SPHERICAL BROWNIAN DIFFUSION FOR LANGUAGE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous diffusion language models involve two coupled design choices: how to corrupt continuous token representations and how to invert that corruption at generation time. Recent work has explored language generation on the unit sphere through geodesic and von Mises–Fisher (vMF) paths, raising the question of which forward process to use on this geometry and how much that choice matters relative to inference. We introduce **Spherical Brownian Diffusion**, using Brownian motion as the canonical isotropic diffusion on the sphere. Its heat semigroup provides a principled forward process, an additive dimension-free noise clock, and an exact ancestral bridge, while a clean-token posterior learned with cross-entropy supports multiple samplers. To separate corruption from inference, we derive high-dimensional approximation rates between Brownian and moment-matched vMF corruption and use this relationship for controlled substitutions. On Sudoku, Brownian and moment-matched vMF give nearly identical results across four samplers and five budgets under otherwise identical settings. In contrast, with the corruption process fixed, sampler choice produces a 25.0-point spread at a sampler budget of 180 calls. Using the ancestral bridge, we split this gain into a memory effect from re-noising the state, present on Sudoku and GSM8K, and a token-selection effect on GSM8K and largest at small budgets. Across reasoning tasks, categorical refinement improves performance, whereas unconditional text generation benefits from stochastic updates that preserve diversity.

1 INTRODUCTION

Most language models are autoregressive (Vaswani et al., 2017): they generate tokens sequentially and cannot revise earlier commitments. Diffusion models instead update many positions in parallel and can revise them throughout generation. In continuous diffusion language models (Li et al., 2022; Dieleman et al., 2022; Gulrajani & Hashimoto, 2023; Yang et al., 2026; Chen et al., 2026), tokens are represented in a continuous space. Beyond choosing how to represent these discrete data, two coupled design choices remain: how to corrupt the representations and how to invert that corruption at inference (Dieleman et al., 2022; Deschenaux & Gulcehre, 2026; Chemseddine et al., 2026b).

The representation of discrete data on a continuous space creates a trade-off: one-hot states represent tokens exactly, but their dimension scales with the vocabulary and their distances carry little semantic structure (Deschenaux & Gulcehre, 2026); lower-dimensional embeddings avoid both issues but require choosing a geometry. The unit sphere is a natural bounded choice, and recent spherical models use a geodesic path (S-FLM) (Deschenaux & Gulcehre, 2026) or a von Mises-Fisher (vMF) path (Chemseddine et al., 2026b). Because each pairs its forward process with a different inference procedure, the effect of the corruption kernel itself is hard to isolate. We ask how the kernel interacts with training and inference, and how much it matters relative to the inference rule.

We introduce **Spherical Brownian Diffusion** (SBD), which corrupts token embeddings with Brownian motion on the unit sphere; its transitions are given by the spherical heat kernel, which we call *Heat*. Up to time scaling, this is the canonical isotropic diffusion on the sphere. SBD is trained with cross-entropy to predict the clean-token posterior (Figure 1), giving a single model that supports both continuous reverse-time SDE/ODE sampling and categorical refinement. This lets us vary the inference rule without changing the learned denoiser; for refinement, q-argmax selects the posterior mode and re-noises its embedding with Heat.

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

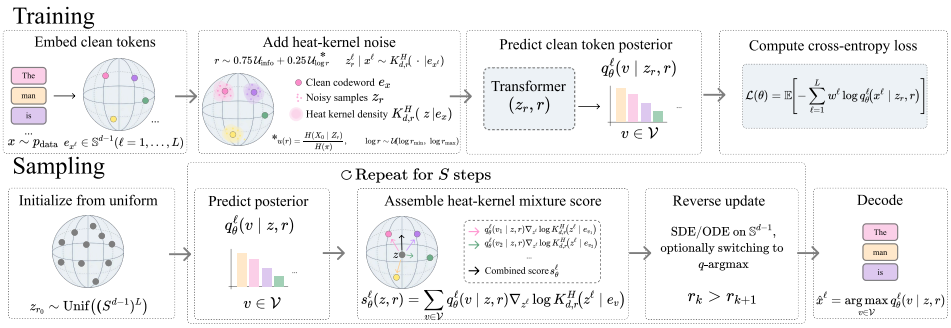

Figure 1: Overview of SBD. **Training (top):** token embeddings are corrupted with the spherical heat kernel, and a Transformer predicts the clean-token posterior with cross-entropy. **Sampling (bottom):** the posterior either builds the score for reverse SDE/ODE updates or drives categorical refinement, which selects tokens and re-noises their codewords.

Heat also admits an exact ancestral bridge: the distribution of a less noisy state given the current state and a clean token. The bridge controls how much of the current state the sampler keeps: we add extra noise before each reverse step and bridge back, so more noise means less memory, at no extra network cost. Other spherical paths we consider do not allow this, so Heat serves both as a reference process and as a tool for separating state memory from token selection. To isolate the effect of the kernel itself, each kernel must be paired with compatible inference dynamics. We therefore introduce a conditional quantile map, built on the radial-transport framework of Chemseddine et al. (2026b), that handles multimodal paths without stiff velocity integration.

Results. Using matched transport, crossed training–sampling substitutions, and the Heat bridge, we study the roles of the corruption kernel in training, continuous transport, and re-noising for categorical refinement (Sections 4.2, 4.3 and 4.4 respectively). We find that when training and sampling use compatible continuous transport, the path itself matters little: Heat, moment-matched vMF, and distant mixture kernels reach similar accuracy. Moment-matched vMF provides a principled nearby control, as our theory shows that it approaches Heat at the critical noise scale. In contrast, sampling with a kernel different from the one used in training is costly: replacing Heat by a mixture kernel on a Heat-trained GSM8K model reduces accuracy to near zero, while a short matched fine-tune largely restores it. Kernel shape also interacts with refinement: mixture re-noising helps q-argmax but hurts confidence gating. Inference has the larger effect overall. With Heat fixed, four inference rules span 25 points on hard Sudoku. The ancestral bridge attributes much of this gap to state memory: retaining the previous state preserves noise on Sudoku and token commitments on GSM8K. This cost grows with finer step grids on both Sudoku and GSM8K. Across Sudoku, GSM8K, and OpenWebText, SBD is competitive with recent continuous models, with the strongest method depending on the inference regime; q-argmax also improves a released S-FLM model, showing that the refinement gain transfers beyond Heat.

Contributions. **(1)** We introduce **SBD**, to our knowledge the first language diffusion model whose forward corruption is given directly by the spherical heat semigroup on token embeddings. Its heat semigroup offers a principled corruption process, a clean-token posterior supporting continuous and categorical sampling, and an exact ancestral bridge controlling trajectory memory (Section 3). **(2)** To compare spherical corruption processes without confounding kernel shape, time parameterization, and inference, we develop a conditional quantile-map integrator, **heat–vMF approximation rates**, and a shape residual (Section 3). **(3)** Using these tools, we show that **inference can matter more than the spherical kernel** in our controlled settings, and isolate state memory as a key factor limiting refinement, with a cost that grows with budget (Sections 4–5).

2 Background

Notation. Let $\mathbf{x} = (x^1, \dots, x^L) \in \mathcal{V}^L$ be a token sequence, with token embeddings $\mathbf{e}_v \in \mathbb{S}^{d-1}$. Then $\mathbf{z} \in (\mathbb{S}^{d-1})^L$, and the tangent projection is $P_{\mathbf{z}} = I - \mathbf{z}\mathbf{z}^\top$.

2.1 Non-autoregressive generation

Autoregressive models generate tokens from left to right. Masked discrete diffusion models instead corrupt tokens and learn to reconstruct them in parallel (Sahoo et al., 2024; Austin et al., 2021).

Continuous alternatives embed tokens in a Euclidean or manifold-valued space and learn a reverse diffusion or flow (Dieleman et al., 2022; Lipman et al., 2023; Zhang, 2026). These models can revise several positions at each step, but their behavior depends on the latent geometry and on how continuous states are mapped back to tokens.

2.2 SPHERICAL PATHS

S-FLM places token embeddings on a sphere and follows deterministic interpolating paths (Deschenaux & Gulcehre, 2026). A related approach uses von Mises–Fisher (vMF) corruption (Chemseddine et al., 2026b). For a clean codeword $\mathbf{e}$, a vMF conditional has density proportional to $\exp(\kappa\langle \mathbf{z}, \mathbf{e}\rangle)$ and score

$$\nabla_{\mathbf{z}} \log p_{\mathrm{vMF}}(\mathbf{z} \mid \mathbf{e}) = \kappa(\mathbf{e} - \langle \mathbf{z}, \mathbf{e}\rangle \mathbf{z}). \tag{1}$$

Its radial coefficient is constant in the cosine (Mardia & Jupp, 1999). Spherical Brownian motion has the same tangent direction but a cosine-dependent coefficient, which is the main geometric difference studied here.

2.3 SPHERICAL BROWNIAN DIFFUSION

Brownian motion is the canonical isotropic diffusion on the sphere. We use the dimension-free coordinate $r = -\log \rho$, where ρ is the retained mean cosine, and intrinsic heat time $\tau_d(r) = 2r/(d-1)$. Here, ρ is the average cosine similarity between noisy and clean embeddings, and τ_d is the elapsed time of standard Brownian motion on the sphere. The coordinate r is comparable across dimensions within one kernel family, but equal r does not imply equal token information across families; analogous time-warping ideas were used in CDCD (Dieleman et al., 2022). Section 4 therefore also uses a measured information clock. In the Brownian clock, the forward process satisfies

$$\mathrm{d}\mathbf{Z}_r = -\mathbf{Z}_r \,\mathrm{d}r + \sqrt{\tfrac{2}{d-1}} P_{\mathbf{Z}_r} \,\mathrm{d}\mathbf{W}_r. \tag{2}$$

Its transition density is the spherical heat kernel $K^H_{d,r}$, which solves $\partial_r K^H_{d,r} = \Delta_{\mathbb{S}^{d-1}} K^H_{d,r}/(d-1)$ and converges to the uniform distribution as $r \to \infty$. Score-based diffusion reverses this process using the score of the noisy marginal (Song et al., 2021; Lou et al., 2023).

For discrete data, the marginal score is a posterior average of conditional heat-kernel scores:

$$\nabla_{\mathbf{z}^\ell} \log p_r(\mathbf{z}) = \sum_{v \in \mathcal{V}} p(x^\ell = v \mid \mathbf{Z}_r = \mathbf{z}) \nabla_{\mathbf{z}^\ell} \log K^H_{d,r}(\mathbf{z}^\ell \mid \mathbf{e}_v). \tag{3}$$

A path-level comparison with S-FLM and vMF is given in Appendix D (Table 4).

3 SPHERICAL BROWNIAN DIFFUSION

This section develops Spherical Brownian Diffusion (SBD) and the tools we use to study the role of the corruption kernel. We first define the Heat forward process and show how a clean-token posterior determines its marginal score (Section 3.1). We then describe training and two ways we use this posterior at inference: standard continuous reverse dynamics and categorical refinement. Our main inference tool is an exact Heat bridge, which lets us vary state memory independently of token selection. Finally, to compare Heat with other spherical corruptions without confounding the forward kernel with the sampler, we introduce matched transport and identify moment-matched vMF as a principled nearby-kernel control. These tools support the controlled kernel study in Section 4.

3.1 FORWARD PROCESS AND MARGINAL SCORE

Our starting point is to use spherical Brownian motion as the tokenwise corruption process. Given a clean sequence $\mathbf{x}$, we corrupt each token embedding independently with the spherical heat kernel introduced in Section 2.3. At noise level r, we have: $\mathbf{z}^\ell \mid x^\ell \sim K^H_{d,r}(\cdot \mid \mathbf{e}_{x^\ell})$. Dependencies between positions are modeled by the Transformer rather than by the forward process.

We sample the heat kernel directly through its cosine distribution. We first draw the cosine c with respect to the clean codeword, sample a uniform tangent direction $\mathbf{v}$, and set $\mathbf{z}^\ell = c\mathbf{e}_{x^\ell} + \sqrt{1-c^2}\,\mathbf{v}$.

Because the heat kernel is isotropic, its conditional score depends only on $c_v^\ell = \langle \mathbf{z}^\ell, \mathbf{e}_v \rangle$:

$$\nabla_{\mathbf{z}^\ell} \log K_{d,r}^H(\mathbf{z}^\ell \mid \mathbf{e}_v) = a_{d,r}^H(c_v^\ell)(\mathbf{e}_v - c_v^\ell \mathbf{z}^\ell), \qquad a_{d,r}^H(c) = \partial_c \log K_{d,r}^H(c). \tag{4}$$

Rather than learning the continuous score directly, we train a Transformer to predict the clean-token posterior $q_\theta^\ell(v \mid \mathbf{z}, r)$ with cross-entropy. This follows the score-interpolation view of continuous diffusion for categorical data (Dieleman et al., 2022): the marginal score is the posterior average of the known conditional scores. On the sphere, the corresponding posterior-weighted tangent-sum representation was also used by Chemseddine et al. (2026b). For Heat,

$$\mathbf{s}_\theta^\ell(\mathbf{z}, r) = \sum_{v \in \mathcal{V}} q_\theta^\ell(v \mid \mathbf{z}, r)\, a_{d,r}^{\mathrm{H}}(c_v^\ell)\big(\mathbf{e}_v - c_v^\ell \mathbf{z}^\ell\big). \tag{5}$$

Thus, a single categorical posterior provides the score required by continuous samplers and can also be used directly for categorical refinement.

3.2 Training and codebooks

We train this posterior directly by denoising classification. For clean data $\mathbf{x}$ and a sampled Heat-corrupted state $\mathbf{z}$, where $p_r^H(\mathbf{z} \mid \mathbf{x}) = \prod_{\ell=1}^L K_{d,r}^H(\mathbf{z}^\ell \mid \mathbf{e}_{x^\ell})$, we predict the clean tokens with cross-entropy,

$$\mathcal{L}_{\mathrm{CE}}(\theta) = \mathbb{E}_{\mathbf{x},\, r,\, \mathbf{z}}\Big[-\textstyle\sum_{\ell=1}^L w_\ell \log q_\theta^\ell(x^\ell \mid \mathbf{z}, r)\Big], \tag{6}$$

where w_ℓ is a task-dependent loss weight. We use frozen or learned unit-norm codebooks; their construction and the task-specific noise distributions are described in Appendix D.

3.3 Continuous dynamics and posterior refinement

Having learned q_θ, we next describe how SBD generates. We consider two uses of the same posterior. First, it can be converted into a marginal score and used with standard reverse-SDE or probability-flow ODE dynamics. Second, it can act directly on token predictions through categorical refinement. We introduce q-argmax for the latter and, exploiting the Heat semigroup, an ancestral bridge that lets us interpolate between retaining and discarding the current continuous state.

Continuous reverse dynamics. As a standard continuous baseline, we convert the posterior into the marginal score using Equation 5. We then integrate either the reverse SDE or the probability-flow ODE on the sphere. The reverse SDE is stochastic, whereas the ODE is deterministic given its initialization. Implementation details are given in Appendix D. These are the standard score-based reverse-time dynamics for a diffusion process (Song et al., 2021).

