498 (8)	0.572 (13)	0.609 (22)	0.614 (32)	0.642 (64)

Table 13: Sudoku validity (↑) under the confidence-based sampler, with the measured NFE in parentheses. τ is the confidence threshold; S is a budget, not a cost, since decoding halts once nothing is masked.

E.3 LM1B

	gen. PPL (↓) (*H* (↑))				
Method	PC$_{H\geq 4.34}$	PC$_{H\geq 4.30}$	PC$_{H\geq 4.25}$	PC$_{H\geq 4.20}$	PC$_{H\geq 4.15}$
vMF	165.4 (4.36)	112.7 (4.30)	66.0 (4.25)	52.4 (4.21)	45.1 (4.16)
Linear	138.4 (4.35)	138.4 (4.35)	138.4 (4.35)	138.4 (4.35)	138.4 (4.35)
VE	151.7 (4.35)	151.7 (4.35)	151.7 (4.35)	151.7 (4.35)	151.7 (4.35)
+ *self-conditioning* ($p = 0.25$)					
vMF	**77.5** (4.34)	**55.6** (4.30)	**38.4** (4.25)	**33.8** (4.21)	**27.0** (4.17)
Linear	88.9 (4.34)	83.1 (4.33)	83.1 (4.33)	83.1 (4.33)	83.1 (4.33)
VE	164.6 (4.36)	164.6 (4.36)	164.6 (4.36)	164.6 (4.36)	164.6 (4.36)

Table 14: **LM1B** predictor-corrector (PC) at NFE $= 128$: best gen. PPL per entropy floor, the leading column held to $H_{\text{data}} = 4.34$. Best per column in bold. Configurations in Table 16.

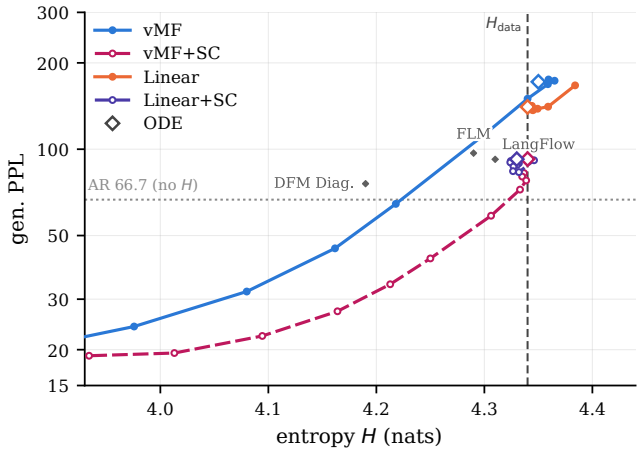


Figure 3: **LM1B** gen. PPL against entropy at NFE $= 128$, sweeping ε; every curve is predictor-corrector. Diamonds: the same models under the ODE (Table 16). References that report an entropy are placed at it, AR is drawn as a horizontal line. Data entropy H_{data} as a vertical dashed line. Configurations in Table 16.

Table 14 gives all PC entropy floors. The LM1B sweep follows the grids of Table 6. Figure 3 draws the $(64, 1)$ split with warp-aware spacing and damping. The non-time-conditioned models use the wider grid $\varepsilon \in \{10^{-5}, \ldots, 10^{-1}\}$, extended to $\varepsilon = 0.3$ for Linear. Table 15 reports their results, Table 17 the budget ablation, and Table 16 the configuration behind each predictor-corrector entry.

Document-separator alignment with reference dataloaders. Our data preparation script uses `[SEP]` (id 102) as the document separator, while the reference dataloaders of [45, 11] derive the

marker via `tokenizer.encode()[0]` and use `[CLS]` (id 101). The two tokens re-tokenize to different GPT-2 BPE sequences and would otherwise bias the perplexity comparison. We remap $102 \to 101$ at decode time before scoring.

Method	Sampler	gen. PPL (↓) (H (↑))
Non-time-conditioned		
vMF	ODE	205.5 (4.38)
	PC$_{H\geq 4.30}$	104.8 (4.31)
	PC$_{H\geq 4.15}$	77.2 (4.28)
Geodesic	ODE	215.5 (4.36)
Linear	ODE	138.8 (4.34)
	PC$_{H\geq 4.30}$	109.4 (4.31)
	PC$_{H\geq 4.25}$	88.3 (4.26)
	PC$_{H\geq 4.20}$	77.0 (4.20)
	PC$_{H\geq 4.15}$	70.7 (4.15)
VE	ODE	195.1 (4.40)
	PC$_{H\geq 4.30}$	190.3 (4.37)
	PC$_{H\geq 4.15}$	176.7 (4.29)

Table 15: LM1B comparison at $\mathrm{NFE} = 128$ for the non-time-conditioned models, companion to Table 2 and Figure 3. ODE is the Euler sampler at $n = 128$; PC$_{H\geq h}$ is the best PC configuration with entropy at or above threshold.

Method	Sampler	gen. PPL (↓) (H (↑))	Configuration
Time-conditioned			
vMF	ODE	171.8 (4.35)	warp-aware, $n = 128$
	PC, $H \geq 4.34$	165.4 (4.36)	$(32, 3)$, $\varepsilon = 3 \cdot 10^{-4}$, wd
	PC, $H \geq 4.30$	112.7 (4.30)	$(64, 1)$, $\varepsilon = 10^{-3}$, wd
	PC, $H \geq 4.25$	66.0 (4.25)	$(32, 3)$, $\varepsilon = 10^{-3}$, –
	PC, $H \geq 4.20$	52.4 (4.21)	$(32, 3)$, $\varepsilon = 10^{-3}$, w
	PC, $H \geq 4.15$	45.1 (4.16)	$(64, 1)$, $\varepsilon = 3 \cdot 10^{-3}$, wd
Geodesic	ODE	167.9 (4.34)	warp-aware, $n = 128$
Linear	ODE	140.4 (4.34)	warp-aware, $n = 128$
	PC, $H \geq 4.15$	138.4 (4.35)	$(64, 1)$, $\varepsilon = 10^{-2}$, wd
VE	ODE	149.7 (4.35)	warp-aware, $n = 128$
	PC, $H \geq 4.15$	151.7 (4.35)	$(64, 1)$, $\varepsilon = 10^{-4}$, wd
+ *self-conditioning* ($p = 0.25$)			
vMF	ODE	92.4 (4.34)	warp-aware, $n = 128$
	PC, $H \geq 4.34$	77.5 (4.34)	$(16, 7)$, $\varepsilon = 3 \cdot 10^{-4}$, wd
	PC, $H \geq 4.30$	55.6 (4.30)	$(32, 3)$, $\varepsilon = 10^{-3}$, wd
	PC, $H \geq 4.25$	38.4 (4.25)	$(32, 3)$, $\varepsilon = 10^{-3}$, w
	PC, $H \geq 4.20$	33.8 (4.21)	$(64, 1)$, $\varepsilon = 3 \cdot 10^{-3}$, wd
	PC, $H \geq 4.15$	27.0 (4.17)	$(64, 1)$, $\varepsilon = 3 \cdot 10^{-3}$, w
Geodesic	ODE	107.1 (4.33)	warp-aware, $n = 128$
Linear	ODE	92.3 (4.33)	warp-aware, $n = 128$
	PC, $H \geq 4.34$	88.9 (4.34)	$(16, 7)$, $\varepsilon = 3 \cdot 10^{-3}$, wd
	PC, $H \geq 4.15$	83.1 (4.33)	$(16, 7)$, $\varepsilon = 10^{-2}$, wd
VE	ODE	163.4 (4.36)	warp-aware, $n = 128$
	PC, $H \geq 4.15$	164.6 (4.36)	$(64, 1)$, $\varepsilon = 10^{-3}$, wd
Non-time-conditioned			
vMF	ODE	205.5 (4.38)	warp-aware, $n = 128$
	PC, $H \geq 4.30$	104.8 (4.31)	$(16, 7)$, $\varepsilon = 10^{-2}$, d
	PC, $H \geq 4.15$	77.2 (4.28)	$(16, 7)$, $\varepsilon = 10^{-2}$, wd
Geodesic	ODE	215.5 (4.36)	warp-aware, $n = 128$
Linear	ODE	138.8 (4.34)	warp-aware, $n = 128$
	PC, $H \geq 4.30$	109.4 (4.31)	$(64, 1)$, $\varepsilon = 10^{-1}$, wd
	PC, $H \geq 4.25$	88.3 (4.26)	$(64, 1)$, $\varepsilon = 0.3$, wd
	PC, $H \geq 4.20$	77.0 (4.20)	$(32, 3)$, $\varepsilon = 0.3$, wd
	PC, $H \geq 4.15$	70.7 (4.15)	$(16, 7)$, $\varepsilon = 0.3$, wd
VE	ODE	195.1 (4.40)	warp-aware, $n = 128$
	PC, $H \geq 4.30$	190.3 (4.37)	$(32, 3)$, $\varepsilon = 10^{-5}$, d
	PC, $H \geq 4.15$	176.7 (4.29)	$(16, 7)$, $\varepsilon = 10^{-3}$, d