Categorical posterior refinement. Rather than converting the posterior into a continuous score, we can act on its token predictions directly. We introduce *q-argmax*, which selects the most likely token at each position and re-noises its codeword at the next radius:

$$\widehat{x}_k^\ell = \arg\max_{v \in \mathcal{V}} q_\theta^\ell(v \mid \mathbf{z}_k, r_k), \qquad \mathbf{z}_{k+1}^\ell \sim K_{d,r_{k+1}}^H\big(\cdot \mid \mathbf{e}_{\widehat{x}_k^\ell}\big). \tag{7}$$

The token choice is deterministic given the current state, while re-noising keeps the trajectory stochastic. Recomputing the posterior at every step allows later iterations to revise earlier predictions. This is related to predict-project-re-noise updates and to Diffusion-LM's repeated return to token embeddings, but here the projection is explicitly the clean-token posterior mode followed by a fresh Heat draw (Sahoo et al., 2024; Li et al., 2022).

Ancestral bridge and restart. q-argmax changes two aspects of ancestral sampling at once: it replaces posterior sampling by a mode decision and discards the previous continuous state by re-noising from scratch. To separate these effects, we exploit the Heat semigroup to construct an exact ancestral bridge. For Heat, the semigroup gives the *bridge* density $B_{s \leftarrow R}^H$ of the state at radius $s < R$, given the current state y at radius R and the clean token x:

$$B_{s \leftarrow R}^H(\mathbf{z} \mid \mathbf{y}, \mathbf{e}_x) = K_{d,s}^H(\mathbf{z} \mid \mathbf{e}_x)\, K_{d,R-s}^H(\mathbf{y} \mid \mathbf{z}) \,/\, K_{d,R}^H(\mathbf{y} \mid \mathbf{e}_x). \tag{8}$$

Sampling x from the exact posterior and then from this bridge gives the exact finite-time reverse conditional; numerically, this is exact up to the heat-table accuracy.

We use the bridge to control how much state memory is retained. After evaluating $q_\theta(\cdot \mid \mathbf{z}_{r_k}, r_k)$, we select $\widehat{x}_k$ either by argmax or by temperature-scaled posterior sampling, add $\Delta \geq 0$ forward noise, and bridge back:

$$\widetilde{\mathbf{z}} \sim K^H_{d,\Delta}(\cdot \mid \mathbf{z}_{r_k}), \qquad \mathbf{z}_{r_{k+1}} \sim B^H_{r_{k+1} \leftarrow r_k + \Delta}(\cdot \mid \widetilde{\mathbf{z}}, \mathbf{e}_{\widehat{x}_k}). \tag{9}$$

This requires no additional network evaluations. $\Delta = 0$ retains the ancestral bridge state, while increasing Δ removes memory and approaches the memoryless refinement update. Token temperature independently controls whether selection is stochastic ($T = 1$) or approaches argmax ($T \to 0$).

We also consider **confidence-gated refinement**, which keeps sufficiently confident tokens fixed while continuing to refine the remaining positions. This is related to confidence-based remasking and token retention in masked diffusion, but our gate uses the clean-token posterior at the current noisy state to decide which positions to hold fixed (Sahoo et al., 2024; Agarwal et al., 2026). Unless stated otherwise, the final sequence is decoded from one last posterior evaluation, $\widehat{x}^\ell = \arg\max_{v \in \mathcal{V}} q^\ell_\theta(v \mid \mathbf{z}, r_{\min})$. Further sampler details and NFE accounting are given in Appendix D.

3.4 Matched transport across isotropic kernels

The previous sections define SBD for Heat. To determine whether its behavior comes from the corruption kernel itself, however, we need to compare Heat with alternative spherical kernels under compatible inference dynamics. Using one kernel with another kernel's reverse dynamics would confound the forward process with sampler mismatch. We therefore construct matched transport for each isotropic kernel.

We build on Riemannian flow matching and the radial-transport framework for isotropic spherical paths (Chen & Lipman, 2024; Chemseddine et al., 2026b). For sharply multimodal paths, for which direct velocity integration becomes difficult, we introduce a conditional quantile-map update, $c \longmapsto F^{-1}_{r'}(F_r(c))$, which is the exact one-dimensional monotone transport for a single conditional kernel. Averaging the resulting spherical displacements under the clean-token posterior gives a first-order update. Derivations and validation are in Appendix E.3.

3.5 Heat–vMF comparison as a nearby-kernel control

To isolate the effect of the corruption kernel, we need a principled nearby alternative to Heat. vMF is a natural choice: it is used by recent spherical language models and shares the same isotropic geometry, while differing in the shape of its conditional distribution. We therefore ask whether moment-matched vMF approaches Heat in high dimension, providing a controlled comparison for testing whether local kernel similarity translates to similar learned generation.

A vMF conditional has score $\kappa(\mathbf{e} - c\mathbf{z})$, with a radial coefficient that is constant in c, whereas the heat-kernel coefficient varies with the cosine. Intuitively, the most relevant regime for comparing their shapes is the critical noise scale, where the remaining alignment with the original token becomes comparable to random fluctuations on the sphere, both of order $1/\sqrt{d}$ (Vershynin, 2018). We compare the two kernels at the critical noise radius $r_d = (\log d + \xi)/2$, where the retained cosine is $m/\sqrt{d}$ for fixed ξ and $m = e^{-\xi/2}$. Let $t_d = (\log d + \xi)/(d-1)$ and choose the vMF concentration $\kappa^\star_d$ to match that retained cosine. Under Heat, write $C = \langle \mathbf{Z}, \mathbf{e} \rangle$ for the noisy-clean cosine.

Proposition 1 (Critical-scale heat-vMF approximation)**.** *For fixed ξ as $d \to \infty$,*

$$D_{\mathrm{KL}}\big(H_{d,r_d} \| V_{d,\kappa^\star_d}\big) = \tfrac{m^4}{4} t_d^2 (1 + o(1)), \qquad \mathbb{E}_H\big[(a^H_{d,r_d}(C)/\kappa^\star_d - 1)^2\big]^{1/2} = m\, t_d (1 + o(1)). \tag{10}$$

In particular, these rates are $O((\log d/d)^2)$ and $O(\log d/d)$.

Moment matching cancels the leading linear discrepancy, leaving a quadratic term that controls both rates. The measured shape residual follows the high-dimensional trend: its maximum falls from 0.025 at $d = 128$ to 0.010 at $d = 768$, though not uniformly over noise levels. Appendix C gives the expansion, proof, and scope of the critical-path result. Appendix Figure 15 tests the two predicted coefficients independently across eight dimensions and three critical-noise paths, without fitting their common target $m = e^{-\xi/2}$. Both estimates approach that target as d grows, although the finite-dimensional discrepancy shrinks slowly, approximately as $1/\log d$.

Takeaway. At the critical noise scale, moment-matched vMF is close to Heat in both conditional likelihood and relative score. This gives the kernel study (Section 4) a principled nearby-kernel comparison; the experiments test how that local agreement carries over to learned generation. The proposition concerns the critical path $r_d = (\log d + \xi)/2$, not the full sampling trajectory.

4 WHICH PARTS OF SBD DEPEND ON THE BROWNIAN KERNEL?

SBD uses the corruption kernel in three roles: as the continuous probability path (Section 4.2), as the training channel whose posterior the sampler relies on (Section 4.3), and as the re-noising operator for categorical refinement (Section 4.4). We vary these roles separately, then use the Heat bridge to separate state memory from token selection effects while sampling (Section 4.5).

4.1 SETUP

We pick kernels at increasing distance from Heat: moment-matched vMF, which our theory predicts is close (Section 3.5); SLERP, the path used by S-FLM; and two vMF mixtures whose bimodal shape deliberately departs from Heat. At retained cosine $m = e^{-r}$, a mixture places weights m and $1 - m$ on vMF components with mean cosines $m + \lambda(1 - m)$ and $(1 - \lambda)m$, so its mean cosine is m while $\lambda \in \{0.5, 0.9\}$ changes its shape (Equation 54). The main SLERP comparison truncates the path at $r_{\min} = 0.942$, where its information is exhausted; Appendix E.4 also reports the full path. Because equal mean cosine need not mean equal token uncertainty, we also compare kernels on a measured information clock, and we report a local shape residual that is zero when a kernel differs from Heat only by a time change (Table 1, definitions in Appendix E.3). Unless stated otherwise, Sudoku kernel experiments use hard puzzles, $d = 128$, 20k training updates, 2,000 evaluation puzzles, and three training seeds.

4.2 THE KERNEL AS A CONTINUOUS PROBABILITY PATH

We first ask whether the path itself matters when everything else is matched: each kernel is trained and sampled with its own valid transport. On hard Sudoku at $B = 180$ (181 NFEs with posterior decoding), Heat, moment-matched vMF, truncated SLERP, and the $.5/.9$ mixtures reach 54.5%, 54.7%, 51.4%, 53.3%, and 55.3%, respectively (Tables 28 and 33). Increasing the sampler budget to $B = 720$ (721 NFEs) does not materially change the Heat/vMF/mixture comparison within seed variation. The full SLERP path reaches $56.2 \pm 2.2\%$ at $B = 180$.

The stability depends on matching the dynamics to the kernel. For the $\lambda = .9$ mixture, inserting the mixture score into Brownian ODE dynamics with nearest-codeword decoding gives 0.0% accuracy at 180 NFEs. At the same sampler budget $B = 180$, posterior decoding alone raises this to $15.3 \pm 3.6\%$ at 181 NFEs, and valid transport alone to $11.2 \pm 0.7\%$ at 180 NFEs; using both reaches $55.3 \pm 2.3\%$ at 181 NFEs (Table 31). Thus, changing the kernel requires both compatible transport and compatible decoding rather than a direct score substitution. The full valid-transport comparison is in Table 28; the decoding-by-dynamics control is in Table 31, with transport checks in Appendix E.3.

4.3 POSTERIOR MISMATCH COMPOUNDS ON SEQUENCE GENERATION

Section 4.2 matched training and sampling. In practice they can be decoupled: a trained model may be sampled with another kernel or schedule. We therefore ask how much a posterior trained with one kernel loses on states from another, and whether the loss comes from kernel shape or only its noise level. To separate the two, we optionally relabel the network input using the information clock.

Table 1: Shape-residual ranges.

Kernel	Residual
$d = 128$	
Heat	$\leq .001$
moment vMF	.001–.025
Mix .5	2.0–16
Mix .9	9.2–98
$d = 768$	
info-vMF	.000–.010

Small shape differences are a clock effect. On GSM8K, moment-matched vMF changes q-argmax accuracy by only -0.9 ± 0.4 points at 65 NFEs ($B = 64$ plus posterior decoding). An information-matched vMF loses 3.8 ± 0.4 points, but relabelling by mean cosine reduces the gap to 0.5 ± 0.7 (Table 38). Both kernels have small shape residuals (Table 1), so the difference comes from the noise parameterization, not the shape.

Large shape differences are not. On hard Sudoku, mismatch cost follows the shape residual. Relabelling reduces the $\lambda = .9$ loss from 14.6 ± 2.9 to 9.6 ± 0.8 points but does not help $\lambda = .5$ ($4.4 \pm$

$0.6 \rightarrow 4.8 \pm 0.3$; Figure 8, Table 30). On GSM8K, mismatch is severe: Heat-trained weights sampled with either mixture reach at most 0.5% accuracy, and relabelling lifts $\lambda = .5$ only from 0.03% to 8.7% (Table 38). The cost grows with sequence length. The mixture-to-Heat success ratio drops from 0.47 for programs of mean length 96 to 0.03 at length 189, with correlation -0.96 (Figure 2, right; Table 41). This is consistent with early token errors corrupting the context of later tokens.

Matched training restores accuracy. We fine-tune the same Heat checkpoint for 25k steps on Heat and on the $\lambda = .5$ mixture, and sample each with its own kernel. At 257 NFEs ($B = 256$ plus posterior decoding), the mixture model reaches $26.23 \pm 1.30\%$, close to the Heat control ($27.32 \pm 0.52\%$) and far above the mismatched mixture (0.03%; Figure 2, left; Table 42). Each arm uses one training seed; error bars are over three inference seeds.

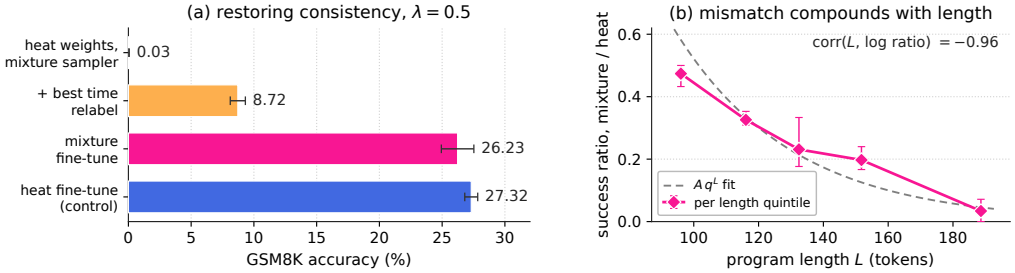

Figure 2: GSM8K consistency restoration at $B = 256$ (257 NFEs, left) and mismatch by program length (right), both with posterior decoding. Error bars show three inference seeds; fine-tunes use one training seed per arm.

4.4 Kernel shape affects categorical refinement

Matched training removes the mismatch cost, but does the kernel become irrelevant? Not for categorical refinement, where it determines how selected tokens are re-noised. On hard Sudoku, matched mixture training and sampling improve q-argmax by 6.0 ± 1.7 points for $\lambda = .5$ and 5.2 ± 2.5 points for $\lambda = .9$. Confidence-gated refinement behaves differently: the $\lambda = .5$ mixture reduces accuracy by 5.8 ± 1.8 points. These effects use five independent training seeds per kernel at $B = 180$ (181 NFEs with posterior decoding). In contrast, Heat and moment-matched vMF remain within roughly one point under tuned gating across the tested budgets. Full kernel results are given in Appendix E.3, where the q-argmax sweep and crossed training/re-noising comparison are reported. Table 32 gives the seed-level gating effects, and Table 24 shows the matched Heat–vMF comparison across inference rules.

The mixture-specific effect does not transfer directly to GSM8K. In the single-training-seed text fine-tune, the matched $\lambda = .5$ model differs from Heat by -1.09 ± 1.80 points with EMA and posterior decoding at 257 NFEs (paired over three inference seeds). Separately, the S-FLM transfer result shows that categorical refinement can carry across a different spherical forward process. Kernel shape therefore has no task-independent ordering: its effect depends on how the kernel is used and on the inference procedure.

4.5 Ancestral bridges separate memory from token selection

Why separate the two. Changing the kernel moves accuracy by only a few points once training and sampling match. With Heat fixed, four inference rules span about 25 points on hard Sudoku. However, q-argmax changes both state memory, through fresh re-noising, and token selection, by replacing posterior sampling with argmax. Either could explain its advantage, and they call for different fixes: decorrelating successive network inputs, or lowering the decoding temperature. Because an exact posterior would make ancestral sampling exact, any gain from departing from it also measures how errors of the learned posterior compound along a trajectory.

How we vary them. The Heat bridge (Equation 8) has two parameters. The restart size Δ adds extra noise before bridging back: $\Delta = 0$ is exact ancestral sampling, which keeps the full state, and $\Delta \rightarrow \infty$ forgets it, as q-argmax does. The temperature T moves token choice from sampling ($T = 1$) to argmax. Neither parameter adds network calls.

Memory helps on both tasks; argmax helps on GSM8K. Removing memory gains 17.7 ± 1.4 points on Sudoku at 129 NFEs ($B = 128$ plus posterior decoding). In contrast, memoryless $T = 1$

sampling is slightly better than argmax there, by 0.5–1.4 points across the reported budgets (Table 2). On GSM8K, removing memory gains 8.1 ± 0.6 points at $B = 64$, while argmax selection adds 11.4 ± 1.0 points at $B = 16$ but only 4.7 ± 2.0 at $B = 256$.

The memory cost grows with budget. As the grid is refined, the bridge deficit grows from 6.4 to 18.4 points on Sudoku and from 3.6 to 9.5 on GSM8K (Table 10). The channel differs by task: redrawing the noise alone recovers most of the Sudoku gap but worsens GSM8K, where retaining old token commitments matters more (Tables 9 and 12). This refines FMLM+'s observation that fresh re-noising avoids repeated rejected predictions (Agarwal et al., 2026).

Table 2: Paired q-argmax gains from removing state memory or using argmax token selection

Task (B)	Memory	Token
Sudoku (16)	$+6.4 \pm 1.0$	-1.3 ± 1.0
Sudoku (180)	$+18.4 \pm 1.8$	-1.4 ± 1.0
GSM8K (16)	$+3.6 \pm 0.3$	$+11.4 \pm 1.0$
GSM8K (256)	$+9.5 \pm 1.4$	$+4.7 \pm 2.0$

Why Heat. The diagnostic needs a time-homogeneous Markov diffusion with a tractable semigroup, so that adding Δ noise and bridging back preserves the Heat marginals. Among the paths studied here, SLERP gives only a deterministic bridge, and vMF marginals give no tractable stochastic one. The spherical Poisson kernel also has a semigroup, but it is a jump process rather than a diffusion.

Takeaway. Kernel choice has no sampler-independent ranking. With compatible continuous transport, all five kernels perform similarly on Sudoku, but a mismatched posterior is costly, especially on long GSM8K programs. In categorical refinement, a mixture can help q-argmax through re-noising yet hurt confidence gating through training. Across both tasks, retaining state memory limits refinement, and its cost grows with budget.

5 Benchmarks

We benchmark SBD on Sudoku, GSM8K, and OpenWebText, spanning reasoning and open-ended generation. SBD is competitive with recent continuous models, with the strongest method depending on the inference regime. We use categorical refinement for reasoning and a hybrid sampler for text (Section 4).

Table 3: Sudoku whole-grid accuracy (%) at $B = 180$ sampler calls (gated SBD: 181 NFEs). FMLM+ and SBD: mean $\pm$ std. over five sampler seeds.

Model	Easy	Medium	Hard
AR (greedy)	14.6	5.1	1.0
MDLM	92.0	77.1	30.2
Duo	96.3	84.7	58.4
CANDI	79.3	45.9	16.7
FLM	94.2	82.7	44.5
S-FLM	94.8	85.2	45.0
FMLM+ PR	$97.77 \pm .13$	$92.99 \pm .33$	$81.65 \pm .36$
Ours: q-argmax	$98.69 \pm .07$	$94.57 \pm .21$	$79.70 \pm .75$
Ours: gated	$\mathbf{99.11 \pm .12}$	$\mathbf{96.66 \pm .15}$	$\mathbf{83.36 \pm .42}$

5.1 Sudoku

Experimental setup. We use the S-FLM Sudoku benchmark with 48k training and 2k evaluation puzzles per difficulty (Alp, 2024; Deschenaux & Gulcehre, 2026). SBD uses a frozen max-min $d = 128$ codebook, with one Heat model trained per difficulty. We evaluate each SBD sampler with five random seeds and report whole-grid exact-match accuracy at a sampler budget of $B = 180$ calls. Full budget sweeps and baseline details are given in Appendix E.2.

Results. Table 3 reports the main comparison. On hard Sudoku, gated SBD reaches 83.36%, slightly above FMLM+ (81.65%), and is the strongest reported continuous/flow model at this budget. On easy and medium puzzles, gated SBD reaches 99.11% and 96.66%, respectively. Full budget sweeps are reported in Appendix E.2.

5.2 Reasoning on GSM8K

Experimental setup. Following prior work, we train on TinyGSM and evaluate on the 1,319 GSM8K test problems (Liu et al., 2023; Cobbe et al., 2021). The model generates one Python program per problem, which we execute to obtain the final answer. SBD uses a learned $d = 768$ codebook and is trained for 250k updates. We compare against recent discrete and continuous non-autoregressive models trained from scratch, together with the reported vMF predictor–corrector baseline with-

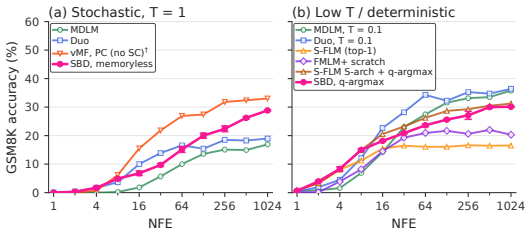

Figure 3: GSM8K accuracy vs. NFE. **(a)** $T = 1$. **(b)** q-argmax for SBD/S-FLM, $T = 0.1$ for MDLM/Duo, endpoint argmax for FMLM+. †: published vMF PC. SBD shows mean $\pm$ std. over three seeds; baselines are single runs.

out self-conditioning. Teacher-assisted and self-conditioned variants are reported separately in Appendix E.6. The controlled decoding curves appear in Figure 3; the full-budget comparison and complete NFE ladder are in Appendix Tables 35 and 36.

Results. At matched $T = 1$, SBD beats MDLM at each budget and Duo from 128 NFEs; the vMF predictor–corrector is stronger, but uses a different training and sampling recipe. Figure 3 shows both decoding regimes. Under low-temperature or deterministic decoding, SBD q-argmax leads only through 8 NFEs; Duo overtakes by 16 and MDLM by 32. At 1,024 NFEs, SBD trails Duo by about 6 points, narrowing the 20-point Duo–S-FLM top-1 gap. Applying q-argmax to S-FLM S-arch reaches 31.2% versus 30.1% for SBD at 1,024 NFEs (one evaluation seed); this is transfer evidence, not a controlled kernel comparison, since schedule, truncation, embedding handling, and training differ. Full baseline provenance is in Appendix E.6.

5.3 Unconditional OpenWebText generation

Experimental setup. We train SBD on OpenWebText with a learned $d = 768$ codebook for 1M updates and generate length-1,024 sequences (Gokaslan & Cohen, 2019; Radford et al., 2019). Following prior work, we evaluate sample quality with GPT-2-Large generative perplexity and diversity with unigram entropy, sweeping temperature across NFE budgets. For open-ended generation, SBD uses a hybrid sampler combining stochastic score updates followed by categorical refinement. The schedule and sampler settings are detailed in Appendix E.11; the temperature-matched k ablation is in Table 45, and all NFE frontiers in Figure 11.

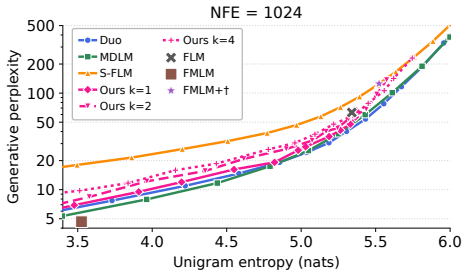

Figure 4: OpenWebText frontier at 1,024 NFEs. Lower generative perplexity is better at matched unigram entropy.

Results. At 1,024 NFEs (Figure 4), SBD improves over S-FLM, while MDLM and Duo retain a better quality–diversity trade-off over much of the range. Less frequent categorical refinement raises diversity but lowers quality, so stochastic updates remain useful for open-ended generation.

6 Related Work

Language diffusion spans discrete models such as MDLM and Duo, and continuous embedding or simplex approaches including Diffusion-LM, CDCD, RePlaid, Dirichlet Flow Matching, and LangFlow (Sahoo et al., 2024; 2025; Li et al., 2022; Dieleman et al., 2022; Yang et al., 2026; Stark et al., 2024; Chen et al., 2026); CDCD's time warping motivates our information clock. Geometric approaches include Riemannian heat-kernel and Fisher–Rao models, S-FLM, vMF spherical flows, and RDLM (Bortoli et al., 2022; Huang et al., 2022; Davis et al., 2024; Cheng et al., 2024; Deschenaux & Gulcehre, 2026; Chemseddine et al., 2026b; Jo & Hwang, 2025). Our conditional quantile map extends spherical flow matching to multimodal cosine laws (Chemseddine et al., 2026b), while q-argmax and confidence gating relate to posterior refinement and token retention (Li et al., 2022; Agarwal et al., 2026). Restart and revision methods likewise modify sampling dynamics or state retention (Xu et al., 2023; Yu et al., 2026). Our Heat bridge gives state-memory interventions a semigroup-based construction; fuller positioning is in Appendix B.

7 Conclusion

We use spherical Brownian motion both as a language model and as a tool for studying continuous language generation. Four lessons emerge. (i) The isotropic kernel matters little when training and sampling are matched; mismatch is severe but largely repaired by matched fine-tuning. (ii) Inference matters more: with Heat fixed, samplers span 25 points, and posterior-based token refinement outperforms continuous integration, with gains that transfer to S-FLM. (iii) Refinement benefits from fresh states: retaining trajectory memory preserves noise and token commitments, with a growing cost at finer step grids. (iv) The best sampler is task-dependent: token commitment helps reasoning, whereas open-ended text benefits from stochastic updates that preserve diversity. Progress may therefore depend less on new forward processes than on inference rules that decorrelate successive states and commit to tokens at the right time.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing, software code and figure preparation, and to independently check mathematical derivations and proofs.

Generative AI was not used to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, or interpret results. The following tasks were not applicable to this work: assist with translation; clean and reformat dataset; support qualitative and thematic data analysis.

All AI-assisted text, code, figures, and mathematical checks were reviewed and verified by the authors. In particular, the authors independently checked the final mathematical derivations and proofs and reviewed and tested the resulting code for correctness. We take responsibility for the final content of this work, including all text, claims, code, figures, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies a generative modeling method using Sudoku, synthetic mathematical problems, and OpenWebText. Models trained on web text can inherit biases and produce inaccurate or harmful content. Generated programs should be executed in a restricted environment. These experiments do not establish that the models are safe for deployment.

REPRODUCIBILITY STATEMENT

We report benchmark-specific model architectures, training budgets, corruption schedules, sampling rules, NFE accounting, and evaluation protocols. The appendices document data splits, checkpoint-selection procedures, seed counts and baseline provenance; where we report variability, we specify whether it is over training or sampler seeds.

ACKNOWLEDGEMENTS

Acknowledgements are withheld during anonymous review.

REFERENCES

Manan Agarwal, Sheel Shah, Chanhyuk Lee, Jaehoon Yoo, Jerry Huang, Seunghoon Hong, Aditi Raghunathan, Jinwoo Kim, and Nicholas M Boffi. Posterior refinement: Fast language generation via any-order flow maps. *arXiv preprint arXiv:2606.24773*, 2026.

Ali Alp. Sudoku puzzle generator. https://github.com/alicommit-malp/sudoku, 2024.

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

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=WBcBhT1NKO.

Nicholas Matthew Boffi, Michael Samuel Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=cqDH0e6ak2.

Valentin De Bortoli, Emile Mathieu, Michael John Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=oDRQGo8I7P.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=DmT862YAieY.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11305–11315. IEEE, 2022.

Jannis Chemseddine, Gregor Kornhardt, Richard Duong, and Gabriele Steidl. Adapting noise to data: Generative flows from learned 1d processes. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=vLQO6nrpYq.

Jannis Chemseddine, Gregor Kornhardt, and Gabriele Steidl. Spherical flows for sampling categorical data. 2026b. URL https://arxiv.org/abs/2605.05629.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g7ohDlTITL.

Ting Chen, Ruixiang ZHANG, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3itjR9QxFw.

Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu. Langflow: Continuous diffusion rivals discrete in language modeling. *arXiv preprint arXiv:2604.11748*, 2026.

Chaoran Cheng, Jiahan Li, Jian Peng, and Ge Liu. Categorical flow matching on statistical manifolds. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=5fybcQZ0g4.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Oscar Davis, Samuel Kessler, Mircea Petrache, Ismail Ilkan Ceylan, Michael M. Bronstein, and Joey Bose. Fisher flow matching for generative modeling over discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=6jOScqwdHU.

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

Justin Deschenaux, Caglar Gulcehre, and Subham Sekhar Sahoo. The diffusion duality, chapter II: ψ-samplers. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=RSIoYWIzaP.

Justin Deschenaux, Lan Tran, and Caglar Gulcehre. Partition generative modeling: Masked modeling without masks. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=vEh1ceS154.

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

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

Ishaan Gulrajani and Tatsunori Hashimoto. Likelihood-based diffusion language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=e2MCL6hObn.

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

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. Elf: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian diffusion models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ecevn9kPm4.

Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 97394–97430. Curran Associates, Inc., 2025. doi: 10.52202/085713-3256. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8c4046a93293442e88f54e19a5849d68-Paper-Conference.pdf.

Mingyu Jo, Jaesik Yoon, Justin Deschenaux, Caglar Gulcehre, and Sungjin Ahn. Loopholing discrete diffusion: Deterministic bypass of the sampling wall. In *International Conference on Learning Representations*, volume 2026, pp. 93890–93918, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DjJmre5IkP.

Jaeyeon Kim, Seunggeun Kim, Taekyun Lee, David Z. Pan, Hyeji Kim, Sham M. Kakade, and Sitan Chen. Fine-tuning masked diffusion for provable self-correction. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=2LRaNghfoM.

Seunggeun Kim, Jaeyeon Kim, Taekyun Lee, Yuyuan Chen, Yilun Du, Sham Kakade, and Sitan Chen. From interface to inference: Eliciting any-order inference from any-order models. *arXiv preprint arXiv:2607.26504*, 2026b.

Brett L. Kotschwar. Hamilton's gradient estimate for the heat kernel on complete manifolds. *Proceedings of the American Mathematical Society*, 135(9):3013–3019, 2007. ISSN 00029939, 10886826. URL http://www.jstor.org/stable/20534916.

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. *arXiv preprint arXiv:2602.16813*, 2026.

David Li, Nikita Gushchin, Dmitry Abulkhanov, Eric Moulines, Ivan Oseledets, Maxim Panov, and Alexander Korotin. IDLM: Inverse-distilled diffusion language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bS8qU2wJxu.

Peter Li and Shing-Tung Yau. On the parabolic kernel of the schrödinger operator. *Acta Mathematica*, 156:153–201, 1986. URL https://api.semanticscholar.org/CorpusID:120354778.

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

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

Bingbin Liu, Sebastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. TinyGSM: achieving >80% on GSM8K with small language models. 2023. URL https://arxiv.org/abs/2312.09241.

Aaron Lou, Minkai Xu, Adam Farris, and Stefano Ermon. Scaling riemannian diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=FLTg8uA5xI.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Seo Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=NKdtztladR.

Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E Peters, and Arman Cohan. Tess: Text-to-text self-conditioned simplex diffusion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2347–2361, 2024.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1999. ISBN 9780471953333. doi: 10.1002/9780470316979.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Scaling up masked diffusion models on text. In *International Conference on Learning Representations*, volume 2025, pp. 82974–82997, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Peter Potaptchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S Albergo. Discrete flow maps. *arXiv preprint arXiv:2604.09784*, 2026.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. CANDI: Hybrid discrete-continuous diffusion models. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=03KQGUvw1e.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Gabriel Raya, Bac Nguyen, Georgios Batzolis, Yuhta Takida, Dejan Stancevic, Naoki Murata, Chieh-Hsin Lai, Yuki Mitsufuji, and Luca Ambrogioni. Noise scheduling as information-guided allocation in diffusion training. *arXiv preprint arXiv:2602.18647*, 2026.

Daan Roos, Oscar Davis, Floor Eijkelboom, Michael M. Bronstein, Max Welling, İsmail İlkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=lK3AWzD7MS.

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

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T Chiu, and Volodymyr Kuleshov. The diffusion duality. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9P9Y8FOSOk.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=xcqSOfHt4g.