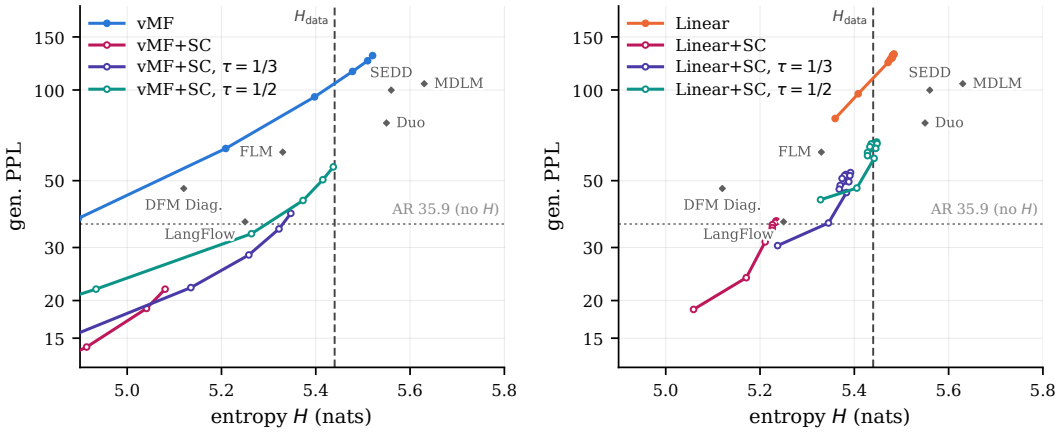
Table 16: Configurations behind every (method, sampler, entropy floor) cell of Tables 2, 14, and 15. Methods that plateau carry one PC row, at the lowest floor they reach.

	vMF		vMF +SC	
NFE	ODE	PC	ODE	PC
8	243.7 (4.40)	132.7 (4.31)	234.5 (4.40)	163.2 (4.32)
16	203.7 (4.37)	190.7 (4.36)	154.9 (4.38)	95.4 (4.30)
32	185.0 (4.36)	150.5 (4.34)	119.3 (4.36)	86.2 (4.30)
64	176.3 (4.35)	123.1 (4.32)	101.7 (4.34)	76.7 (4.32)
128	171.8 (4.35)	112.7 (4.30)	92.4 (4.34)	55.6 (4.30)

Table 17: Sampling-budget ablation on LM1B for time-conditioned vMF, without and with self-conditioning; cells report gen. PPL (H). PC is the best configuration subject to $H \geq 4.30$ (Appendix E.1).

E.4 OpenWebText

Against the references of Table 2, self-conditioned vMF under the ODE (36.0 at $H = 5.20$) is close to LangFlow (36.5 at 5.25) and ahead of DFM Diagonal (47.1 at 5.12). The delayed-onset ODE at the data entropy (73.9 at NFE $= 1024$, 70.7 at 128, Table 21) is ahead of SEDD Uniform (99.9) and MDLM (104.9), which sample above it. AR gives 35.9 without a reported entropy. At lower floors PC passes LangFlow and DFM Diagonal as well, though Linear overtakes vMF there at NFE $= 1024$.



(a) vMF (b) Linear

Figure 4: **OpenWebText** gen. PPL against entropy at NFE $= 1024$, sweeping ε at self-conditioning onsets $\tau \in \{0, 1/3, 1/2\}$; every curve is predictor-corrector. References that report an entropy are placed at it; AR reports none and is drawn as a line. Dashed: H_{data}. Both panels share axes. Configurations in Table 19.

	gen. PPL (↓) (*H* (↑))				
Method	$\text{PC}_{H\geq 5.44}$	$\text{PC}_{H\geq 5.40}$	$\text{PC}_{H\geq 5.35}$	$\text{PC}_{H\geq 5.25}$	$\text{PC}_{H\geq 5.15}$
vMF	97.7 (5.44)	94.9 (5.40)	88.5 (5.36)	61.1 (5.26)	51.1 (5.17)
Linear	117.1 (5.47)	94.6 (5.41)	80.4 (5.36)	70.1 (5.31)	70.1 (5.31)
+ *self-conditioning* ($p = 0.25$)					
vMF	**55.5** (5.44)	49.5 (5.40)	38.9 (5.35)	33.3 (5.26)	32.2 (5.22)
Linear	59.3 (5.44)	**47.3** (5.41)	**36.2** (5.35)	**32.4** (5.31)	**21.8** (5.16)

Table 18: **OpenWebText** predictor-corrector (PC) at NFE $= 1024$: best gen. PPL per entropy floor, the leading column held to $H_{\text{data}} = 5.44$. Self-conditioned cells select over the onset $\tau \in \{0, 1/3, 1/2\}$. Best per column in bold. Configurations in Table 19.

The OpenWebText sweep follows the grids of Table 6, at NFE $\in \{128, 1024\}$. Figure 4 draws the $(512, 1)$ split with warp-aware spacing. Tables 18 and 20 report the predictor-corrector results per entropy floor, Table 21 the ODE results over both spacings and all onsets, and Table 19 the configurations behind the reported entries.

Method	Cell	gen. PPL (↓) (H (↑))	Configuration	Onset
NFE $= 1024$				
vMF	PC, $H \geq 5.44$	97.7 (5.44)	$(128, 7)$, $\varepsilon = 10^{-4}$, d	0
	PC, $H \geq 5.40$	94.9 (5.40)	$(512, 1)$, $\varepsilon = 10^{-4}$, wd	0
	PC, $H \geq 5.35$	88.5 (5.36)	$(128, 7)$, $\varepsilon = 10^{-4}$, wd	0
	PC, $H \geq 5.25$	61.1 (5.26)	$(512, 1)$, $\varepsilon = 3 \cdot 10^{-4}$, d	0
	PC, $H \geq 5.15$	51.1 (5.17)	$(256, 3)$, $\varepsilon = 3 \cdot 10^{-4}$, d	0
Linear	PC, $H \geq 5.44$	117.1 (5.47)	$(128, 7)$, $\varepsilon = 10^{-3}$, wd	0
	PC, $H \geq 5.40$	94.6 (5.41)	$(512, 1)$, $\varepsilon = 10^{-1}$, d	0
	PC, $H \geq 5.35$	80.4 (5.36)	$(512, 1)$, $\varepsilon = 10^{-1}$, wd	0
	PC, $H \geq 5.25$	70.1 (5.31)	$(128, 7)$, $\varepsilon = 10^{-1}$, wd	0
vMF +SC	PC, $H \geq 5.44$	55.5 (5.44)	$(512, 1)$, $\varepsilon = 10^{-5}$, wd	1/2
	PC, $H \geq 5.40$	49.5 (5.40)	$(256, 3)$, $\varepsilon = 10^{-5}$, wd	1/2
	PC, $H \geq 5.35$	38.9 (5.35)	$(512, 1)$, $\varepsilon = 10^{-5}$, wd	1/3
	PC, $H \geq 5.25$	33.3 (5.26)	$(512, 1)$, $\varepsilon = 10^{-4}$, wd	1/2
	PC, $H \geq 5.15$	32.2 (5.22)	$(128, 7)$, $\varepsilon = 10^{-5}$, d	0
Linear +SC	PC, $H \geq 5.44$	59.3 (5.44)	$(512, 1)$, $\varepsilon = 10^{-3}$, wd	1/2
	PC, $H \geq 5.40$	47.3 (5.41)	$(512, 1)$, $\varepsilon = 10^{-2}$, wd	1/2
	PC, $H \geq 5.35$	36.2 (5.35)	$(512, 1)$, $\varepsilon = 10^{-2}$, wd	1/3
	PC, $H \geq 5.25$	32.4 (5.31)	$(128, 7)$, $\varepsilon = 10^{-2}$, wd	1/3
	PC, $H \geq 5.15$	21.8 (5.16)	$(128, 7)$, $\varepsilon = 10^{-2}$, wd	0
NFE $= 128$				
vMF	PC, $H \geq 5.44$	100.4 (5.44)	$(32, 3)$, $\varepsilon = 3 \cdot 10^{-4}$, wd	0
	PC, $H \geq 5.40$	98.3 (5.42)	$(16, 7)$, $\varepsilon = 3 \cdot 10^{-4}$, wd	0
	PC, $H \geq 5.25$	52.8 (5.25)	$(16, 7)$, $\varepsilon = 10^{-3}$, d	0
Linear	PC, $H \geq 5.40$	122.5 (5.50)	$(32, 3)$, $\varepsilon = 10^{-1}$, wd	0
vMF +SC	PC, $H \geq 5.44$	52.7 (5.44)	$(32, 3)$, $\varepsilon = 10^{-4}$, wd	1/3
	PC, $H \geq 5.35$	42.8 (5.37)	$(16, 7)$, $\varepsilon = 10^{-4}$, wd	0
	PC, $H \geq 5.25$	33.8 (5.27)	$(16, 7)$, $\varepsilon = 3 \cdot 10^{-4}$, wd	0
	PC, $H \geq 5.15$	32.6 (5.24)	$(32, 3)$, $\varepsilon = 3 \cdot 10^{-4}$, wd	0
Linear +SC	PC, $H \geq 5.44$	54.1 (5.44)	$(16, 7)$, $\varepsilon = 10^{-3}$, wd	0
	PC, $H \geq 5.40$	46.8 (5.42)	$(16, 7)$, $\varepsilon = 10^{-2}$, wd	0
	PC, $H \geq 5.35$	45.2 (5.39)	$(32, 3)$, $\varepsilon = 10^{-2}$, wd	0

Table 19: Configurations behind every PC cell of Tables 18 and 20. A row is used for every floor its entropy meets, down to the next row. Notation as in Section E.1; the onset column gives the fraction τ of the trajectory over which the self-conditioning input is suppressed. Corrector damping is on for all OpenWebText cells.

	gen. PPL (↓) (H (↑))				
Method	PC$_{H \geq 5.44}$	PC$_{H \geq 5.40}$	PC$_{H \geq 5.35}$	PC$_{H \geq 5.25}$	PC$_{H \geq 5.15}$
vMF	100.4 (5.44)	98.3 (5.42)	98.3 (5.42)	52.8 (5.25)	52.8 (5.25)
Linear	122.5 (5.50)	122.5 (5.50)	122.5 (5.50)	122.5 (5.50)	122.5 (5.50)
+ *self-conditioning* ($p = 0.25$)					
vMF	**52.7** (5.44)	52.7 (5.44)	**42.8** (5.37)	**33.8** (5.27)	**32.6** (5.24)
Linear	54.1 (5.44)	**46.8** (5.42)	45.2 (5.39)	45.2 (5.39)	45.2 (5.39)

Table 20: **OpenWebText** predictor-corrector (PC) at NFE $= 128$; otherwise as Table 18.