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to DNA sequence design. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46495–46513. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/stark24b.html.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, and Rémi Leblond. Self-conditioned embedding diffusion for text generation. 2022. URL https://arxiv.org/abs/2211.04236.

Jaesung Tae, Hamish Ivison, Sachin Kumar, and Arman Cohan. Tess 2: A large-scale generalist diffusion language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21171–21188, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Veličković, Federico Barbero, Christos Perivolaropoulos, Simon Osindero, and Razvan Pascanu. Perplexity cannot always tell right from wrong. *arXiv preprint arXiv:2601.22950*, 2026.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Guanghan Wang, Yair Schiff, Subham Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 147282–147339. Curran Associates, Inc., 2025. doi: 10.52202/085713-4927. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d877ea0dd78bdbe54830670618c1de09-Paper-Conference.pdf.

Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi S. Jaakkola. Restart sampling for improving generative processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=wFuemocyHZ.

Zhihan Yang, Wei Guo, Shuibai Zhang, Subham Sekhar Sahoo, Yongxin Chen, Arash Vahdat, Morteza Mardani, and John Thickstun. Continuous diffusion scales competitively with discrete diffusion for language. *arXiv preprint arXiv:2605.18530*, 2026.

Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Beyond autoregression: Discrete diffusion for complex reasoning and planning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NRYgUzSPZz.

Longxuan Yu, Shaorong Zhang, Yu Fu, Hui Liu, Yue Dong, and Greg Ver Steeg. Revise, don't freeze: Sampler-matched training for self-correcting masked diffusion language models. *arXiv preprint arXiv:2606.01026*, 2026.

De Shuai Zhang. When latent geometry is not enough: Draft-conditioned latent refinement for non-autoregressive text generation. *arXiv preprint arXiv:2605.15557*, 2026.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 63186–63227, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9e3b203e72c4e058de26d02a92a81844-Paper-Conference.pdf.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=PEQFHRUFca.

APPENDIX CONTENTS

A Limitations **18**

B Additional related work **18**

C Poisson-kernel reduction for the heat-vMF comparison **19**

C.1 Heat to Poisson 19

C.2 Poisson to moment-matched vMF 20

C.3 From the two comparisons to score and KL 20

C.4 Finite-dimensional refinement 21

C.5 Scope and marginal score fields 22

D Implementation details **22**

D.1 Heat-kernel tables 23

D.2 Full-vocabulary mixture 23

D.3 Codebook construction 24

D.4 Adaptive schedules and solver grids 24

D.5 Sampling details 25

D.6 Exact ancestral sampling for Heat 25

D.7 Training cost 29

D.8 Sampling latency and throughput 30

E Additional experimental details **30**

E.1 Model and training setup 31

E.2 Sudoku benchmark protocol 32

E.3 Kernel-consistency study 34

E.4 SLERP comparison 39

E.5 Published Sudoku baselines 39

E.6 GSM8K setup and metric 40

E.7 Full GSM8K comparison 41

E.8 S-FLM S-arch q-argmax transfer 43

E.9 GSM8K kernel-consistency study 43

E.10 GSM8K checkpoint sweep 45

E.11 OpenWebText evaluation 46

E.12 Discrete commitment across tasks 50

E.13 Learned codebook geometry 52

E.14 Numerical validation of the heat kernel 55

F Technical estimates for the heat-vMF proof **56**

F.1 Coordinate moments and concentration 56

F.2 Proof of the heat-Poisson estimate 57
F.3 Explicit Poisson-vMF calibration 59
F.4 Exponentially weighted likelihood tails 59

G OpenWebText generated samples **60**
G.1 Target entropy $H = 4.5$ nats 60
G.2 Target entropy $H = 5.0$ nats 62
G.3 Target entropy $H = 5.25$ nats 63
G.4 Target entropy $H = 5.5$ nats 65

A LIMITATIONS

Our experiments focus on a set of controlled settings designed to isolate the effects of corruption geometry and inference. The kernel study uses hard Sudoku with a fixed $d = 128$ codebook and 20k training updates, providing a focused setting for studying categorical shape effects. The text mixture experiment uses a single-seed 25k-step fine-tuning run rather than from-scratch training. Benchmark error bars reflect sampler variation on fixed checkpoints, while training-seed variation is reported for selected kernel experiments. The crossed evaluations further suggest that the observed gating behavior is encoded in the mixture-trained weights, although a more detailed study of its calibration remains open.

The theoretical analysis focuses on the critical-noise asymptotic regime. While the conditional cosine quantile map is exact, converting it to a spherical displacement and averaging under the posterior introduces a first-order approximation. At larger GSM8K budgets, low-temperature MDLM and Duo remain strong baselines, while SBD shows smaller gains from deterministic token selection. Our restart-size (Δ) analyses also target representative settings: the GSM8K sweep uses a fixed NFE budget, and the Sudoku sweep is performed on the tuning split.

B ADDITIONAL RELATED WORK

Geometric and continuous models. Several works extend diffusion and flow-based generative modeling to non-Euclidean spaces. Riemannian score models use heat-kernel corruption on manifolds (Bortoli et al., 2022; Huang et al., 2022). RDLM is a continuous language diffusion model on a hyperspherical statistical manifold and uses Brownian motion within logarithm-bridge processes (Jo & Hwang, 2025). In contrast, SBD directly uses the spherical heat semigroup as the forward corruption process for token embeddings.

Spherical flows are especially close to our setting. They use vMF paths and derive the corresponding dynamics through a radial continuity equation and a one-dimensional transport construction (Chemseddine et al., 2026b). Our quantile-map construction follows the same radial view, but applies it to more general isotropic spherical kernels, including the bimodal mixtures used in our experiments. Related ideas also appear in work on learned one-dimensional processes, flow-map matching, and consistency models (Chemseddine et al., 2026a; Boffi et al., 2025; Song et al., 2023).

Other continuous language models study different ways of defining the noise process or the representation space. CDCD uses time warping to distribute training effort across noise levels, while LangFlow studies information-uniform schedules and an ODE-based NLL bound (Dieleman et al., 2022; Chen et al., 2026). Continuous and flow-based language models also include TESS, TESS 2, self-conditioned embedding diffusion, latent language diffusion, Categorical Flow Maps, Discrete Flow Maps, and ELF (Mahabadi et al., 2024; Tae et al., 2025; Strudel et al., 2022; Lovelace et al., 2023; Roos et al., 2026; Potaptchik et al., 2026; Hu et al., 2026). Fisher Flow Matching and Statistical Flow Matching instead use Fisher–Rao geometry for categorical distributions, while SSD-LM and Dirichlet Flow Matching use simplex- or Dirichlet-valued representations (Davis et al., 2024; Cheng et al., 2024; Han et al., 2023; Stark et al., 2024). Our backbone follows DiT and the text-diffusion architecture used by SEDD. Analog Bits is an earlier example of self-conditioning for diffusion over discrete data (Peebles & Xie, 2023; Lou et al., 2024; Chen et al., 2023).

Discrete refinement and memory. SBD is also related to discrete diffusion and iterative token-refinement methods. These include D3PM, continuous-time discrete denoising, SEDD, MD4, MDLM, and scaled masked language models (Austin et al., 2021; Campbell et al., 2022; Lou et al., 2024; Shi et al., 2024; Sahoo et al., 2024; Nie et al., 2025). Partition Generative Modeling is a related any-order approach (Deschenaux et al., 2026b). Confidence-based token ordering has also been explored in MaskGIT, reparameterized discrete diffusion, and adaptive token-ordering methods (Chang et al., 2022; Zheng et al., 2024; Ye et al., 2025; Kim et al., 2025). For Sudoku, recent work studies self-correction and entropy-bounded unmasking (Kim et al., 2026a; Ben-Hamu et al., 2025).

Several methods repeatedly move between continuous or noisy states and token predictions. Diffusion-LM projects embedding trajectories back toward token embeddings, FMLM+ accepts endpoints produced by learned flow maps, and ReMDM alternates between token predictions and

remasking. Restart and revision methods similarly modify how information from earlier sampler states is retained (Li et al., 2022; Agarwal et al., 2026; Sahoo et al., 2024; Xu et al., 2023; Wang et al., 2025; Yu et al., 2026). Our Heat bridge is useful for studying this effect because it lets us vary state memory while keeping the token-selection rule fixed.

This is different from carrying denoiser features across sampling steps, as in Loopholing (Jo et al., 2026). Fresh-noise interventions are also related to stochastic churn in EDM (Karras et al., 2022). FMLM+ reports that reusing the initial noise when re-noising cells lowers GSM8K accuracy from 19.0% to 14.4% at 32 NFEs (Agarwal et al., 2026). Our bridge experiments separate this noise-persistence effect from persistence of the previously selected token.

Finally, sampler comparisons depend strongly on the evaluation regime. Generative perplexity can reward low-diversity samples: time-agnostic analyses of masked models trace misleadingly low generative perplexity to an effective temperature reduction caused by inaccurate categorical sampling (Zheng et al., 2025), and generative frontiers make the trade-off explicit by treating generative perplexity and entropy as two components of a divergence to a reference distribution (Pynadath et al., 2026b). Perplexity can also assign low values to incorrect outputs and need not select the more accurate model (Veličković et al., 2026). Verifiable tasks such as Sudoku and GSM8K sidestep both issues by providing direct correctness signals, complementing entropy-aware evaluation of unconditional text. The Diffusion Duality series similarly studies the interaction between training objectives and inference procedures for uniform-state diffusion, including few-step generation and predictor–corrector sampling (Sahoo et al., 2025; Deschenaux et al., 2026a). These works motivate our emphasis on treating the forward process and the inference rule as separate design choices rather than attributing generation behavior to either component in isolation.

C Poisson-kernel reduction for the heat-vMF comparison

This appendix gives the analytic reduction behind Proposition 1. Throughout, we fix $\xi \in \mathbb{R}$ and write

$$t_d = \frac{\log d + \xi}{d-1}, \qquad r_d = \frac{d-1}{2} t_d, \qquad m = e^{-\xi/2}, \qquad U = \sqrt{d}\, C, \qquad \bar{t}_d = t_d - \frac{2}{d}. \tag{11}$$

Thus $q_d = e^{-r_d} = m/\sqrt{d}$ and $\mathrm{He}_2(x) = x^2 - 1$. All constants below may depend on the fixed operating point ξ, but not on d. Densities are taken relative to the uniform measure on $\mathbb{S}^{d-1}$, or equivalently to the uniform coordinate law of U.

The proof has two comparison steps: heat $\to$ Poisson $\to$ moment-matched vMF. Both leave the same quadratic Hermite mode, so their coefficients can be added directly. We then obtain KL from the variance of the centered log-likelihood ratio. Appendix F gives the differentiated expansions and tail estimates used in these steps.

C.1 Heat to Poisson

Let $P_{n,d}$ be the normalized zonal Gegenbauer polynomial of degree n and $h_{n,d}$ its multiplicity. For $D = q\partial_q$, the spherical Poisson kernel is

$$P_{q,d}(u) = \frac{1-q^2}{(1 - 2qu/\sqrt{d} + q^2)^{d/2}} = \sum_{n \geq 0} h_{n,d} q^n P_{n,d}(u/\sqrt{d}). \tag{12}$$

The heat coefficient at intrinsic time t_d factors exactly as

$$\exp\left[-\frac{1}{2} n(n+d-2) t_d\right] = q_d^n \exp\left[-\frac{1}{2} n(n-1) t_d\right]. \tag{13}$$

Consequently,

$$K^H_{d,r_d}(u/\sqrt{d}) = \exp\left[-\frac{t_d}{2} D(D-1)\right] P_{q,d}(u)\Big|_{q=q_d}. \tag{14}$$

Equation 14 is an identity of the convergent zonal expansions; no asymptotic approximation has yet been made.

Lemma 1 (Central C^1 heat-Poisson estimate)**.** *Fix $\xi \in \mathbb{R}$. There are remainders R_d^H and S_d^H such that, under the heat law of U,*

$$\log \frac{K_{d,r_d}^H(U/\sqrt{d})}{P_{q_d,d}(U)} = -\frac{m^2 t_d}{2} \operatorname{He}_2(U-m) + R_d^H(U), \tag{15}$$

$$\partial_u \log \frac{K_{d,r_d}^H(u/\sqrt{d})}{P_{q_d,d}(u)} = -m^2 t_d (u-m) + S_d^H(u), \tag{16}$$

with

$$\|R_d^H - \mathbb{E}_H R_d^H\|_{L^2(H)} + \|R_d^H - \mathbb{E}_H R_d^H\|_{L^3(H)} = O(t_d^2 + t_d/d), \tag{17}$$

$$\|S_d^H\|_{L^2(H)} = O(t_d^2 + t_d/d). \tag{18}$$

The corresponding uniform C^1 estimate on a growing central window is given in Equation 76.

Proof outline. Equation 14 isolates the small heat perturbation of an explicit Poisson kernel. Its first-order term is $-t_d D(D-1)P/2$, and $D(D-1)P/P = m^2 \operatorname{He}_2(u-m) + O((1+|u|^4)/d)$. An exact complex-Gaussian representation controls the Taylor remainder before and after one u derivative. Sub-Gaussian coordinate tails and global heat-kernel bounds then give the displayed heat-weighted norms. The full proof is in Appendix F.2; this argument does not differentiate an uncontrolled density approximation. □

C.2 Poisson to moment-matched vMF

Lemma 2 (Moment-matched Poisson-vMF expansion)**.** *At $q_d = m/\sqrt{d}$, on each growing central window $|u-m| \le A\sqrt{\log d}$ with fixed A, the log-density ratio and its derivative satisfy*

$$\log \frac{P_{q_d,d}(u)}{K_{d,\kappa_d^\star}^V(u/\sqrt{d})} = \frac{m^2}{d} \operatorname{He}_2(u-m) + O\left(\frac{1+|u|^6}{d^2}\right), \tag{19}$$

$$\partial_u \log \frac{P_{q_d,d}(u)}{K_{d,\kappa_d^\star}^V(u/\sqrt{d})} = \frac{2m^2}{d}(u-m) + O\left(\frac{1+|u|^5}{d^2}\right). \tag{20}$$

The density remainder is $O(d^{-2})$ in $L^2(H)$ and $L^3(H)$, and the derivative remainder is $O(d^{-2})$ in $L^2(H)$.

Proof. The vMF coordinate law is an exponential tilt of the uniform spherical coordinate law. Its log-partition function satisfies $\psi_d(\lambda) = \lambda^2/2 - \lambda^4/(4d) + O(d^{-2})$. Matching the heat mean, $\psi_d'(\lambda_d^\star) = m$, gives $\lambda_d^\star = m + m^3/d + O(d^{-2})$. Expanding the explicit Poisson density and this matched vMF density, then subtracting, leaves $m^2 \operatorname{He}_2(u-m)/d$. Differentiating the same explicit formulas gives the score term. Appendix F.3 records the expansions and the tail argument that makes their remainders valid in the stated norms. □

C.3 From the two comparisons to score and KL

Proof of Proposition 1. Combining Lemma 1 with Lemma 2 gives

$$L_d(U) := \log \frac{K_{d,r_d}^H(U/\sqrt{d})}{K_{d,\kappa_d^\star}^V(U/\sqrt{d})} = -\frac{m^2 \bar{t}_d}{2} \operatorname{He}_2(U-m) + \rho_d(U), \tag{21}$$

where

$$\|\rho_d - \mathbb{E}_H \rho_d\|_{L^2(H)} + \|\rho_d - \mathbb{E}_H \rho_d\|_{L^3(H)} = O(t_d^2 + t_d/d + d^{-2}). \tag{22}$$

The score expansion is obtained independently from the differentiated estimate:

$$\frac{a_{d,r_d}^H(U/\sqrt{d}) - \kappa_d^\star}{\sqrt{d}} = -m^2 \bar{t}_d (U-m) + \sigma_d(U), \qquad \|\sigma_d\|_{L^2(H)} = O(t_d^2 + t_d/d + d^{-2}). \tag{23}$$

Here we used $\partial_u \log K(u/\sqrt{d}) = a(u/\sqrt{d})/\sqrt{d}$. Equation 99 then gives

$$\frac{a^H_{d,r_d}(U/\sqrt{d}) - \kappa^\star_d}{\kappa^\star_d} = -m\bar{t}_d(U-m) + O_{L^2(H)}(t_d^2 + t_d/d + d^{-2}). \tag{24}$$

The explicit $2/d$ term is lower order than t_d, but it determines the leading finite-dimensional correction along the critical path. Absorbing this term into the remainder gives the simpler likelihood and score expansions in Proposition 1.