Method	NFE	Onset	gen. PPL (↓) (H (↑)) ODE (uniform)	ODE (warp-aware)
Ours				
vMF	1024	0	140.4 (5.518)	**139.3** (5.518)
	128	0	156.8 (5.549)	**146.9** (5.537)
Geodesic	1024	0	149.9 (5.537)	**139.0** (5.514)
	128	0	249.8 (5.684)	**144.4** (5.536)
Linear	1024	0	123.5 (5.455)	**123.0** (5.456)
	128	0	134.2 (5.473)	**128.6** (5.480)
+ *self-conditioning* ($p = 0.25$)				
vMF	1024	0	41.6 (5.244)	**36.0** (5.198)
		$1/3$	116.1 (5.533)	55.3 (5.381)
		$1/2$	146.1 (5.559)	**73.9** (5.456)
	128	0	77.0 (5.436)	**53.6** (5.350)
		$1/3$	138.4 (5.581)	**70.7** (5.453)
		$1/2$	166.1 (5.600)	88.4 (5.506)
Geodesic	1024	0	64.9 (5.409)	**36.3** (5.233)
		$1/3$	159.0 (5.563)	52.7 (5.386)
		$1/2$	159.0 (5.563)	**70.3** (5.449)
	128	0	258.8 (5.726)	**55.0** (5.378)
		$1/3$	275.0 (5.714)	**67.5** (5.447)
		$1/2$	275.0 (5.714)	83.1 (5.491)
Linear	1024	0	38.1 (5.229)	**36.3** (5.224)
		$1/3$	38.6 (5.236)	52.5 (5.372)
		$1/2$	41.6 (5.278)	**66.2** (5.429)
	128	0	60.9 (5.361)	**50.1** (5.348)
		$1/3$	61.3 (5.355)	63.0 (5.433)
		$1/2$	64.3 (5.384)	**76.1** (5.475)

Table 21: Per-variant OpenWebText ODE results: uniform and warp-aware spacing at $n \in \{128, 1024\}$, across the self-conditioning onset τ. Bold marks the better spacing at onset 0 and, among the delayed-onset cells, the one reported at the data entropy: the lowest gen. PPL with $H \geq H_{\text{data}} = 5.44$, or the closest to it for Linear at $\text{NFE} = 1024$.

E.5 TinyGSM

The TinyGSM sweep follows the grids of Table 6. Table 23 reports the accuracy at every budget, and Table 22 lists the configuration behind each predictor-corrector entry.

Method	Cell	Acc. (%)	Configuration
*Table 3 (*NFE $= 1024$*)*			
vMF	PC	33.0	$(512, 1)$, $\varepsilon = 10^{-2}$, w
Linear	PC	12.7	$(512, 1)$, $\varepsilon = 10^{-1}$, wd
VE	PC	1.9	$(512, 1)$, $\varepsilon = 10^{-3}$, d
+ *self-conditioning (*$p = 0.25$*)*			
vMF	PC	43.5	$(512, 1)$, $\varepsilon = 10^{-2}$, w
Linear	PC	28.5	$(512, 1)$, $\varepsilon = 10^{-1}$, d
VE	PC	4.1	$(512, 1)$, $\varepsilon = 10^{-1}$, d
Table 23 ladder; the NFE $= 1024$ *rows repeat the vMF cells above*			
vMF	PC, NFE $= 2$	0.0	$(1, 1)$, $\varepsilon = 10^{-1}$, d
	PC, NFE $= 4$	0.7	$(2, 1)$, $\varepsilon = 10^{-3}$, w
	PC, NFE $= 8$	6.2	$(4, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 16$	15.5	$(8, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 32$	21.8	$(16, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 64$	26.9	$(32, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 128$	27.4	$(64, 1)$, $\varepsilon = 10^{-2}$, wd
	PC, NFE $= 256$	31.8	$(128, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 512$	32.4	$(256, 1)$, $\varepsilon = 10^{-2}$, w
vMF +SC	PC, NFE $= 2$	0.0	$(1, 1)$, $\varepsilon = 10^{-3}$, w
	PC, NFE $= 4$	1.1	$(2, 1)$, $\varepsilon = 10^{-3}$, w
	PC, NFE $= 8$	9.7	$(4, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 16$	21.9	$(8, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 32$	30.9	$(16, 1)$, $\varepsilon = 10^{-2}$, wd
	PC, NFE $= 64$	35.6	$(32, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 128$	39.7	$(64, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 256$	39.4	$(128, 1)$, $\varepsilon = 10^{-2}$, w
	PC, NFE $= 512$	41.7	$(256, 1)$, $\varepsilon = 10^{-2}$, w

Table 22: Configurations behind the PC cells of Tables 3 and 23. Notation as in Section E.1.

	vMF		vMF +SC	
NFE	ODE	PC	ODE	PC
1	0.0	–	0.0	–
2	0.4	0.0	0.2	0.0
4	2.0	0.7	2.7	1.1
8	3.9	6.2	7.1	9.7
16	5.8	15.5	11.6	21.9
32	6.2	21.8	17.4	30.9
64	6.1	26.9	19.9	35.6
128	6.4	27.4	22.5	39.7
256	6.7	31.8	23.0	39.4
512	6.4	32.4	23.7	41.7
1024	6.4	33.0	23.7	43.5

Table 23: Sampling-budget ablation on GSM8K: test accuracy (↑, %) for time-conditioned vMF, without and with self-conditioning. PC is the best configuration per budget (Appendix E.1).

Sampling temperature. Table 24 repeats the entries of Table 3 at temperature $T = 0.1$, which divides the logits by T before every softmax of the sampler, as in [15].

Method	ODE		PC	
	$T = 1$	$T = 0.1$	$T = 1$	$T = 0.1$
vMF	6.4	10.5	33.0	30.9
Geodesic	5.0	8.5	–	–
Linear	7.7	13.4	12.7	15.8
VE	2.1	4.7	1.9	4.6
+ self-conditioning ($p = 0.25$)				
vMF	23.7	29.9	43.5	40.0
Geodesic	14.9	19.1	–	–
Linear	18.3	23.5	28.5	31.8
VE	3.5	8.2	4.1	8.0
Reference [15], $T = 1$ / $T = 0.1$				
MDLM	18.0 / ≈ 33			
Duo	17.2 / 36.0			
$\mathbb{S}$-FLM (top-k=1)	18.0 / ≈ 18			

Table 24: **GSM8K** test accuracy (↑, %) at sampling temperature $T = 1$ (Table 3) and $T = 0.1$, NFE $= 1024$. Reference values from [15], where $\approx$ marks values given there only approximately.

E.6 SDE Sampler

Protocol. We evaluate Algorithm 1 on the same checkpoints and under the protocol of Section 5: gen. PPL and entropy H on LM1B and OpenWebText, GSM8K accuracy after training on TinyGSM. The number of Euler–Maruyama steps equals the budget, $S = \mathrm{NFE}$; there is no predictor-corrector split. Steps are placed uniformly in the warp coordinate, on the grid clamped below $u = 1$ (up to $1 - 10^{-3}$), and tokens are decoded from the posterior at the final state, so no step to $\kappa_{\max}$ is taken.

Sweep. We sweep the diffusion scale $\sigma \in \{0.003, 0.01, 0.03, 0.05, 0.1, 0.2, 0.3, 0.5, 1, 3\}$ and the schedule exponent $q \in \{1, 2\}$ of $\sigma_t = \sigma(1 - u)^q$, with the path progress variable u of Section E.1 ($u = \kappa_t/\kappa_{\max}$ for vMF, $u = t$ for Linear, $u = 1 - \sigma_t/\sigma_{\max}$ for VE). Entropy floors, sample counts and scoring follow the per-dataset protocol of Appendix E.1.

Reporting. All SDE runs use a single seed. Table 25 through Table 27 report the results (per entropy floor for LM1B and OpenWebText), and Table 28 lists the winning (σ, q) behind each entry. Geodesic admits no closed-form score and is omitted.

Method	gen. PPL (↓) (H (↑))				
	$\mathrm{SDE}_{H \geq 4.34}$	$\mathrm{SDE}_{H \geq 4.30}$	$\mathrm{SDE}_{H \geq 4.25}$	$\mathrm{SDE}_{H \geq 4.20}$	$\mathrm{SDE}_{H \geq 4.15}$
vMF	136.1 (4.35)	136.1 (4.35)	77.1 (4.28)	77.1 (4.28)	77.1 (4.28)
Linear	137.3 (4.34)	137.3 (4.34)	137.3 (4.34)	137.3 (4.34)	137.3 (4.34)
VE	145.3 (4.35)	145.3 (4.35)	145.3 (4.35)	145.3 (4.35)	145.3 (4.35)
+ self-conditioning ($p = 0.25$)					
vMF	**70.0** (4.34)	**49.7** (4.30)	**49.7** (4.30)	**37.2** (4.24)	**26.9** (4.18)
Linear	118.2 (4.35)	91.8 (4.32)	91.8 (4.32)	91.8 (4.32)	91.8 (4.32)
VE	160.2 (4.36)	160.2 (4.36)	160.2 (4.36)	160.2 (4.36)	160.2 (4.36)

Table 25: LM1B SDE results at NFE $= 128$: best gen. PPL (H) per entropy floor; methods that plateau repeat the same value at every floor. Best per column in bold.

Across all three tasks the SDE is competitive with the predictor-corrector results at matched budgets, though not at every entropy floor, with σ acting as the same quality–entropy trade-off knob as the corrector step size. On GSM8K for example it reaches 43.1% at NFE $= 512$, matching the predictor-corrector headline (43.5% at NFE $= 1024$, Table 3). The path ordering is unchanged: vMF leads at nearly every floor and budget, Linear requires large σ and trails on GSM8K, and VE remains flat.

Method	NFE	gen. PPL (↓) (H (↑))				
		SDE$_{H\geq 5.44}$	SDE$_{H\geq 5.40}$	SDE$_{H\geq 5.35}$	SDE$_{H\geq 5.25}$	SDE$_{H\geq 5.15}$
vMF	1024	**112.7** (5.47)	**93.6** (5.43)	88.1 (5.36)	88.1 (5.36)	49.4 (5.19)
	128	118.9 (5.51)	118.9 (5.51)	68.3 (5.38)	68.3 (5.38)	52.8 (5.20)
Linear	1024	119.4 (5.44)	105.0 (5.42)	105.0 (5.42)	105.0 (5.42)	105.0 (5.42)
	128	128.7 (5.48)	128.7 (5.48)	128.7 (5.48)	128.7 (5.48)	128.7 (5.48)
+ self-conditioning ($p = 0.25$)						
vMF	1024	–	–	–	–	33.0 (5.17)
	128	–	–	**43.5** (5.35)	**37.4** (5.30)	27.1 (5.21)
Linear	1024	–	–	–	–	**26.3** (5.17)
	128	149.3 (5.54)	149.3 (5.54)	50.9 (5.35)	50.9 (5.35)	50.9 (5.35)

Table 26: OpenWebText SDE results at NFE $\in \{1024, 128\}$: best gen. PPL (H) per entropy floor; methods that plateau repeat the same value at every floor. Empty cells (–) mark floors that no configuration in the sweep stays above. Best per column in bold.

Method	NFE $= 1024$	NFE $= 512$
vMF	33.4	32.6
Linear	9.6	9.2
VE	2.8	2.5
+ self-conditioning ($p = 0.25$)		
vMF	**42.4**	**43.1**
Linear	22.6	22.7
VE	3.9	4.5

Table 27: GSM8K test accuracy (↑, %) after training on TinyGSM, under the SDE sampler at NFE $\in \{1024, 512\}$. Best per column in bold.

Dataset	Method	Floor / budget	Value (H)	σ	q
LM1B (128)	vMF (TC)	$H \geq 4.34$ / 4.30	136.1 (4.35)	0.03	1
	vMF (TC)	$H \geq 4.25$	77.1 (4.28)	0.05	1
	vMF (SC)	$H \geq 4.34$	70.0 (4.34)	0.03	1
	vMF (SC)	$H \geq 4.30$ / 4.25	49.7 (4.30)	0.05	1
	vMF (SC)	$H \geq 4.20$	37.2 (4.24)	0.1	2
	vMF (SC)	$H \geq 4.15$	26.9 (4.18)	0.1	1
	Linear (TC)	all floors	137.3 (4.34)	0.2	1
	Linear (SC)	$H \geq 4.34$	118.2 (4.35)	3	1
	Linear (SC)	$H \geq 4.30$ and below	91.8 (4.32)	0.1	1
	VE (TC / SC)	all floors	145.3 (4.35) / 160.2 (4.36)	0.1 / 0.003	1
OWT	vMF (TC), 1024	$H \geq 5.44$	112.7 (5.47)	0.03	2
	vMF (TC), 1024	$H \geq 5.40$ / 5.35 / 5.15	93.6 (5.43) / 88.1 (5.36) / 49.4 (5.19)	0.03 / 0.05 / 0.05	1 / 2 / 1
	vMF (TC), 128	$H \geq 5.40$ / 5.35 / 5.15	118.9 (5.51) / 68.3 (5.38) / 52.8 (5.20)	0.03 / 0.03 / 0.05	2 / 1 / 2
	vMF (SC), 1024	$H \geq 5.15$	33.0 (5.17)	0.003	1
	vMF (SC), 128	$H \geq 5.35$ / 5.25 / 5.15	43.5 (5.35) / 37.4 (5.30) / 27.1 (5.21)	0.01 / 0.03 / 0.03	1 / 2 / 1
	Linear (TC), 1024	$H \geq 5.44$	119.4 (5.44)	1	1
	Linear (TC), 1024	$H \geq 5.40$ and below	105.0 (5.42)	3	1
	Linear (TC), 128	all floors	128.7 (5.48)	0.1	1
	Linear (SC), 1024	$H \geq 5.15$	26.3 (5.17)	3	1
	Linear (SC), 128	$H \geq 5.44$ / 5.40	149.3 (5.54)	3	1
	Linear (SC), 128	$H \geq 5.35$	50.9 (5.35)	0.1	1
GSM8K	vMF (TC)	1024 / 512	33.4% / 32.6%	0.5 / 0.3	2 / 1
	vMF (SC)	1024 / 512	42.4% / 43.1%	0.5 / 0.2	2 / 1
	Linear (TC)	1024 / 512	9.6% / 9.2%	3	1
	Linear (SC)	1024 / 512	22.6% / 22.7%	3	1
	VE (TC)	1024 / 512	2.8% / 2.5%	0.01 / 3	1
	VE (SC)	1024 / 512	3.9% / 4.5%	0.03 / 3	1

Table 28: Winning (σ, q) configurations behind each SDE cell of Tables 25–27. Values that plateau across floors are listed once. Linear selects large σ throughout GSM8K and on self-conditioned OpenWebText, except at $H \geq 5.35$ for NFE $= 128$.

F Additional Evaluations

F.1 Inference-Seed Variability

Path	Setting	Seed 42	Mean ± s.d.
LM1B, ODE, NFE $= 128$ *(Table 16)*			
vMF	ODE	171.8	171.4 ± 2.0
Geodesic	ODE	167.9	168.9 ± 1.5
Linear	ODE	140.4	139.4 ± 1.0
VE	ODE	149.7	147.7 ± 1.9
vMF+SC	ODE	92.4	93.1 ± 0.8
Geodesic+SC	ODE	107.1	106.9 ± 1.5
Linear+SC	ODE	92.3	92.1 ± 1.3
VE+SC	ODE	163.4	164.4 ± 1.4
LM1B, PC, NFE $= 128$ *(Table 14)*			
vMF	PC, $H \geq 4.30$	112.7	112.4 ± 2.2
vMF	PC, $H \geq 4.25$	66.0	66.2 ± 0.5
vMF	PC, $H \geq 4.20$	52.4	51.6 ± 0.5
vMF	PC, $H \geq 4.15$	45.1	44.5 ± 0.4
Linear	PC, $H \geq 4.30$–4.15	138.4	139.3 ± 1.8
VE	PC, $H \geq 4.30$–4.15	151.7	151.0 ± 2.5
vMF+SC	PC, $H \geq 4.30$	55.6	56.0 ± 0.7
vMF+SC	PC, $H \geq 4.25$	38.4	38.6 ± 0.2
vMF+SC	PC, $H \geq 4.20$	33.8	33.9 ± 0.1
vMF+SC	PC, $H \geq 4.15$	27.0	26.7 ± 0.2
Linear+SC	PC, $H \geq 4.30$–4.15	83.1	83.5 ± 0.3
VE+SC	PC, $H \geq 4.30$–4.15	164.6	167.2 ± 1.9
OpenWebText, ODE, NFE $= 1024$ *(Table 21)*			
vMF	ODE	139.3	137.9 ± 1.2
Geodesic	ODE	139.0	140.3 ± 2.7
Linear	ODE	123.0	124.7 ± 1.5
vMF+SC	ODE	36.0	36.0 ± 0.5
Geodesic+SC	ODE	36.3	35.7 ± 0.7
Linear+SC	ODE	36.3	37.0 ± 0.5
vMF+SC	ODE, delayed onset	73.9	75.3 ± 0.8
Geodesic+SC	ODE, delayed onset	70.3	71.5 ± 1.0
Linear+SC	ODE, delayed onset	66.2	66.1 ± 0.7
OpenWebText, PC, NFE $= 1024$ *(Table 18)*			
vMF	PC, $H \geq 5.15$	51.1	49.9 ± 0.9
Linear	PC, $H \geq 5.15$	70.1	68.0 ± 1.6
vMF+SC	PC, $H \geq 5.15$	32.2	31.7 ± 0.4
Linear+SC	PC, $H \geq 5.15$	21.8	22.0 ± 0.3
TinyGSM, NFE $= 1024$*, accuracy in % (Table 3)*			
vMF	ODE	6.4	6.6 ± 0.3
Geodesic	ODE	5.0	5.3 ± 0.3
Linear	ODE	7.7	7.6 ± 0.5
VE	ODE	2.1	2.5 ± 0.4
vMF	PC	33.0	32.9 ± 0.4
Linear	PC	12.7	12.1 ± 0.6
VE	PC	1.9	2.4 ± 0.5
vMF+SC	ODE	23.7	23.6 ± 1.0
Geodesic+SC	ODE	14.9	15.1 ± 0.2
Linear+SC	ODE	18.3	18.7 ± 0.8
VE+SC	ODE	3.5	4.1 ± 0.4
vMF+SC	PC	43.5	42.2 ± 1.0
Linear+SC	PC	28.5	28.2 ± 1.1
VE+SC	PC	4.1	4.0 ± 0.3