Let $X_d = L_d - \mathbb{E}_H L_d$. Define the centered quadratic mode and residual by

$$H_{2,d} = \mathrm{He}_2(U-m) - \mathbb{E}_H\,\mathrm{He}_2(U-m), \qquad \eta_d = \rho_d - \mathbb{E}_H \rho_d. \tag{25}$$

Then

$$X_d = -\frac{m^2\bar{t}_d}{2} H_{2,d} + \eta_d. \tag{26}$$

Equations 22 and 72 give

$$\mathrm{Var}_H(X_d) = \frac{m^4}{2}\bar{t}_d^2 + O(t_d^3 + t_d^2/d). \tag{27}$$

Because $L_d = \log(p_H/p_V)$,

$$\mathbb{E}_H[e^{-L_d}] = 1, \qquad D_{\mathrm{KL}}(H\|V) = \log \mathbb{E}_H[e^{-X_d}]. \tag{28}$$

The uniform Taylor expansion on $G_d = \{|U-m| \le A\sqrt{\log d}\}$, together with the exponentially weighted tail control in Appendix F.4, therefore yields

$$D_{\mathrm{KL}}(H_{d,r_d}\|V_{d,\kappa^\star_d}) = \frac{m^4}{4}\left(t_d - \frac{2}{d}\right)^2 + O(t_d^3 + t_d^2/d). \tag{29}$$

In particular,

$$D_{\mathrm{KL}}(H_{d,r_d}\|V_{d,\kappa^\star_d}) = \frac{m^4}{4} t_d^2\left[1 + O((\log d)^{-1})\right], \tag{30}$$

which proves Equation 10. The moment recursion in Appendix F.1 also gives $\mathbb{E}_H[(U-m)^2] = 1 + O(t_d + d^{-1})$. Taking the $L^2(H)$ norm in Equation 24 therefore gives the RMS relative-score error $mt_d(1+o(1))$. □

Takeaway. Moment matching removes the leading linear mismatch. The remaining quadratic mode gives an $O(t_d)$ relative score error and an $O(t_d^2)$ conditional KL. The score claim follows from a differentiated estimate, while the KL claim follows from a likelihood-weighted tail bound beyond the Gaussian limit.

C.4 Finite-dimensional refinement

Combining Equation 24 with the second-moment estimate used above, and taking the fourth root in Equation 29, gives the finite-dimensional correction

$$\frac{\widehat{m}_{\mathrm{score}}}{m} = 1 - \frac{2}{dt_d} + o((\log d)^{-1}), \tag{31}$$

$$\frac{\widehat{m}_{\mathrm{KL}}}{m} = \sqrt{1 - \frac{2}{dt_d}} + o((\log d)^{-1}), \tag{32}$$

where $\widehat{m}_{\mathrm{score}}$ is recovered from the RMS relative-score error divided by t_d (equivalently, from the Gaussian-calibrated $\mathrm{MAD}/[t_d\sqrt{2/\pi}]$), and $\widehat{m}_{\mathrm{KL}} = (4D_{\mathrm{KL}}/t_d^2)^{1/4}$. Since $dt_d = \log d + \xi + o(1)$, the predicted leading corrections are respectively $-2/(\log d + \xi)$ and $-1/(\log d + \xi)$. These constants follow from the explicit $2/d$ term; the numerical sweep supports them but is not used in the proof.

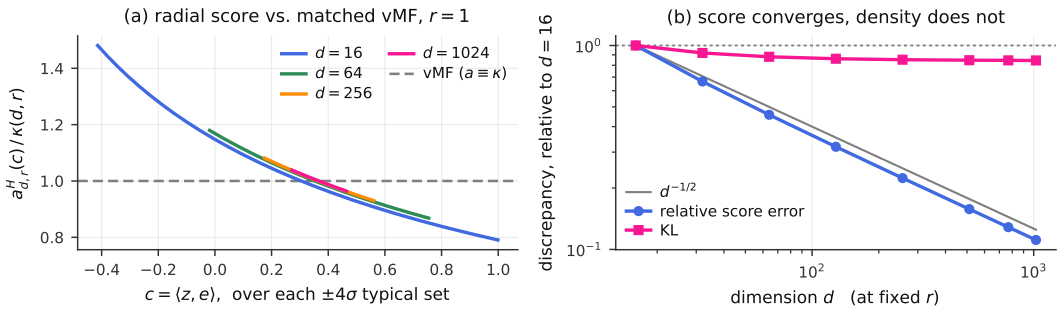

Figure 5: Fixed-noise heat-vMF diagnostic at $r = 1$, which is distinct from the critical scaling in Proposition 1. **Left:** at fixed r, the moment-matched vMF radial score becomes a better local approximation to the heat score as dimension increases. **Right:** both discrepancies are normalized by their values at $d = 16$. The RMS relative radial-score error decreases approximately as $d^{-1/2}$, while the conditional channel KL remains nearly constant. This does not contradict Equation 29: at fixed r the retained cosine is constant, whereas the proposition follows $r_d = (\log d + \xi)/2$, for which the retained cosine is $m/\sqrt{d}$.

C.5 Scope and marginal score fields

For a codebook $\{\mathbf{e}_x\}$, the heat and vMF marginal scores can be written as

$$s_H(\mathbf{z}) = \sum_x P_H(x \mid \mathbf{z}) a_H(c_x)(\mathbf{e}_x - c_x \mathbf{z}), \tag{33}$$

$$s_V(\mathbf{z}) = \kappa \sum_x P_V(x \mid \mathbf{z})(\mathbf{e}_x - c_x \mathbf{z}), \tag{34}$$

where $c_x = \langle \mathbf{z}, \mathbf{e}_x \rangle$. Since $\|\mathbf{e}_x - c_x \mathbf{z}\| \le 1$,

$$\|s_H(\mathbf{z}) - s_V(\mathbf{z})\| \le \mathbb{E}_{P_H(x|\mathbf{z})} |a_H(c_x) - \kappa| + \kappa \|P_H(\cdot \mid \mathbf{z}) - P_V(\cdot \mid \mathbf{z})\|_1. \tag{35}$$

Rotational symmetry reduces the expectation in the first term to the single-channel heat law. After averaging over the heat marginal, the KL chain rule bounds the posterior KL by the single-channel KL. Pinsker's inequality and Jensen's inequality therefore give

$$\frac{1}{\kappa} \mathbb{E} \|s_H(\mathbf{Z}) - s_V(\mathbf{Z})\| \le \mathbb{E}_H \left| \frac{a_H(C)}{\kappa} - 1 \right| + \sqrt{2 D_{\mathrm{KL}}(H \| V_\kappa)}. \tag{36}$$

At $\kappa = \kappa^\star_d$, both terms are $O(t_d)$ under Lemma 1.

For context, Figure 5 shows that the fixed-radius regime behaves differently from the critical path analyzed above.

Takeaway. The critical-scale comparison also controls the exact marginal score field for a fixed codebook. It does not guarantee a learned posterior or better generation quality. The fixed-r diagnostic shows improving score agreement without a similar decrease in KL; the theorem applies only to the critical noise path.

Scope of the critical-path result. Proposition 1 establishes the approximation at $r_d = (\log d + \xi)/2$, corresponding to the critical regime of the sampling trajectory. At $d = 128$ and $\xi = 0$, $r_d \approx 2.43$, near the high-noise edge of the informative Heat/vMF band in Table 25; the result is therefore local. Up to a time change, Heat is the isotropic diffusion for which the standard Brownian score-reversal dynamics are exact; other kernel shapes require compatible transport.

D Implementation details

Why Brownian motion is the canonical isotropic diffusion. For completeness, consider a non-degenerate, time-homogeneous continuous diffusion on $\mathbb{S}^{d-1}$, with $d \ge 3$, whose law is invariant

under the action of $\mathrm{SO}(d)$. Its generator has the local form

$$\mathcal{L}f = b \cdot \nabla f + \frac{1}{2}\,\mathrm{tr}\big(A\nabla^2 f\big)\,, \tag{37}$$

where b is a tangent vector field and A is a positive-definite symmetric tensor on the tangent bundle. Rotation invariance forces b to be an $\mathrm{SO}(d)$-invariant vector field. At any point, invariance under the isotropy subgroup $\mathrm{SO}(d-1)$ therefore gives $b = 0$. The same isotropy group acts irreducibly on the tangent space, so A must be a scalar multiple of the Riemannian metric. Transitivity of $\mathrm{SO}(d)$ makes this scalar constant over the sphere. Hence

$$\mathcal{L} = c\,\Delta_{\mathbb{S}^{d-1}}, \qquad c > 0, \tag{38}$$

which is spherical Brownian motion up to a rescaling of time. Its transition kernels consequently form the heat semigroup.

Table 4: Comparison with S-FLM (Deschenaux & Gulcehre, 2026) and vMF spherical flows (Chemseddine et al., 2026b). Brownian motion defines a Markov heat semigroup, while the other methods specify a path or concentration schedule.

Method	Forward family	Noise parameterization	Semigroup / path
S-FLM	Geodesic interpolation	Interpolation schedule α_t	Designed path
vMF	vMF family	Concentration κ_t	Chosen concentration schedule
Ours	Spherical heat kernel	Retained-cosine clock $r = -\log \rho$	Heat semigroup

D.1 Heat-kernel tables

The score in Equation 5 requires $a^H_{d,r}(c) = \partial_c \log K^H_{d,r}(c)$. At intrinsic heat time τ, the kernel has the Gegenbauer expansion

$$\overline{K}^H_{d,\tau}(c) \propto \sum_{j\geq 0} m_{j,d} e^{-\lambda_{j,d}\tau/2} P^{(d)}_j(c), \qquad \lambda_{j,d} = j(j+d-2), \tag{39}$$

where $m_{j,d}$ is the dimension of the degree-j spherical-harmonic space. We scale terms by the largest term while accumulating the series; the relative size of the final sum then measures numerical cancellation.

The series is not reliable over the full domain. At $d = 128$, it suffers severe cancellation around $r \in [0.4, 1.8]$ in the region containing most probability mass, while the small-time approximation is already inaccurate. Splicing the two approximations produces errors of up to 8% under the identity

$$\mathbb{E}[a^H_{d,r}(C)(1-C^2)] = (d-1)\mathbb{E}[C]. \tag{40}$$

We therefore solve the heat equation. For $f(c,\tau) = \log \overline{K}^H_{d,\tau}(c)$,

$$\partial_\tau f = \frac{1}{2}\left[(1-c^2)(f_{cc} + f_c^2) - (d-1)c f_c\right]. \tag{41}$$

The diffusion coefficient vanishes at $c = \pm 1$, so both endpoints are outflow boundaries. Working with the log density avoids underflow. The solver starts at $r = 0.05$, where the small-time approximation has a Stein residual of 8×10^{-5}, i.e., a relative violation of 8×10^{-5} in Equation 40, which the exact heat kernel satisfies. It stores a snapshot at each required radius.

At runtime, $\log a^H_{d,r}(c)$ is bilinearly interpolated from a float32 grid uniform in $(\log r, c)$. The same table supplies the radial marginal for direct forward sampling. We check agreement with the spectral series where it is reliable, Equation 40, unit mass, and insensitivity to the solver grid.

D.2 Full-vocabulary mixture

For N positions, Equation 5 uses one $[N \times d][d \times |\mathcal{V}|]$ product for all cosines, an elementwise lookup for the radial factors, and one $[N \times |\mathcal{V}|][|\mathcal{V}| \times d]$ product for the weighted codewords. We evaluate the exact vocabulary sum without top-k truncation. The contraction is split into chunks of 2,048 codewords and performed in two passes, first for the log normalizer and then for the weighted sum, so no additional dense vocabulary-sized tensor is retained beyond the logits.

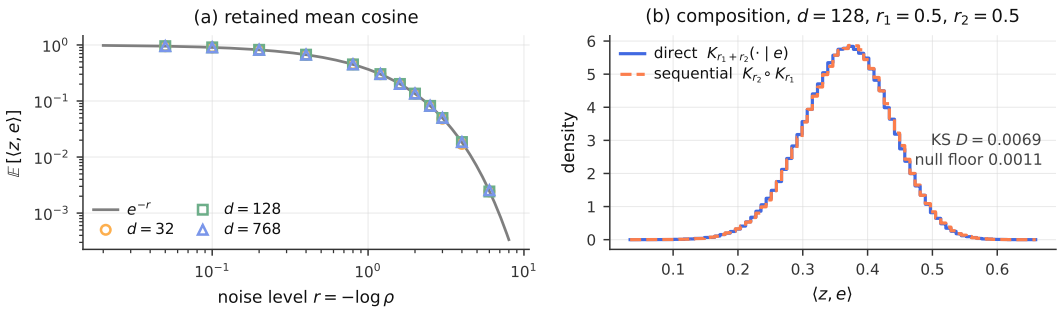

Figure 6: Numerical validation of the Brownian clock and forward sampler. **(a)** For $d \in \{32, 128, 768\}$, forward samples recover $\mathbb{E}\langle \mathbf{Z}_r, \mathbf{e}\rangle = e^{-r}$ with relative error from 0.002% to 0.1% wherever the target is above the Monte Carlo resolution. Because $r = -\log \rho$, this checks the heat tables and sampler. **(b)** At $d = 128$, direct corruption at $r_1 + r_2$ and sequential corruption at $(r_1, r_2) = (0.5, 0.5)$ differ by KS distance $D = 0.0069$, compared with a null floor of 0.0011 from two independent direct draws. Across all tested dimensions and seven radius pairs, $D \leq 0.023$. The exact Brownian semigroup identity is therefore only approximate in the numerical sampler.

D.3 Codebook construction

Frozen and learned codebooks share a max-min initialization. Starting from Gaussian vectors projected to the sphere, we minimize

$$\mathcal{L}_{\text{code}} = \mathbb{E}_v \left[\beta \operatorname{logsumexp}_{w \neq v} \left(\frac{\langle \mathbf{e}_v, \mathbf{e}_w \rangle}{\beta} \right) \right], \tag{42}$$

annealing β from 0.08 to 0.02 and renormalizing after every update.

For learned codebooks, we optionally use

$$\mathcal{R} = \lambda_{\text{anchor}} \sum_v \|\mathbf{e}_v - \mathbf{a}_v\|^2 \tag{43}$$

$$+ \lambda_{\text{sep}} \sum_v \frac{\operatorname{softplus}(\gamma[\cos(\mathbf{e}_v, \mathrm{NN}(v)) - \mu])}{\gamma}. \tag{44}$$

The free configuration uses $\lambda_{\text{anchor}} = 0$, $\lambda_{\text{sep}} = 1$, $\mu = 0.7$, and $\gamma = 20$. Here μ is a cosine threshold, equivalent to a nearest-neighbor angle of $\arccos(0.7) \simeq 0.7954$ radians. The separation term compares a random subset of rows with the complete codebook, and the codebook learning rate is one tenth of the backbone rate. The separate $d = 128$ OWT geometry diagnostic at 150k updates has median displacement 63.7 degrees and median nearest-neighbor angle 0.970 radians (Appendix E.13). Setting $\lambda_{\text{anchor}} = 3 \times 10^{-3}$ keeps them within 8 degrees of the initialization.

D.4 Adaptive schedules and solver grids

One training-noise ablation mixes the information clock $u(r) = H(X_0 \mid Z_r)/H(\pi)$ with a log-uniform radius draw:

$$r \sim 0.75\,\mathcal{U}_{\text{info}} + 0.25\,\mathcal{U}_{\log r}. \tag{45}$$

All positions contribute to the cross-entropy objective; content and padding losses are normalized separately. These variants focus training examples or solver points on noise levels where the posterior changes, without changing the forward process or training objective.

For fixed-entropy training, we measure $H_\theta(r) = \mathbb{E}[\mathbb{H}(q_\theta(\cdot \mid \mathbf{z}, r))]$ once and draw

$$q(r) \propto (1 - \eta) \left| \frac{\mathrm{d}H_\theta}{\mathrm{d}\log r} \right| + \eta \mathcal{U}_{\log r}. \tag{46}$$

For entropy-based solver grids, we invert the CDF of the same density at $S + 1$ equally spaced quantiles and use $\eta = 0.15$. The log-uniform component is needed because the entropy curve is nearly flat at low noise.

For online adaptation, we keep a rolling buffer of $(\log r, \ell)$ pairs: 8,192 for Sudoku, 6,400 for TinyGSM, and 3,200 for OWT. This loss-slope allocation follows S-FLM's adaptive schedule, itself inspired by InfoNoise (Deschenaux & Gulcehre, 2026; Raya et al., 2026). After a 1,000-step warm-up, every 500 steps we fit a regularized cubic smoothing spline with 20 knots and update

$$q(\log r) \propto \max\left(\frac{\mathrm{d}\ell}{\mathrm{d}\log r}, 0\right) + \varepsilon. \tag{47}$$

A uniform floor preserves coverage (mixing weight 10^{-3} for language), and both the buffer and discretized CDF are checkpointed.

For the same vocabulary, the informative interval $u \in [0.1, 0.9]$ occupies 13.9% of the log-radius range at $d = 128$ and 9.3% at $d = 768$. A log-uniform draw therefore allocates relatively fewer examples to informative levels as the dimension grows.

D.5 Sampling details

For a decreasing grid $r_0 > \cdots > r_S$, let $h = r_k - r_{k+1}$ and $g^2 = 2/(d-1)$. The spherical reverse-SDE and probability-flow ODE updates are

$$\mathbf{z} \leftarrow \mathrm{Exp}_{\mathbf{z}}\Big(g^2 s_\theta(\mathbf{z}, r_k)h + g\sqrt{h}\,\boldsymbol{\xi}\Big), \qquad \text{(SDE)}, \tag{48}$$

$$\mathbf{z} \leftarrow \mathrm{Exp}_{\mathbf{z}}\big(\tfrac{1}{2}g^2 s_\theta(\mathbf{z}, r_k)h\big), \qquad \text{(ODE)}, \tag{49}$$

where $\boldsymbol{\xi}$ is isotropic tangent noise and the exponential map keeps every iterate on the sphere.

Prompt positions are fixed at initialization and reset to their clean codewords after every update. Plain q-argmax independently draws fresh heat-kernel noise around each selected codeword at every step; it does not reuse the OWT Gumbel tensor or evaluate the vocabulary score mixture. A first-order predictor with c corrector steps costs $1 + c$ network evaluations per step, Heun's method costs two, and q-argmax costs one. We distinguish the sampler budget B from the total NFE: nearest-codeword or direct token readout uses B calls, while a separate posterior decode uses $B+1$. A Brownian Heun run takes $B/2$ solver steps; q-argmax, the bridge, and quantile transport take B steps. Diagnostic comparisons that hold the grid rather than the call count fixed are identified separately.

D.6 Exact ancestral sampling for Heat

The bridge in Equation 8 follows from Bayes' rule and the Chapman–Kolmogorov identity. For a clean token x, the joint density of a bridge state $\mathbf{z}_s$ and an observed state $\mathbf{y} = \mathbf{z}_R$ is $K^H_{d,s}(\mathbf{z}_s \mid \mathbf{e}_x)K^H_{d,R-s}(\mathbf{y} \mid \mathbf{z}_s)$; division by the marginal $K^H_{d,R}(\mathbf{y} \mid \mathbf{e}_x)$ normalizes it. Drawing $x \sim p(x \mid \mathbf{y}, R)$ and then drawing from the bridge is an exact reverse step when the posterior is exact. In the numerical implementation, posterior approximation and heat-table interpolation are the sources of error.

Two-coordinate bridge sampler. Let $\eta = \langle \mathbf{e}_x, \mathbf{y}\rangle$, $c_1 = \langle \mathbf{z}_s, \mathbf{e}_x\rangle$, and $c_2 = \langle \mathbf{z}_s, \mathbf{y}\rangle$. The projection of a uniform point on $\mathbb{S}^{d-1}$ onto the span of $\mathbf{e}_x$ and $\mathbf{y}$ has density, up to normalization,

$$J_d(c_1, c_2 \mid \eta) \propto \frac{1}{\sqrt{1-\eta^2}}\left[1 - \frac{c_1^2 - 2\eta c_1 c_2 + c_2^2}{1-\eta^2}\right]_+^{(d-4)/2}. \tag{50}$$

The bridge-coordinate density is proportional to $K^H_{d,s}(c_1)K^H_{d,R-s}(c_2)J_d(c_1, c_2 \mid \eta)$. After drawing (c_1, c_2), we choose a uniform direction in the orthogonal complement and reconstruct $\mathbf{z}_s$. This includes the sphere's volume factor and reduces the bridge draw to a two-dimensional numerical sampling problem.

Restart and comparison kernels. In Equation 9, the posterior is evaluated once at the current state; the selected token is held fixed while the state is re-noised by Δ and bridged back. Heat's semigroup means the restarted state has the correct conditional marginal at radius $r_k + \Delta$. At $\Delta = 0$ this is the ancestral bridge. As Δ grows, its retained cosine $e^{-\Delta}$ decays; when the restart state

is nearly uniform, bridging back approaches fresh heat re-noising of the selected codeword, the memoryless q-argmax update. The separate token rule is argmax or sampling from q_θ at $T = 1$. Both controls add no network calls.

This construction depends on a tractable Markov transition family, not on isotropy alone. Spherical Brownian motion is the canonical isotropic continuous diffusion up to time scaling, as shown above. Among the spherical language-model paths compared here, SLERP is deterministic conditional on the clean token, so its bridge is a point mass. vMF distributions are not closed under composition; although vMF marginals can be generated by transport ODEs or stochastic dynamics, the former is deterministic and the latter does not provide a tractable transition density for an exact bridge. Other semigroup families or couplings may permit related bridge constructions. [1]

The bridge uses the heat tables described above. For radii below 0.05, we use the small-time approximation with Stein residual 8×10^{-5}; for larger radii, the tables are obtained by solving the log heat equation. The tabulated Chapman–Kolmogorov identity is satisfied to about 10^{-4} in log-kernel values, and the conditional mean cosine follows e^{-h} for an increment h. These density-level checks are distinct from Monte Carlo composition of the numerical forward sampler, whose KS errors are reported in Figure 6.

Heat configurations for the restart experiments. Table 5 records the checkpoints and evaluation protocols used below.

Table 5: Model and evaluation configuration for the Heat restart experiments. Sudoku uses $B = 128$ sampler calls; posterior decoding in the bridge comparison gives 129 total NFEs.

Task	Heat model	Evaluation set	Restart protocol
Hard Sudoku	8-layer DiT, width 512; frozen $d = 128$ codebook; 20k updates, batch 256	1,000 tuning and 2,000 disjoint evaluation puzzles	$B = 128$; seed regimes specified per table
GSM8K	Heat-trained online 250k checkpoint; context length 512	All 1,319 test problems	$B = 64$; three paired inference seeds

Sudoku bridge versus q-argmax. Table 6 compares the ancestral bridge with q-argmax on the 2,000-puzzle hard-Sudoku evaluation split at $B = 128$. These are the four-training-seed subset used to compare both gated and ungated variants; the completed five-training-seed memory comparison is in Tables 2 and 10. Posterior decoding adds one final network call, for 129 calls total. In the ungated case, the bridge loses mainly through illegal grids, not by returning legal solutions that violate puzzle clues.

Table 6: Hard-Sudoku exact-match accuracy (%) on the 2,000-puzzle evaluation split. Each sampler uses $B = 128$ sampling calls plus one final posterior-decoding call (129 network calls total). Four paired training seeds, with sampler outputs paired within seed; Δ columns are paired differences from q-argmax. The main-text memory gain uses the completed five-seed comparison instead.

Sampler	Ungated	Δ	Gated (tuned)	Δ
q-argmax	77.1 ± 2.1	–	82.1 ± 1.3	–
Bridge, argmax token	59.4 ± 3.1	-17.8 ± 1.6	79.1 ± 1.9	-3.0 ± 1.1
Bridge, sampled token	60.4 ± 2.2	-16.8 ± 1.3	79.1 ± 1.3	-3.0 ± 0.6
Memoryless, $T = 1$	77.5 ± 2.5	$+0.4 \pm 0.8$ (n.s.)	–	–

The ungated failure breakdown (valid / legal but wrong puzzle / illegal) is 77.1 / 2.6 / 20.2 for q-argmax and 59.4 / 2.6 / 38.1 for the bridge. The same clue-violation rate shows that the bridge's lower exact match comes from illegal grids rather than added clue drift; Table 6 reports the corresponding exact-match rates.

[1] The spherical Poisson kernel also forms a semigroup, but it corresponds to subordinated Brownian motion, a jump process rather than a continuous diffusion.

Restart-size sweep on Sudoku. This family was tuned on a separate 1,000-puzzle split with four sampler seeds, no gate, and $B = 128$. Table 7 shows recovery as added noise erases more state memory; the tuning split is distinct from the evaluation split above.

Table 7: Hard-Sudoku restart sweep on the 1,000-puzzle tuning split at $B = 128$ sampler calls, four sampler seeds, without gating.

Δ	Retained cosine $e^{-\Delta}$	Argmax token	Sampled token ($T = 1$)
0 (bridge)	1.00	59.0	60.0
0.5	0.61	75.9	76.7
1	0.37	76.5	77.8
2.43 ($\frac{1}{2}\log d$)	0.09	76.9	78.0
8 (near-memoryless)	≈ 0	76.6	78.8

Why the ancestral bridge plateaus. The following diagnostics use a separate 1,000-puzzle hard-Sudoku tuning split, three independently trained Heat models (seeds 0–2), and $B = 128$. The bridge diagnostics in Tables 8 and 9 use these three training seeds; the budget sweep in Table 10 instead uses five sampler seeds on 1,000 hard-Sudoku tuning puzzles, and the Sudoku restart-size sweep in Table 7 uses four sampler seeds on the tuning split. These seed regimes and splits are distinct. Table 8 compares each trajectory with a fresh forward sample of the true solution at the same noise level. The measured cosine marginal and codeword margin (i.e. cosine to the nearest codeword minus cosine to the second-nearest codeword, averaged over cells) show no *bridge-specific* shift. Yet the bridge makes fewer wrong-to-right and right-to-wrong changes, halving net revision (Table 9).

Table 8: Hard-Sudoku trajectory diagnostics on 1,000 tuning puzzles at $B = 128$, averaged over the informative noise band and three Heat training seeds. KS uses cosine to the nearest codeword against forward samples of the true solution; the reference codeword margin is 0.115.

Diagnostic	Memoryless	Bridge	Restart $\Delta = 0.5$
KS versus forward cosine marginal	0.004	0.006	0.005
Codeword margin	0.115	0.115	0.116
Calibration error on trajectory states	0.050	0.070	0.053
Cell accuracy after $r \approx 1.9$ (%)	95.0	92.7	94.6

Table 9: Hard-Sudoku memory intervention on the same 1,000 tuning puzzles, three Heat training seeds and $B = 128$. Bridge-fresh retains the previous token choice but redraws the noise. Flip counts are per cell over the informative band; net revision is wrong-to-right minus right-to-wrong.

Diagnostic	Memoryless	Bridge	Bridge-fresh	Restart $\Delta = 0.5$
Valid grids (%)	74.8	58.6	73.1	74.8
Wrong→right	0.291	0.168	1.390	0.273
Right→wrong	0.252	0.148	1.351	0.233
Net revision	0.040	0.020	0.039	0.040
Wrong-cell confidence, trajectory / redrawn	.976/.976	.968/.969	.975/.975	.976/.976
Calibration error, trajectory / redrawn	.052/.052	.068/.067	.055/.055	.052/.052

Bridge-fresh recovers 14.5 of the 16.2 lost points, about 90%, while retaining commitment memory. Its paired accuracy difference from memoryless is -1.70 ± 0.44 points across the three training seeds. Its gross flip count is about five times the memoryless count, but its *net* revision and accuracy are close to memoryless. Re-noising the same committed grid gives almost unchanged confidence and calibration within each sampler. Thus the calibration gap mainly reflects which grids were committed, instead of a different reading of a fixed grid. Confidence is nearly unchanged when evaluated on the same states, suggesting that the calibration gap mainly reflects which grids become committed. The remaining gap measures the commitment-memory cost in this intervention.

Budget scaling on Sudoku and GSM8K. Table 10 uses the 1,000-puzzle Sudoku tuning split and five sampler seeds, unlike the 2,000-puzzle benchmark in Table 19 or the five-training-seed kernel

diagonal in Table 24. Its 78.0% q-argmax result at $B = 180$ therefore need not equal the 77.9% estimate from the latter study; the two use different splits and seed regimes. On Sudoku the bridge is nearly flat while q-argmax gains 12.9 points. GSM8K also shows a growing deficit, but its bridge improves by 3.0 points. For a Heat step of radius h, the old state retains mean cosine e^{-h}; as the grid is refined, successive readings become more correlated. The heuristic sum $\sum_k (1 - e^{-h_k}) \leq \sum_k h_k$ stays bounded by the width of the informative band, rather than growing with NFE. This scaling argument captures the effective number of posterior readings up to approximation. The slight Sudoku bridge increase, $58.8 \to 59.5\%$, is compatible with saturation but does not establish the mechanism on its own.

Table 10: Bridge versus q-argmax across sampler budgets B. Sudoku uses 1,000 hard tuning puzzles and five sampler seeds; GSM8K uses all 1,319 test questions, the Heat 250k checkpoint, three paired inference seeds and common prior draws. Values are mean accuracy (%); differences are bridge minus q-argmax. The last row uses $B = 180$ for Sudoku and $B = 256$ for GSM8K.