Table 29: Inference-seed variability: mean and standard deviation of gen. PPL (TinyGSM: accuracy) over five generation seeds.

The configurations behind the main tables were re-run with five generation seeds (42–46), holding everything else fixed. Seed 42 is the draw printed in the result tables. Table 29 reports the mean

and standard deviation over the five seeds for 59 table entries. A configuration that plateaus across entropy floors is listed once.

F.2 Across-Sample Diversity

Entropy H measures diversity within a sample. To check that the corrector does not buy perplexity with a loss of diversity across samples, we compute distinct-2 [30], the fraction of unique word bigrams over all samples, and self-BLEU-4 [56], each sample scored against all others (leave-one-out, NLTK method-1 smoothing), on $1{,}024$ samples per cell with whitespace tokenization of the decoded texts. Real data is a held-out set of the same size. Table 30 reports selected cells at the budgets of the main tables.

Dataset	Cell	H	distinct-2 (↑)	self-BLEU (↓)
LM1B	real data	4.34	0.668	0.161
	vMF, ODE	4.35	0.667	0.171
	Linear, ODE	4.34	0.658	0.180
	vMF, PC, $\varepsilon = 10^{-5}$	4.35	0.667	0.173
	vMF, PC, $\varepsilon = 3 \cdot 10^{-2}$	4.18	0.460	0.392
OpenWebText	real data	5.44	0.586	0.176
	vMF, ODE	5.52	0.614	0.161
	Linear, ODE	5.45	0.610	0.167
	vMF, PC, $\varepsilon = 10^{-5}$	5.51	0.609	0.167
	vMF +SC, ODE	5.20	0.549	0.184
	vMF +SC, PC	5.19	0.548	0.185

Table 30: Across-sample diversity at $\mathrm{NFE} = 128$ (LM1B) and $\mathrm{NFE} = 1024$ (OpenWebText), $1{,}024$ samples per cell. ODE and PC are compared at equal NFE. Each cell is its own $1{,}024$-sample draw, so entropies can differ slightly from the main tables.

At matched entropy the corrector leaves across-sample diversity unchanged. Where it lowers entropy, it lowers across-sample diversity visibly and in proportion, so the gains at matched entropy do not come from a hidden loss of diversity.

F.3 Distributional Distance (SOL)

Generative perplexity and entropy score single samples. To compare whole sample distributions with the data we additionally report SOL [28], a double-sliced Wasserstein distance between the distributions of contextual token embeddings of a generated and a held-out reference corpus. It needs only samples, so it applies to every model in our tables, and lower is closer. We follow the protocol of [28]: $5{,}000$ samples per corpus, Dream-7B as the encoder with $8{,}192$ token directions and $8{,}192$ Gaussian-process slices, 64 quantile bins and length scale 0.1. On OpenWebText the reference is their held-out GPT-2-packed set (family G) at a $1{,}024$-token horizon. For our models we generate $5{,}000$ fresh samples at the configuration of the printed entry and report a single estimate (direction seed 0, GP seed 0). PC rows in Tables 31 and 32 are labeled by the entropy floor at which their configuration was selected (Tables 18 and 14); the regenerated samples can fall up to 0.02 below it. Reference values are taken from [28], where they are the mean over nine seed pairs. For the rows shown, the standard deviation over these pairs is at most 0.19. Table 31 lists the OpenWebText scores.

On LM1B (Table 32) the reference is $10{,}000$ held-out validation sequences of 128 BERT tokens (decoded with special tokens kept), with the same encoder and settings and the reference subset drawn as for OpenWebText. Our samples are decoded the same way. Reference rows are scored the same way, with H and gen. PPL measured on the scored samples by the official evaluation code.

Model	Setting	H	gen. PPL	SOL (↓)
Real OWT (held out)	–	5.449	14.9	2.07
Reference, values from [28]				
AR [45]	temperature 1.0	5.582	34.9	22.21
MDLM [45]	exact sampling, $T = 1000$	5.445	59.4	36.39
Duo [46]	ancestral, $T = 1024$	5.537	76.9	32.64
LangFlow [11]	NFE $= 128$	5.425	59.0	32.49
LangFlow [11]	NFE $= 1024$	5.257	36.9	30.54
FLM [29]	Euler, 1024 steps	5.345	63.1	48.39
Ours				
vMF	ODE, NFE $= 1024$	5.512	137.6	45.82
Geodesic	ODE, NFE $= 1024$	5.519	140.9	45.71
Linear	ODE, NFE $= 1024$	5.464	125.9	44.19
vMF +SC	ODE, NFE $= 1024$	5.194	35.9	32.15
Geodesic +SC	ODE, NFE $= 1024$	5.227	35.7	32.22
Linear +SC	ODE, NFE $= 1024$	5.227	36.7	31.54
vMF +SC	PC, NFE $= 1024$	5.203	31.8	30.09
Linear +SC	PC, NFE $= 1024$	5.157	21.9	28.77
vMF +SC	SDE, NFE $= 128$	5.210	27.2	28.94
Linear +SC	SDE, NFE $= 128$	5.360	51.3	31.94
Ours, at the data entropy ($H \approx 5.44$, NFE $= 1024$, onset $\tau = 1/2$)				
vMF +SC	ODE, delayed onset	5.462	75.6	33.69
Geodesic +SC	ODE, delayed onset	5.456	72.0	33.70
Linear +SC	ODE, delayed onset	5.435	66.3	32.90
vMF +SC	PC, $H \geq 5.44$	5.429	56.0	30.85
Linear +SC	PC, $H \geq 5.44$	5.424	59.1	31.85

Table 31: **OpenWebText** SOL (Dream-7B encoder, reference G, 5,000 samples, 1,024 tokens). H and gen. PPL of our rows are measured on the same 5,000 samples, those of the references are as reported in [28].

Model	Setting	H	gen. PPL	SOL (↓)
Real LM1B (held out)	–	–	–	1.44
Reference				
LangFlow [11]	NFE $= 128$	4.316	93.1	26.00
MDLM [45]	NFE $= 128$	4.325	114.0	29.83
FLM [29]	NFE $= 128$	4.339	113.0	32.97
FMLM [29]	NFE $= 5$	4.199	93.2	50.17
FMLM [29]	NFE $= 2$	4.188	105.9	55.40
FMLM [29]	NFE $= 1$	4.147	115.8	62.05
Ours, NFE $= 128$				
vMF	ODE	4.350	169.9	44.93
Geodesic	ODE	4.348	168.1	45.74
Linear	ODE	4.344	139.0	38.12
VE	ODE	4.352	146.8	39.39
vMF +SC	ODE	4.331	94.1	29.68
Geodesic +SC	ODE	4.338	108.0	32.90
Linear +SC	ODE	4.325	91.7	28.52
VE +SC	ODE	4.358	164.3	43.24
vMF +SC	PC, $H \geq 4.30$	4.297	56.4	21.22
Linear +SC	PC, $H \geq 4.30$	4.330	83.8	25.92
vMF +SC	PC, $H \geq 4.34$	4.339	78.0	25.30
Linear +SC	PC, $H \geq 4.34$	4.334	88.4	27.30
vMF +SC	PC, $H \geq 4.15$	4.167	26.9	15.46

Table 32: **LM1B** SOL (Dream-7B encoder, 5,000 samples, 128 tokens). H and gen. PPL are measured on the same 5,000 samples. The real-data row is the mean over nine seed pairs of two disjoint held-out halves.

G Qualitative Samples

G.1 LM1B

Decoded LM1B samples (BERT tokenizer, $L = 128$) at $\mathrm{NFE} = 128$, from the runs behind Tables 2 and 14 and Figure 3.

vMF + self-conditioning

ODE, warp-aware, $n = 128$
gen. PPL = 92.4, $H = 4.34$.