B (Sudoku / GSM8K)	Sudoku			GSM8K		
	Bridge	q-argmax	Difference	Bridge	q-argmax	Difference
16	58.8	65.1	-6.4 ± 1.0	14.6	18.2	-3.6 ± 0.3
32	59.1	69.2	-10.1	15.3	20.9	-5.5
64	58.9	73.7	-14.8 ± 0.8	15.5	23.6	-8.1 ± 0.6
128	59.3	77.0	-17.7 ± 1.4	16.6	25.6	-9.0
180 / 256	59.5	78.0	-18.4 ± 1.8	17.6	27.1	-9.5 ± 1.4

The token-selection differences summarized in the main-text bridge table are listed at every evaluated budget in Table 11. Negative values mean that sampling tokens at $T = 1$ is better than argmax. The Sudoku rows use the 2,000-puzzle evaluation split and five training seeds; the GSM8K rows use three paired inference seeds and common prior draws (i.e. the same initial uniform-sphere samples at $r_{\max}$ are reused across compared samplers for each seed (common random numbers), so differences are paired).

Table 11: Paired token-selection effect, measured as q-argmax minus memoryless sampling at $T = 1$ (percentage points; mean $\pm$ sd). Sudoku uses 2,000 evaluation puzzles, five training seeds, and posterior decoding (NFE $= B + 1$). GSM8K uses all 1,319 test questions, three paired inference seeds, and common prior draws. Negative values favor $T = 1$ sampling.

Task (B)	q-argmax $-$ bridge	q-argmax $- T = 1$
Sudoku (16)	$+6.4 \pm 1.0$	-1.3 ± 1.0
Sudoku (64)	$+14.8 \pm 0.8$	-0.6 ± 0.8
Sudoku (128)	$+17.7 \pm 1.4$	-0.5 ± 0.7
Sudoku (180)	$+18.4 \pm 1.8$	-1.4 ± 1.0
GSM8K (16)	$+3.6 \pm 0.3$	$+11.4 \pm 1.0$
GSM8K (64)	$+8.1 \pm 0.6$	$+8.5 \pm 1.3$
GSM8K (256)	$+9.5 \pm 1.4$	$+4.7 \pm 2.0$

The memory channels separate differently on the two tasks. At GSM8K sampler budget $B = 64$, bridge-fresh is worse than the exact bridge (Table 12). A single-seed control that forces the previous token to equal the current choice scores 22.67% against 23.12% for q-argmax on the same seed, and the bridge-fresh run reproduces 14.10%. The argmax changes on about 9.2% of GSM8K positions per step versus about 0.5% of Sudoku cells, with 5.9 versus 0.65 changes per position over a run. Retaining an old token therefore matters more for high-churn GSM8K; the Sudoku intervention isolates the contribution of noise persistence to the observed performance gap of the bridge sampler.

Restart-size sweep on GSM8K. We use the Heat-trained 250k checkpoint and all 1,319 GSM8K test questions, with $B = 64$ and three paired inference seeds. Table 13 shows that memory loss also helps program generation, with slower recovery from partial memory loss than on Sudoku.

Temperature-controlled GSM8K comparisons. Table 14 compares released MDLM and Duo checkpoints with SBD under matched $T = 1$ and low-temperature/deterministic decoding. Base-

Table 12: GSM8K memory intervention at sampler budget $B = 64$ on all 1,319 test questions: one Heat 250k checkpoint, three paired inference seeds and common prior draws. Differences are paired against q-argmax.

Sampler	Accuracy (%)	Difference (points)
q-argmax	23.63	–
Exact bridge	15.49	-8.1 ± 0.6
Bridge-fresh	13.34	-10.3 ± 1.3

Table 13: GSM8K accuracy for the Heat restart sweep at $B = 64$. Results use the 250k Heat checkpoint and three paired inference seeds.

Δ	GSM8K accuracy (%)	Difference from q-argmax (points)
0 (exact ancestral)	15.49	-8.14 ± 0.61
0.5	21.89	-1.74 ± 0.39
8	23.60	-0.03 ± 0.95
∞ (q-argmax)	23.63	–

line values are single runs; SBD means are over three inference seeds. Table 15 gives the longer memoryless $T = 1$ and q-argmax curves used in Figure 3.

Table 14: GSM8K accuracy (%) by NFE and decoding temperature. MDLM and Duo values are single runs; SBD values are means over three inference seeds.

NFE	Duo $T = 1$	MDLM $T = 1$	SBD memoryless $T = 1$	Duo $T = .1$	MDLM $T = .1$	SBD q-argmax
16	10.0	1.8	6.80	22.7	14.3	18.15
32	13.9	5.7	9.70	28.2	23.0	20.85
64	16.6	10.0	15.14	34.3	27.5	23.63
128	15.4	13.5	19.99	32.2	31.5	25.60
256	18.5	15.1	22.39	35.3	33.1	27.12

Table 15: GSM8K accuracy (%) for memoryless SBD at $T = 1$ and SBD q-argmax across the full evaluated ladder. Means are over three inference seeds.

NFE	Memoryless, $T = 1$	q-argmax
1	0.0	0.7
4	1.7	8.3
16	6.8	18.2
64	15.1	23.6
256	22.4	27.1
512	26.2	30.0
1,024	28.8	30.1

D.7 TRAINING COST

Table 16 compares training throughput and total cost for 250k updates on TinyGSM and 1M updates on OpenWebText. The AR, MDLM, Duo, CANDI, FLM, and S-FLM values, including the S-FLM variants, are taken directly from Table 3 of Deschenaux & Gulcehre (2026); we do not remeasure these baselines. Their measurements use 8 NVIDIA H100 GPUs for TinyGSM and 16 for OWT. Our measurements use NVIDIA GH200 GPUs, so this is not a hardware-matched comparison of method efficiency. Baseline totals are copied as reported, instead of being recomputed from their rounded throughput.

Table 16: Training cost on TinyGSM (250k updates) and OpenWebText (1M updates). Baseline values are from Table 3 of Deschenaux & Gulcehre (2026). Our duration and GPU-hours for TinyGSM refer to the complete run. OWT GPU-hours are implied by the reported 4.219 updates/s over 1M updates on 16 GPUs; the full-run wall time was not recorded. Throughput is measured over training segments, not averaged over the full run. Baselines use H100s and our runs use GH200s. NR indicates an unreported full-run duration.

	Updates/s		Duration (h)		GPU-hours	
Model	TinyGSM	OWT	TinyGSM	OWT	TinyGSM	OWT
AR	5.4	6.0	12.8	46.6	102.1	745.7
MDLM	3.0	3.5	22.8	80.3	182.7	1,284.5
Duo	3.0	3.5	22.8	80.3	182.7	1,284.5
CANDI	3.0	2.5	23.5	108.9	188.3	1,742.9
FLM	3.2	3.3	21.7	84.2	173.6	1,346.8
S-FLM	4.2	4.5	16.5	61.7	132.3	987.7
S-FLM + adapt.	4.2	4.5	16.5	61.7	132.3	987.7
S-FLM + adapt. + $\mathbb{S}$-arch	2.6	2.8	26.7	99.2	213.7	1,587.3
Ours	3.465	4.219	20.0	NR	160.3	1,053

D.8 Sampling latency and throughput

Table 17 reports measurements on the same single NVIDIA GH200 GPU, using released baseline checkpoints and our 1M-update checkpoint. All methods generate 1,024-token sequences at $T = 1$ where applicable and 256 actual network evaluations, with the call counts independently verified. Each configuration uses two warmup runs followed by seven timed repeats. Latency is measured at batch size 1 in seconds per sample; throughput is measured at batch size 4 in generated tokens per second. Peak memory is also reported at batch size 4. These values report newly measured inference costs, complementing the training costs reported above.

Table 17: Sampling cost on one NVIDIA GH200 GPU at 256 NFEs, sequence length 1,024, and $T = 1$ where applicable. All configurations use two warmups and seven timed repeats. Latency uses batch size 1; throughput and peak memory use batch size 4.

Method	Latency (s/sample)	Throughput (tokens/s)	Peak memory (GiB)
MDLM	3.40	979	3.73
Duo	3.21	1,259	3.73
S-FLM	3.62	620	5.59
Ours: q-argmax	3.37	1,221	2.31
Ours: Heun SDE	7.54	257	2.73
Ours: OWT hybrid, $k = 1$	7.03	277	5.80

At matched NFE, q-argmax has latency comparable to MDLM and Duo, with throughput close to Duo and the lowest measured peak memory. The Heun SDE and OWT hybrid are slower despite the same number of network calls, showing that matched NFE need not imply matched wall-clock cost. These timings characterize inference efficiency under the evaluated settings, with generation quality assessed separately.

E Additional experimental details

Models, comparisons, and validation checks. We refer to trained models by their configuration. Unless stated otherwise, a *learned codebook* is initialized at the max-min packing (Equation 42) and trained jointly with the backbone, and the *adaptive noise schedule* is the online loss-adaptive r-schedule of Appendix D.4 (Equation 47). A learned codebook is *unanchored* when $\lambda_{\text{anchor}} = 0$ in Equation 44; the anchored variant uses $\lambda_{\text{anchor}} = 3 \times 10^{-3}$ and is otherwise identical (Appendix D.3). The OpenWebText model (Appendix E.11) is trained for 10^6 updates with an unanchored learned codebook ($d = 768$) and the adaptive noise schedule (run

owt_d768_online_free_lr3e4; Table 18). We call a model *matched* when it is sampled with the kernel it was trained on, and *crossed* when it is sampled with a different kernel (Section 4.3; Tables 30 and 38). Paired Heat–vMF comparisons (Tables 24 and 39) use the same seeds, puzzles, and sampler for both training kernels. For the error-compounding analysis on GSM8K (Appendix E.9), we use two denoising probes that involve no sampling. The *code-token probe* measures top-1 accuracy of the Heat posterior on program tokens only (Table 40). The *full-answer probe* measures per-token accuracy over the whole answer region (program and padding) and supplies the per-token rate p for the p^L check (Table 41 and Figure 2). We validate every transport with three checks (Appendix E.3). The *Brownian-compatibility index* (Equation 55, with the L^2 variant in Equation 56) measures how far a kernel family is from being generated by Brownian motion; it is zero for Heat. The *conditional-law check* starts the sampler from pure noise with a single codeword and tests, by a Kolmogorov–Smirnov (KS) test, whether it reproduces the kernel's conditional law K_r. The *intermediate-marginal check* uses an exact posterior on a 10-codeword toy problem and tests the marginals at several radii (KS) and the final token frequencies (total variation).

E.1 Model and training setup

Our Sudoku experiments use an 8-layer Diffusion Transformer (DiT) with hidden width 512 and eight attention heads. GSM8K and OWT use a 12-block spectral DiT with hidden width 768, 12 attention heads, and conditioning dimension 128 (Peebles & Xie, 2023; Lou et al., 2024). The codebook dimension d is separate from the backbone embedding: we compare $d = 128$ and $d = 768$ as specified per task. The posterior network receives the normalized log-radius $\widetilde{r} = (\log r - \log r_{\min})/(\log r_{\max} - \log r_{\min})$. In every training and sampling run, $r_{\min} = 0.02$ and $r_{\max} = 8$, corresponding to retained mean cosines $e^{-0.02} \simeq 0.98$ and $e^{-8} \simeq 3.35 \times 10^{-4}$. The sole exception is the truncated SLERP ablation, which uses $r_{\min} = 0.942$. Unconditioned positions are initialized uniformly on the sphere. The exact Heat marginal at $r = 8$ is not uniform, but its retained mean cosine is only $e^{-8} \simeq 3.35 \times 10^{-4}$ and it is numerically close to stationarity. Uniform initialization therefore introduces a small finite-horizon approximation instead of changing the limiting process. We use AdamW with $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, gradient clipping at 1.0, and a backbone learning rate of 3×10^{-4}. Sudoku uses weight decay 0.01, a one-cycle schedule with 5% warm-up and cosine decay, fp32 training, and no EMA. Language models use zero weight decay, a constant learning rate after 2,500 warm-up updates, bf16 training, and EMA decay 0.9999; language evaluation uses EMA weights. Dropout is 0.1 during training and disabled during evaluation. Table 18 summarizes the task-specific settings. These are our training budgets, not the budgets used to train the released baseline checkpoints.

Table 18: Main-model training configuration. Sequence length includes the prompt in conditional tasks. Language parameter counts separate the backbone from the trainable codebook; all models omit self-conditioning.

Setting	Sudoku	TinyGSM → GSM8K	OpenWebText
Backbone	8 layers, width 512, 8 heads	12 blocks, width 768, 12 heads; conditioning width 128	Same as GSM8K
Parameters	26.2M	130.8M + 37.7M codebook	131.6M + 38.6M codebook
Codebook d	128	768	768
Tokenizer	Sudoku digits	SmolLM-135M; 49,152 + PAD	GPT-2; 50,257 + 1
Sequence length	180	512	1,024, wrapped
Training updates	20k	250k	1M
Global batch	256 (4×64)	512 (8×64)	512 (16×32)
Weight decay	0.01	0	0
Learning rate	3×10^{-4}; one-cycle	3×10^{-4}; constant after warm-up	Same as GSM8K
LR warm-up	5% of updates	2,500 updates	2,500 updates
Precision / EMA	fp32 / none	bf16 / 0.9999	bf16 / 0.9999
Adaptive buffer	8,192	6,400	3,200
Codebook	Frozen max-min	Learned from max-min anchor; LR 3×10^{-5}	Same as GSM8K

Codebooks and noise schedules. The online loss-adaptive schedule starts after 1,000 updates and is refit every 500 updates; language runs mix in a uniform component of weight 10^{-3}. Learned language codebooks are parameterized around a fixed max-min anchor but have no anchor penalty. They use a separation penalty of weight 1.0 and cosine margin 0.7, with a codebook learning rate equal to one tenth of the backbone rate (Appendix D). Frozen codebooks retain their max-min initialization. Learned codebooks are updated jointly with the backbone; their initialization and regularization are described in Appendix D. The main models use the online loss-adaptive schedule above. A separate schedule ablation mixes a uniform information-clock draw with a log-uniform draw, with weights 0.75 and 0.25 (Equation 45). Fixed-entropy and online adaptive schedules instead allocate more training examples to informative noise levels. The fixed-entropy schedule is calibrated once, whereas the online schedule is updated from the recent cross-entropy profile. We distinguish the training noise distribution from the sampling grid: the *entropy + log sampling* ablation uses an entropy training schedule and a log-radius solver grid, while *entropy + entropy sampling* also uses an entropy-based solver grid.

Conditional generation and decoding. The five GSM8K checkpoint ablations and Sudoku q-argmax runs use Equation 7 at every noise level. These runs have no score-update phase, sampler switch, or Gumbel perturbation. Each step selects the most likely clean token and independently re-noises its embedding at the next radius. Prompt positions remain fixed. The final token at each position is the most likely token under the predicted clean-token posterior. Only OWT uses the score/q-argmax hybrid. Argmax makes each token choice deterministic given the current state, but the trajectory remains stochastic because initialization and re-noising are random. Confidence-gated Sudoku sampling also retains accepted tokens and stops re-noising their positions; unfrozen positions continue with the same q-argmax rule. We do not use self-conditioning in our models.

Sampling budget and uncertainty. We count network evaluations (NFEs) as Transformer forward passes during generation. A q-argmax step and a first-order ODE or SDE predictor each use one evaluation; a Heun step uses two. A predictor with c corrector updates uses $1 + c$ evaluations. Kernel-table lookup and heat-kernel re-noising do not call the Transformer and do not add NFEs. NFE therefore measures model calls, not total wall-clock cost. Sudoku benchmark error bars vary five sampler seeds on one trained checkpoint per difficulty, except the low-budget SDE/ODE entries, which use three. Kernel-study error bars vary independent training seeds where stated. GSM8K fine-tune error bars vary three inference seeds on one training seed per arm.