[CLS] [CLS] the office of the consumer agreed that workers must still be in danger of falling incomes. [CLS] opec, which currently has a 25 cent stake in opel, has continued to face up to a " horrisome " auction yesterday that has been dominated by optimism about production costs in recent years with a general bid commitment. [CLS] boutvinges - reid was scheduled to attend the opening hearing friday and fetch a hearing later on wednesday morning. [CLS] but her father, overlooking, told reporters in australia she didn ' t want to find out anything about her medal numbers. [CLS] (ap) - alynn communications said thursday that it [CLS]

[CLS] - time music star, who died monday in a train crash from a orange county, calif., hospital after her real sister was injured. [CLS] he said more than a quarter of asset operations held by shareholders in hbos were " fundamentally uncontrolled. " [CLS] other than just new jersey ' s reflay, mancini is flying to redkna as a potential derby to impress. [CLS] dozens of people are reportedly still dead every day. [CLS] my skills with maximum curiosity and aptitude sabotage affects the utter order were also with greater resonance. [CLS] " the six - hourors " hosts a series of family - dance [CLS]

[CLS] kids have been putting through something. " [CLS] west yorkshire police said a 17 - year - old man was being treated for very serious injuries after a woman was snatched in st george ' s address in august last year. [CLS] a second trial, which began in may 2007, involves unlawfully fleeting lehman brothers from spain last year, but say mr ijyu is unlikely to stay in line with citigroup ' s mainholders. [CLS] this article was made by glitj global markets (bst on friday 27 july 2010. bbc home city part al. humek, kan. [CLS] i ' m just a hawk. [CLS]

vMF + self-conditioning

PC, $(32, 3)$**,** $\varepsilon = 10^{-3}$**,** wd
gen. PPL = 55.6, $H = 4.30$.

[CLS] the most basic surroundings. [CLS] nasoe ' s brilliant strike from tevez gave arsenal a crucial lead but didier drogba, after watching a frustrating equaliser only nine minutes later, sent his side back into the net. [CLS] the franchise retail stores (www. thewinepark. com / fairfield hd. com) is a leading consumer electronics retailer of specialty apparel, accessories, apparel, accessories, and accessories. walmart. com offers a wide array of consumer products, including specialtywear, shoes, necklaces, game rings. [CLS] in afternoon trading on thursday, the ftse eurofirst 300 added 7. 58 [CLS]

[CLS] authorities say suspect umar farouk abdulmutallab and his deputy, umar farouk jabayiri, were terror suspects who had been blamed for amsterdam ' s september 11 terror attacks on yemen. [CLS] mr thomas, 51, from harrow, east london, was awarded the medal. [CLS] wachovia, which has been hit hard by the subprime mortgage crisis, is one of the few banks to step up deposits that make it easier for the borrowers to borrow, a move that could help them buy into their own homes and refinance them in exchange for at least two years. [CLS] sarkozy, who has [CLS]

[CLS] tribal areas near the border with afghanistan and along pakistan ’ s border border. [CLS] " the country will travel to copenhagen this month to travel to london to set up a new strategy to tackle global warming and fight climate change, " he said. [CLS] the obamas ’ spending plan will provide up to $ 114 billion to help them avert large failures in the country ’ s financial markets. [CLS] u. s. officials have questioned amiri ’ s suggestion that it would not only stop refining oil and gasoline in iran, and help strengthen his government ’ s courage and line the international community for imposing sanctions on tehran. [CLS] they sometimes discussed [CLS]

Geodesic + self-conditioning

ODE, warp-aware, $n = 128$
gen. PPL = 107.1, $H = 4.33$.

[CLS], egypt has not yet endorsed the arab civil board. [CLS] an authority official said he concluded the bomb was not being fully treated. [CLS] guilgivichs - - whom she later joined from her college training - - fled to sweden after her refusal to take in the equipment. [CLS] the same thing in which everyone wants to stick has spoken. [CLS] church schools in the city are also on the hook to rebunk the fragile social divide between gay and lesbian women and the catholic community. [CLS] ochoa, who started pulling back on a stretcher, got edging on the next hole. [CLS] (cbs) a bad air [CLS]

[CLS] that doesn ’ t mean. [CLS] but its identity is unusually low. [CLS] pelosi, reid and feinstein broke talking quickly wednesday to republicans and skeptical that they would not create a $ 1 billion spending bill enacted at the end of july. [CLS] these rules are bills that would really require. [CLS] the move, inscores the response of lloyds, has triggered concerns almost taken by the credit card agency to an overwhelm the company ’ s control. [CLS] after their jabing against the bag, " the likelihood of a single child dropped dramatically, " he explained, only the 49 - year - old smorist [CLS]

[CLS] on the number of posts who could have been working at children ’ s allowance? [CLS] hutesens chose newly groomed secretary - president blika el saytim to have a victory over government - dominated election friday ’ s gathering to try to reduce fears that the earthquake could be caused by apparent problems in kenya. [CLS] also years later, those were identified as unfair encountering "orro racer, " whose name is friends of the u. s. and who emigrated to the united states. [CLS] which has experience fuel that by simply not understanding what the chinese principles they would be very careful. [CLS] 12 (bna) - iraq [CLS]

Linear + self-conditioning

PC, $(16, 7)$**,** $\varepsilon = 10^{-2}$**,** wd
gen. PPL = 83.1, $H = 4.33$.

[CLS] are usually not to be discrued upon garnett ’ s success. [CLS] a document which leads to the rule of speech, such as law and literature, is more more important than any practice or constitution. [CLS] that contest was 6 - 0, against berkeley state. [CLS] two packages each sells for $ 3. [CLS] the burmese state media appointed suu kyi ’ s chief military adviser as the former military director in late 2005 when he was released on trial. [CLS] in fact, the chocolate isn ’ t missing by snogakers. [CLS] the discovery was based on other models of advanced data to create safe, fragmented communities [CLS]

[CLS] ft, the company said that according to the daily telegraph, sir ronnie prepared for his review of financial rules, " recounting " the recent report on the corporation ’ s annual council tax package for a maximum of five years which presents a " grave and grave concern for the economy. [CLS] she was confirmed dead in serious condition at the hospital at fort bellalco on oct. [CLS] " they are supported by the external state which can involve people at other disciplines in london. [CLS] new tools will allow employers to make final monthly annuity checks monthly but also, if they don ’ t save their husbands or pay off their loan, [CLS]

[CLS] to the all - star roster. [CLS] he put his face in the lungs and had three broken fingers at his leg. [CLS] " this is not about a prank, " foreign secretary david miliband said on saturday. [CLS] assets bearing price on non - performing interests increased by $ 5. 5 million to $ 1. 0 million during the second quarter of 2009. the write - bearing loss on commercial investment banking investments was $ 17. 8 million to $ 11. 5 million in the first quarter of 2009. these increase resulted in the decrease of these deferred tax assets during the first nine months of fiscal 2008. the decrease [CLS]

G.2 OpenWebText

Complete decoded OpenWebText samples (GPT-2 tokenizer, $L = 1024$), from the runs behind Tables 21, 18, and 20; each header gives its own sampler and budget. Each is one unedited sequence of 1024 tokens, shown in full; `<|endoftext|>` is the document boundary emitted by the model. Sixteen samples were drawn per setting and one is reproduced here. The cards span the entropy range of the self-conditioned models: the first two are delayed-onset configurations that hold the data entropy $H_{\text{data}} = 5.44$, the last two sit well below it.

vMF + self-conditioning

ODE, warp-aware, $n = 128$**, delayed onset** $\tau = 1/3$
gen. PPL = 70.7, $H = 5.45$.

<|endoftext|>, but it's expected that next week 100 people will be forced to reunite to the nuclear plant in Fukushima Laeu.
Last June, Japan's Daima Electric Plant experienced a nuclear leak from the Monroe River's northern-east Port. Three days afterwards, the New Zealand Electric Company's Tokichi plant attempted to impueance 15,000 electrical service workers.
The much-caated Fukushima catastrophe also toppled 1,400 jobs.
READ MORE: Sugo in New Zealand, despite New boom firing on
Tamushima will be bankruptcy on Tuesday, more than a year and two weeks after taking in.
"Based on orders received from one company active, Japanese Ministry of Defence, there is a planned shutdown regarding the Tok nuclear power plant operation," the company said in a statement.
"The shutdown is very effective. Severe alarm systems will continue to operate in the course of days, although additional stations will be shut down wherever appropriate.
"We take a precaution for incidents to affect the safety of serviceages and prevent disruption of our radio networks."
There would be "no limit on whether a grid construct makes short malfunctions and Japan will do everything it takes to ensure our power safety and levels security".
Japan spokesman Adrian The Earth Group, who represents the environment and an environment group campaigning for nuclear regulation, said on Tuesday the plant released gigigrams of nuclear water, causing the loss of 2,700 lives.
READ MORE: Fukushima warns in Fukushima Laeu plant
Mr Gani released a statement on Wednesday saying that "defcipled parties are sworn to blame".
"Our tribe and colleagues were appointed for the democratic votes, while others may have failed popular campaigns for atomic power construction already signed.
"Our peoples rely on telling the facts to make decisions, which is why we are at the heart of this tragedy running up as it seems is the thousands of deaths that have been reported across Australia and Europe.
“Our indigenous staff in the country have recently informed others we could do our part to finish, design, engineer, repair or mitigate this accident."
“Some of the financial studies anticipated in the work will be prescribed red leather faces and hardware over time and will be added to public order in perhaps few weeks."
Mr Gani said he was "poursing to get as much support here as possible from the New Zealand Liberal government".
"We will act here to help against this, and this impoverishes this island," he said. “To us, as everybody else, we've always been a people, and we're proud to be a part that's a New thing.”<|endoftext|>“I'm proud to be heard,” - Donald Donald Trump
Hundreds of thousands of ad members attended a Jupiter, Florida fundraiser and tickets to Donald J. Trump seeking hearings range from $1,500 to $10,000, according to WSB-TV.The Jupiter fundraising sign-up was buzzing: “Our Tickets Go Nice to support our fellow Chovers does not support Donald Trump.”Unfortunately, the crowds were overflowing in the Byzantine fundraising Hall.“Many people spend time with family and want to hear Inside trust about their life and sit in the mirror, to who plays and

what they love," said Susan Marallova, Florida fundraiser chairman.The fundraiser quoted the former entrepreneur: "We're aware of a lot of big people to us as a donors. They like to hear from each other, and we're always proud to be heard." (Watch ABC News)"And if you're a political example if you run into a [...] plan Social Security, we're also going to make you pro-life. So you have to make sure your values are strictly," according to the 68-year-old.When Does Trump admitted that he thought he'd use the fundraiser a lot of cash and decided to shell out Trump junior "promises," he said, "But we don't make a lot of money."<|endoftext|>Scientists wonder how arms could be applied to a human robots
We could pull the challenge if we could interpose humans into another Predator
Once in life, humans are able to recognize or recognize objects from their materials. To deal with this challenge is to make the robot a human.
Earlier this year, researchers at the Google UCSF National Institute for Advanced Science demonstrated the need to patent a robot-in-lectrol chip to humans recognize more objects like their arms via an accelerated brain operation by their brains.
E' Xiao and his undergraduate colleagues, a Research Science in Biology for the Robotics and Engineering Science Management Office at Carnegie Mellon University, predict current developments in Autonomous A Human (VID). The findings were published last month in the journal Science (Feb. 6<|endoftext|>

vMF + self-conditioning

PC, $(32, 3)$**,** $\varepsilon = 10^{-4}$**,** wd**, delayed onset** $\tau = 1/3$
gen. PPL = 52.7, $H = 5.44$.

<|endoftext|>
Criticism [edit]
For conspiracy scientists, there is a fundamental difference between admitting and examining several more explanations and not logically distinguishing what type of findings one feels as authentic. Thus, the proponents of conspiracy theories should abandon neither religious or religious truths and selectively believe the truth is all being "telling".[1][2][3][4] They must not counter the idea that conspiracy theories are fringe theories, or beliefs away from them. Inspire scientific analysis distinguishes conspiracy theories based on science. As a result, it is possible to reconcile their beliefs in a clinical manner.[5]
Psychic professionals, clinicians or even representatives of a general medical team or psychiatric practitioner should be reminded that you must practice or consideration conspiracy theories in order to make informed informed decisions prior to events contained in the evidence being clearly presented. Readers should also agree with the decision based on their beliefs, background, beliefs, and theories of the conspiracy suspects.[6]
Religion [edit]
Islamians feel many Muslims embrace conspiracy theories placing doubt into the "wrong" religion.[6][9]
Religion [edit]
When conspiracy theorists are asked to explain their sectarian beliefs, values and the regime, it is not to say that a Muslim holds an allegiance to Islamic State.[10]
Pro-cychvery [edit]
Contivants believe conspiracy theories focus on science rather than evidence.[10][12] This reasoning is not limited. The main problem with conspiracy theories is that they cannot match credible evidence publicly. For example, Internet searches maintained by the 2017 Federal Bureau of Investigation contain over 900 conspiracy theories and conspiracy theories.[13][13] Federal authorities report millions of them each year.[13] Furthermore, they lack the so-called "logic fallacy"[13] which states that a fallacy fallacy addresses a "unexpected number" possible lengths to disproport "facts from the sources".[16] Some websites promote "rastery claim" when people believe "that conspiracy theories based on false evidence have conflicting viewpoints".[17]
Random evidence shows the success of conspiracy theories on the internet.[17] A significant number of websites rely on experts to further verify such claims, because conspiracy theorists believe that "stages should only be given to read something big".[19] Psychological conspiracy theories have also been dubbed by the media "ill-informed".[21]
Pro-systematic theories [edit]
Popular, secular and statistical beliefs are perhaps most popular pedavaging conspiracy theories on concepts almost exclusively.[22] Although the conspiration is true, the theories tend to respectfully agree with belief ideas.[20][23][24]
For example, Sir Mamhan Kealday Infar, a 13-year-old school student, was allegedly attacked in Amrikhar district in 1999 by the terrorist group "dhhan Shirab" and murdered several girls.[23] The above article

argues that pseudipcience values are pre-corrected and suggests that it is better to accept these pro-systematic, world-conpetual conspiracy theories.[27]
A new body of research from 1998–2010, which classified conspiracy theories as "confidence-mining", changed the term "skeptology" to "independentular science".[28][30][31] Conspiracy theories are esteemed among scientists such as psychologist Ralph Ross.[30][32][33][34] Conspiracy theories are more popular than among non-religious religions such as Islamic Zoroidhs.[32]
Religion in the United States [edit]
Religion and conspiracy theories also popularize regularly. According to a Demosium Group which polled voters in the United States during the preceding 2000 election, 60% of Americans believe that these conspiracy theories "are too harmless for all Americans".
Foci theories & non-religious beliefs [edit]
A 2013 Gallup poll survey showed the majority of Americans as wary of conspiracy theories as either secular or faithless. 72% believe all children and 66% believe in science:[9]
What you should say is that I'm putting on all who the faithly are...I know, because I believe in my own belief, that from death you live the belief you believe comes far from me. But I don't from religion - everything I believe comes far from nature because you choose all the facts."[38]
A 1998-15 survey conducted by the Gervaeb Centre for the Religious Science reported an "open grade on possible belief"[62] with 219 reporting that there was at least no connection between belief in children and adult science.[19][full citation needed]
In 2013 the American Academy of Sciences for Children and Kidcrators stated that there was no connection between belief in children and infants or in human beliefs.[89][11]
Those who believe in conspiracy beliefs and experience difficulties in babies' ability to understand science are more likely to adopt conspiracy theories due to the following:
attention and physical differences
honderance<|endoftext|>

Linear + self-conditioning

PC, $(128, 7)$**,** $\varepsilon = 10^{-2}$**,** wd
gen. PPL = 21.8, $H = 5.16$.

<|endoftext|> far as it goes. We still didn't have enough information to get into this whole thing exactly what we could have been of the two that hadn't, like, read up on a scrap that didn't all wrap up. We didn't have to worry about my brothers', either. So we, we didn't have to worry about the papers of all of that."
But the scrap cracked and he looked down at Martin, stunned. "uuaaaaaya! He was traded in the mid-30s. Obviously he got seriously hurt, we're gonna have a problem with catching the pain a lot on his wrist, a little!"
"Yeah! Joel was right," said Patrick, banging on her phone. "What would you do about that?"
Grind Huslan's face was placed on a finger. "You're such a man. I never asked to do that. I made him bruise. There was a guy who was gonna shoot twenty-five or six. You weren't gonna get paid for it. You didn't use the weapon. He was gonna kill sixteen."
Steven rolled back. "That's the dumbest thing I've ever known."
"No," said Gerry. "Maybe is that wrong with killing?"
"Oh, I'm not aware of it, Billy, but these guys have no problems with it—it's a bit wrong with killing, aren't they?"
"Yeah, Jeff Tummy, who shot twice," Patrick said dismissively, "said, 'I shot twenty-teen, I don't want to tell you about it.' Her bruizer raised her gun at Connie and shouted, 'I shot twenty-seven, eight—how did those guys do that? You shot twenty-three. I shot five and five. Fuck Chuck Huslan's memory!"
Tummy looked at Kelly. "Did that? I shot sixty. I stole 20 pounds of meat, like I didn't do that for any reason. These guys didn't do this either."
Sabotita frowned then. Kelly leaned into the window, slamming down the window.
"Why did you say that?" Mitchell stun pointed at Connie in an agitated manner. "Why didn't they say something like that?"
"Did you say something like that?" Connie asked.
"I won't give credit," Jim said.
"Didn't you give credit."
Gosh, Chuck had nothing to do with that. Harry Lamaine had nothing to do with him.
"He hit me in the back," Tummy quused. "Ay-whooly. How did Billy get shot twice?"
"Nothing," Graham laughed, crying. "I wanted to kill him! Not all of this shit would really red scratch my character. My crew and nobody even wanted to stop him!"

“I don’t want to lock him up, fear cutting—I love catching!”
“Keeping that in mind. These guys were cool,” said Kelly. “I’m gonna do fine. Not bad for shooting these guys.”
“That sounds fucking silly,” said Connie.
“That’s too much.” He paused and took a turn to chant loudly. “That’s too big to go by.”
Kelly swung his fist at Sherrod, who tapped the knife again.
“Throw his gun.”
Centley stepped out of his way, “Jon Barteo’s a badass, right now!”
Mitchell looked at Sherrod’s cart. Timward three co-workers took heads, both hurled at each other.
Tummy started to each other, joined by Sherrod’s jeer, smirking gunsters and guts.
“Listen!” continued Connie, “he’s the gun, okay? I’ve got no control over what I want.”
“Wait, isn’t that what I just said?”
Mitchell gave Mitchell a nod, leaning back at Tummy. “I’ve got no control. I’m having good at home.
Good Christmas stores are really where people would buy it.<|endoftext|>

Geodesic + self-conditioning

ODE, warp-aware, $n = 1024$
gen. PPL = 36.3, $H = 5.23$.

<|endoftext|> so let’s really draw you on her thoughts on all of the things Manziel brings to the franchise.
Podcast: Jon Grick, Eric McCllethe: That’s a National Report dynamic, and all of this history is celebratory, but what do you focus on?
Strkinger Well, man. You know, we don’t care. These are things people all know. We’ve been following for a while. It’s pretty important that we react to those comments as well.
Strkinger: Like, Marty is the kind of quarterback, he’s a burner, he’s a playmaker #4.
But we talk about Johnny. Because Johnny knows what he knows, he sees a great role, but he’s here.
And you’re starting to see what he lacks or nobody else.
But he’s a good, man. He’s a runner, but he has every opportunity. He says he’s a Wonderlinger or not a pass rusher.
And that’s why you’re starting to see him.
That’s why you’re starting to see him.
And that’s why you guys are starting to see him.
Now, this is the reason Michael Manziel runs no.
You can see him, but he runs unnestly.
He seems to give off Tom Roethlisberger in a way. Manziel will build the quarterback that you all know, but if he knows who he’ll be, it’s tough to see him improve.
So when he comes out, deeper Manziel is a tough, athletic, kid. He knows nobody knows what he is.
And he’s a tackling, man. He’s a quarterback, but he has talent.
Rate’s a burner. And he’s a playmaker #4. But he’s dynamic, strong-type kid. He knows what’s special, it’s tough to see him play well.
Now, every minute he comes out. He’s a great, athletic, kid.
And he’s a burner. And he’s a playmaker #4.
He’s a burner, but he definitely knows who he can be.
Or. Johnny Manziel is the kind of quarterback. He’s tough, man. He’s a burner, but he has talent.
And that is inseparable.
He does this. He’s been an explosive. And in Manziel’s life, he’ll be very great. And he’s going to excel in high school.
But that is inseparable.
But does he ever have to work his way. It doesn’t mean he knows he is athletic and yet he knows he’s total and finally smart down. That doesn’t mean he needed a chance to rush because he’s not anything. He’s just tools. He still has to play one-on-one.
And that is inseparable.
He does this. He’s been an explosive. And in Manziel’s life, he’ll be very great. And he’s going to excel in high school.
But that is inseparable.
Well, does he do everything as quarterback. He didn’t open up. And he’s trying to play in a spectacular way.
This guy really fits. And he’s been in the open position. Average, outstanding rating. He just wants to see those suitsers.

Does Johnny Manziel have everything around. He's all around. He just wants to see those suits.
Well, does Manziel have everything around. He does everything as quarterback. He didn't open up. And he is trying to play in a spectacular way.
This guy really fits. He's been in the open position. Average, outstanding rating. He just wants to see those suitsers.
And he's doing everything. And he's doing everything. And I think he's going to have an incredible receiver. And he's going to strip and a great top 1 threat. And it's not all he scares, but he's going to have an amazing receiver. And he's going to need some wide receivers for him in the offseason. And that's going to be a solid receiver.
And I'd hate to hear right now how the Colts use the combination of the AFC mix at the same time. He's going to need something bigger than Peyton Manning, he can take on all the plays and the competition. And that's going to be in the mix. And it's not all he gets. Incover and great top 1 threat. And I think he's going to have an easy defense. And he'll be a threat to everybody, but he's going to need some wide receivers for him in the offseason. And he's going to be a solid receiver. And that's going to counterbalance over his upside, which anyone is aware. And that's going to be in the mix.
Hell, I'd<|endoftext|>

G.3 TinyGSM

Generated programs at $\mathrm{NFE} = 1024$ on the full 1,319-problem GSM8K test set, under the configurations of Table 3. The Geodesic card uses uniform spacing, the only Geodesic run whose programs were saved.

vMF + self-conditioning

PC, $(512, 1)$**,** $\varepsilon = 10^{-2}$**,** w**, seed** 46
accuracy $= 43.0\%$ $(567/1319)$.

Test item 0, correct.

```
def simple_math_problem() -> int:
    """
    Janet's ducks lay 16 eggs per day.
    She eats three for breakfast every morning and bakes muffins for her
friends every day with four.
    She sells the remainder at the farmers' market daily for $2 per fresh duck
egg.
    How much in dollars does she make every day at the farmers' market?
    """
    eggs_left = 16 - 3 - 4

    revenue = eggs_left * 2

    result = revenue

    return result
```

Test item 1, correct.

```
def simple_math_problem() -> int:
    '''
    A robe takes 2 bolts of blue fiber and half that much white fiber.
     How many bolts in total does it take?
    '''
    blue_bolts = 2
    white_bolts = blue_bolts / 2
    total_bolts = blue_bolts + white_bolts
    result = total_bolts
    return result
```

vMF + self-conditioning

ODE, warp-aware, $n = 1024$**, seed** 45
accuracy $= 23.0\%$ $(303/1319)$.

Test item 0, wrong *— returns* -72*, gold* 18*.*

```
def simple_math_problem() -> int:
    '''
    Janet's ducks lay 16 eggs per day.
    She eats three for breakfast every morning and bakes muffins for her
friends every day with four.
    She sells the remainder at the farmers' market daily for $2 per fresh duck
egg.
    How much in dollars does she make every day at the farmers' market?
    '''
    eggs_sold = (16 - 3) * 4
    eggs_left = 16 - eggs_sold
    money_made = eggs_left * 2
    result = money_made
    return result
```

Test item 1, correct.

```
def simple_math_problem() -> int:
    '''
    A robe takes 2 bolts of blue fiber and half that much white fiber.
     How many bolts in total does it take?
    '''
    blue_bolts = 2
    white_bolts = blue_bolts / 2
    total_bolts = blue_bolts + white_bolts
    result = total_bolts
    return result
```

Geodesic + self-conditioning

ODE, uniform spacing, $n = 1024$
accuracy $= 10.8\%$ $(143/1319)$.

Test item 14, correct.

```
def simple_math_problem() -> int:
    '''
    In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of
the remaining enrolled in jazz dance, and the rest enrolled in hip-hop dance.
    What percentage of the entire students enrolled in hip-hop dance?
    '''
    total_students = 20
    contemporary_students = total_students * 0.20
    remaining_students = total_students - contemporary_students
    jazz_students = remaining_students * 25/100
    hiphop_students = total_students - contemporary_students - jazz_students
    percentage_hiphip_hop_students = (hiphop_students / total_students) * 100
    result = percentage_hiphip_hop_students
    return result
```

Test item 15, wrong *— returns* -4992.0*, gold* 125*.*

```
def simple_math_problem() -> int:
    """
```

```
    A merchant wants to make a choice of purchase between 2 purchase plans:
jewelry worth $5,000 or electronic gadgets worth $8,000.
    His financial advisor speculates that the jewelry market will go up 2.5%
while the electronic gadgets market will rise 1.2% within the same month.
    If the merchant is looking to maximize profit at the end of this month by
making a choice, how much profit would this be?
    """
    jewelry_price = 5000000*2
    electronic_price = 80000*1.012
    new_price = 5000 + electronic_price
    total_price = 8000/1000 + electronic_price
    profit = total_price - new_price
    result = profit

    return result
```

Linear + self-conditioning

PC, $(512, 1)$**,** $\varepsilon = 10^{-1}$**,** d**, seed** 46
accuracy $= 28.6\%$ $(377/1319)$.

Test item 1, correct.

```
def simple_math_problem() -> int:
    '''
    A robe takes 2 bolts of blue fiber and half that much white fiber.
     many many bolts in total does it take?
    '''
    blue_bolts = 2
    white_bolts = blue_bolts / 2
    total_bolts = blue_bolts + white_bolts
    result = total_bolts
    return result
```

Test item 3, correct.

```
def simple_math_problem() -> int:
    '''
    James decides to run 3 sprints 3 times a week.
     He runs 60 meters each sprint.
     many many total meters does he run a week?
    '''
    meters_per_sprint = 60
    sprints_per_week = 3
    num_sprints_per_week = 3
    total_meters_run = meters_per_sprint * sprints_per_week *
num_sprints_per_week
    result = total_meters_run
    return result
```

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: We describe the general problem setting. The approach, the theoretical contributions and then exactly what our experimental results validate.

   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: Limitations are discussed in Appendix A.

   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: All proofs are given in the Appendix with assumptions.

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 use public datasets. Describe algorithm and method in detail. The code is public on GitHub (link in the abstract).

   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: The code is public on GitHub (link in the abstract), with scripts for evaluating to get the results in the paper.

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: In Section Implementation Details and in Experiments we outline this.

   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: Appendix F.1 reports the mean and standard deviation over five generation seeds for the LM1B, OpenWebText, and TinyGSM result-table entries. Retraining with several training seeds is too computationally expensive.

   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: Outlined in Implementation Details.

   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: Yes we follow standard practice.

   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: [Yes]

    Justification: Social impacts are discussed in Appendix A.

    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 training scale is not large enough for the model to pose a risk.

    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: Owners are properly credited and licenses are respected.

    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: The released code (GitHub link in the abstract) is documented in its README.

    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: This paper does not use crowdsourcing 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: This 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: No LLMs have been used in any core component of this paper.

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.