E.2 Sudoku benchmark protocol

We use 48k training and 2k evaluation puzzles per difficulty from the S-FLM generator (Alp, 2024; Deschenaux & Gulcehre, 2026). Easy, medium, and hard puzzles reveal 40, 35, and 30 cells. One frozen max-min $d = 128$ Heat model is trained for 20k updates per difficulty with the online schedule, cross-entropy, global batch 256, and a fixed puzzle prefix. All benchmark entries use dropout-free evaluation and explicit network-call accounting. The information-grid entries use five sampler seeds on the same weights; low-budget SDE/ODE instead use the log-radius grid and three sampler seeds, as indicated in Table 19.

Confidence thresholds are selected separately for each difficulty and seed on 1,000 held-out puzzles from a generator seed disjoint from training and evaluation. Seven candidates span 0.99 to 0.9999999. Posterior decoding uses one last evaluation of $q_\theta(v \mid \mathbf{z}, r_{\min})$. It costs one extra call, so gated rows at $B = 180$ are charged 181 NFEs. It helps gating by 0.08, 0.69, and 0.77 points on easy, medium, and hard puzzles and is neutral for the transport samplers: the largest SDE/ODE gain is 0.27 points and the typical gain is 0.03 points.

Table 19 gives the full sweep. The solver grid is chosen per sampler on held-out puzzles; SDE/ODE use the log-radius grid below $B = 64$. True NFE counts network calls, so $B = 16$ and $B = 32$ correspond to 8 and 16 two-call Heun steps, respectively, while posterior decoding is charged one additional call.

Table 19: Sudoku exact-match accuracy (%). The solver grid is chosen per sampler on held-out puzzles; SDE/ODE use the log-radius grid below $B = 64$. Low-budget SDE/ODE entries average three sampler seeds; all other entries average five sampler seeds on one checkpoint per difficulty. Error bars do not include training variance. Column headings are sampler calls B; gated posterior decoding uses $B + 1$ total NFEs, and the other rows use B.

Difficulty	Sampler / B	16	32	64	128	180
Easy	SDE	$95.32 \pm .26$	$96.93 \pm .28$	$97.20 \pm .18$	$97.32 \pm .33$	$97.10 \pm .15$
	ODE	$96.33 \pm .63$	$96.38 \pm .20$	$96.24 \pm .21$	$96.54 \pm .15$	$96.42 \pm .21$
	q-argmax	$97.87 \pm .18$	$98.24 \pm .10$	$98.50 \pm .08$	$98.68 \pm .11$	$98.69 \pm .07$
	gated q-argmax	$98.85 \pm .11$	$98.94 \pm .09$	$98.96 \pm .13$	$99.14 \pm .11$	$99.11 \pm .12$
Medium	SDE	$84.83 \pm .13$	$88.70 \pm .48$	$89.15 \pm .57$	$89.35 \pm .53$	$89.75 \pm .40$
	ODE	$86.82 \pm .38$	$87.32 \pm .95$	$87.29 \pm .25$	$87.92 \pm .22$	$88.15 \pm .57$
	q-argmax	$91.17 \pm .42$	$92.56 \pm .23$	$93.33 \pm .33$	$94.11 \pm .11$	$94.57 \pm .21$
	gated q-argmax	$94.62 \pm .40$	$95.74 \pm .33$	$96.21 \pm .33$	$96.47 \pm .22$	$96.66 \pm .15$
Hard	SDE	$50.45 \pm .41$	$60.63 \pm .92$	$62.27 \pm .73$	$63.28 \pm .93$	$63.25 \pm .36$
	ODE	$56.02 \pm .75$	$57.95 \pm .74$	$58.50 \pm .88$	$58.72 \pm .51$	$58.24 \pm .82$
	q-argmax	$68.96 \pm .76$	$72.23 \pm .81$	$75.58 \pm .17$	$78.51 \pm .36$	$79.70 \pm .75$
	gated q-argmax	$74.80 \pm .95$	$78.55 \pm .53$	$81.08 \pm .43$	$82.76 \pm .84$	$83.36 \pm .42$

Provenance of hard-Sudoku q-argmax results. Table 20 separates four Heat q-argmax results that should not be treated as repeated measurements of one setting. Only the benchmark row is used in the main Sudoku comparison. The three-seed Heat–vMF comparison in Table 24 uses an identical paired protocol for both training arms, with inference settings matched within each sampler. The Heat arm used 20k updates and the log-radius grid. Its 73.88% estimate is 5.82 points below the benchmark's 79.70%, which used the information grid. This is consistent with the later observed 5.6-point gain from information spacing for hard-Sudoku q-argmax at $B = 180$. The seed regimes also differ: three training seeds here versus five sampler seeds on one checkpoint in the benchmark.

Table 20: Protocol provenance for four Heat q-argmax results (%) on 2,000 hard puzzles at a sampler budget of $B = 180$ calls. Standard deviations reflect the seed type in the last column. Gated runs include a final posterior decode (+1 call).

Study	Accuracy	Training	Grid	Decode	Seeds
Benchmark	$79.70 \pm .75$	online, 20k	information	none, 180 calls	5 sampler; 1 checkpoint
Heat–vMF contrast	$73.88 \pm .41$	20k updates	log-radius	q-argmax	3 paired training seeds
Kernel Heat diagonal	77.9 ± 1.8	online, 20k	information-matched	posterior, 181 calls	5 training seeds
Frozen-codebook ablation	$79.68 \pm .42$	online, 20k	information	posterior, 181 calls	3 sampler; 1 checkpoint

At $B = 64$, both grids are available for continuous transport. Table 21 reports five sampler seeds per grid on the same 2,000 puzzles and one fixed Heat checkpoint per difficulty, with nearest-codeword decoding (64 total NFEs) and dropout disabled. Differences are small and have mixed signs: two favor the information grid and four favor the log grid. The largest is -0.92 points for medium/ODE, compared with a quadrature-combined standard deviation of about 0.54; without paired differences we do not assign a paired significance level. In contrast, the information grid adds 5.6 points to q-argmax on hard Sudoku at $B = 180$. This supports a larger grid effect for categorical refinement than for the transport samplers in this control.

Table 21: NFE-64 transport control with nearest-codeword decoding: exact-match accuracy (%), mean $\pm$ sd over five sampler seeds per grid on one fixed Heat checkpoint per difficulty and 2,000 puzzles.

Difficulty	SDE log	SDE information	ODE log	ODE information
Easy	97.10 ± 0.41	97.20 ± 0.18	96.47 ± 0.23	96.24 ± 0.21
Medium	89.44 ± 0.31	89.15 ± 0.57	88.21 ± 0.48	87.29 ± 0.25
Hard	62.97 ± 0.76	62.27 ± 0.73	58.34 ± 1.02	58.50 ± 0.88

FMLM+ comparison. We evaluate released checkpoints on the same evaluation puzzles with $p = 0.999$, uniform budget, fresh noise, and five sampler seeds.

Table 22: FMLM+ Sudoku exact-match accuracy (%), mean ± standard deviation over five sampler seeds on fixed released checkpoints.

NFE	Easy	Medium	Hard
4	$97.60 \pm .09$	$91.53 \pm .39$	$70.36 \pm .78$
16	$97.87 \pm .27$	$92.97 \pm .54$	$81.35 \pm .55$
64	$97.87 \pm .29$	$92.99 \pm .42$	$81.44 \pm .71$
180	$97.77 \pm .13$	$92.99 \pm .33$	$81.65 \pm .36$

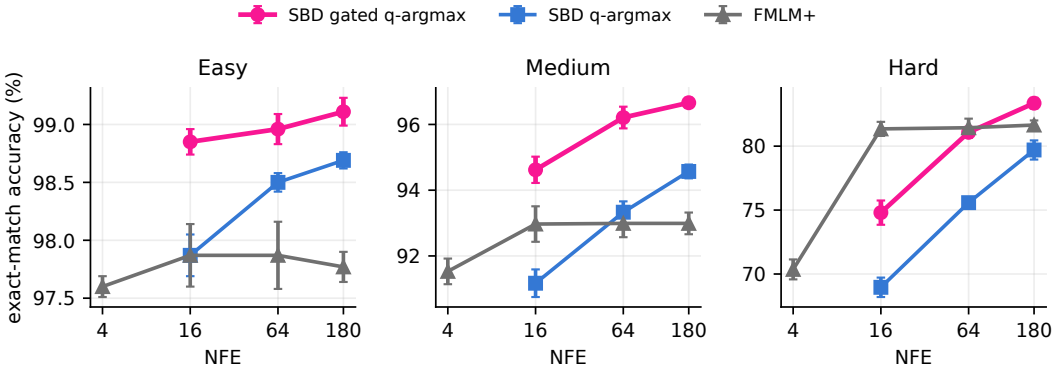

Figure 7: Complete Sudoku accuracy curves for easy, medium, and hard puzzles. Error bars are standard deviations over five sampler seeds on fixed checkpoints, not training seeds. Comparing gated SBD with FMLM+, FMLM+ leads on hard Sudoku at $B = 16$, they are tied within variation at $B = 64$, and SBD leads at $B = 180$. The plotted budget labels are sampler calls B; gated SBD's posterior decode adds one call.

Figure 7 shows the full budget curves behind the fixed-budget comparison.

Failure decomposition. q-argmax produces the most legal grids but can drift from a clue because every position is re-noised. Gating nearly removes this failure but prevents some repairs, while posterior decoding recovers part of the lost legality; the measured breakdown is in Table 23.

Table 23: Hard-Sudoku failure decomposition (%) at $B = 180$ sampler-call budget, averaged over five sampler seeds. The posterior-decoded gated row uses 181 actual calls. "Wrong puzzle" is legal but inconsistent with a clue.

Sampler	Valid	Legal	Wrong puzzle	Rows	Columns	Boxes
SDE	63.25	68.96	5.71	75.01	75.77	81.00
ODE	58.24	62.80	4.56	70.32	71.53	77.75
q-argmax	79.70	86.94	7.24	89.39	89.70	91.92
gated, nearest	82.59	82.66	0.07	85.87	86.27	89.01
gated, posterior	83.36	85.00	1.64	87.98	88.07	90.56

Codebook ablation. We only consider hard Sudoku here. At $d = 128$, frozen codewords reach 79.68 ± 0.42 with q-argmax and 83.70 ± 0.66 with gating; learned codewords reach 76.48 ± 0.16 and 81.03 ± 0.36. We omit $d = 768$ because every seed selects the tightest tested threshold and commits almost every cell early; those rows measure posterior saturation rather than codebook quality. Frozen codebooks suit Sudoku's nine-token vocabulary, whereas language uses learned codebooks for a much larger vocabulary; this ablation does not test the language setting.

E.3 Kernel-consistency study

Radial transport and the quantile map. For an isotropic conditional kernel, let $f_r(c)$ and $F_r(c)$ be the density and CDF of its cosine coordinate. The conditional radial velocity and its spherical lift

are

$$v_r(c) = -\frac{\partial_r F_r(c)}{f_r(c)}, \tag{51}$$

$$\mathbf{u}_r(\mathbf{z} \mid \mathbf{e}) = \frac{v_r(c)}{1-c^2}(\mathbf{e} - c\mathbf{z}). \tag{52}$$

The marginal field is the posterior average of these conditional fields. Given the marginal velocity $\mathbf{u}_\theta$ and score $\mathbf{s}_\theta$, the velocity–score family is

$$\mathbf{z} \leftarrow \mathrm{Exp}_{\mathbf{z}}\Big(h[-\mathbf{u}_\theta + \varepsilon \mathbf{s}_\theta] + \sqrt{2\varepsilon h}\,\boldsymbol{\xi}\Big), \tag{53}$$

with tangent noise $\boldsymbol{\xi}$; $\varepsilon = 0$ gives the ODE, and $\varepsilon = 1/(d-1)$ recovers the Brownian reverse SDE for Heat. The exact conditional quantile map in Section 3.4 avoids integrating through low-density gaps between mixture modes. The subsequent posterior-averaged spherical displacement is first-order.

Table 24: Hard-Sudoku accuracy (%) at $B = 180$ sampler calls across three independent training seeds. The Heat and moment-matched vMF arms use an identical paired protocol, with inference settings matched within each procedure. Entries are mean ± standard deviation. Posterior-decoded gated rows use 181 actual calls.

Training kernel	ODE	SDE	q-argmax	Gated q-argmax
Heat	58.42 ± 0.44	62.83 ± 1.16	73.88 ± 0.41	83.42 ± 0.32
Moment-matched vMF	58.81 ± 0.47	63.11 ± 1.14	73.92 ± 0.50	82.61 ± 0.31
Δ (vMF − heat)	$+0.39$	$+0.28$	$+0.04$	-0.81

Kernel family. For $a \in [0, 1)$, let $\mathrm{vMF}_d(a \mid \mathbf{e})$ denote the von Mises–Fisher distribution on $\mathbb{S}^{d-1}$ centered at $\mathbf{e}$ with concentration $\kappa_d(a)$ chosen to give mean cosine a; equivalently, $I_{d/2}(\kappa_d(a))/I_{d/2-1}(\kappa_d(a)) = a$. For $m = e^{-r}$, define

$$K^{\mathrm{mix}}_{r,\lambda}(\cdot \mid \mathbf{e}) = m\ \mathrm{vMF}_d(m + \lambda(1-m) \mid \mathbf{e}) + (1-m)\ \mathrm{vMF}_d((1-\lambda)m \mid \mathbf{e}). \tag{54}$$

The component weights are m and $1 - m$, and the corresponding component mean cosines are $m + \lambda(1 - m)$ and $(1 - \lambda)m$. Their weighted mean is m, so λ changes the conditional shape without changing the retained mean cosine. At $r = 1.6$, the cosine standard deviation is 0.08 for Heat, 0.22 for $\lambda = .5$, and 0.37 for $\lambda = .9$.

Table 25: Radius at remaining token-entropy fraction u. A smaller radius reveals the token later during generation. The full SLERP path matches Heat within $|\Delta u| \le 0.016$. This comparison uses the full-path clock; the truncated variant is evaluated separately. Section 4.2 reports SLERP truncated at $r_{\min} = 0.942$; Table 33 compares both paths.

u	Heat	vMF	Mix .5	Mix .9
.05	1.05	1.03	.47	.056
.25	1.31	1.31	.81	.31
.50	1.56	1.55	1.18	.73
.90	2.39	2.38	2.51	2.30
SLERP (full): $\lvert\Delta u\rvert \le 0.016$ relative to Heat; separate truncation ablation: $r_{\min} = 0.942$.				

Valid transport checks. Write $V = v_r(c)/(1-c^2)$, $S = a_r(c)/(d-1)$, and $w = 1 - c^2$. Our primary raw compatibility diagnostic is the tangent-norm L^1 ratio

$$T_0^{(1)} = \frac{\mathbb{E}[|V+S|\sqrt{w}]}{\mathbb{E}[|V|\sqrt{w}]}. \tag{55}$$

Table 26 estimates it with 20,000 draws per radius. The L^2 diagnostic in Table 29 instead uses 50,000 draws per radius and estimates

$$T_0^{(2)} = \frac{\sqrt{\mathbb{E}[(V+S)^2 w]}}{\sqrt{\mathbb{E}[V^2 w]}}. \tag{56}$$

The two estimators differ even on the same samples. We use $T_0^{(1)}$ for raw-compatibility comparisons: the order of magnitude is about 10^{-4} for Heat, 10^{-2} for moment-matched vMF, and order one for the mixtures. Raw T_0 also counts a pure change of speed, so the fitted residual below separates time change from shape. Valid-ODE accuracy is compared in Table 28.

Table 26: Raw $T_0^{(1)}$: the L^1 estimator in Equation 55, using 20,000 draws per radius at $d = 128$. The mixture estimates are most informative at the level of order of magnitude because of estimator variability.

r	Heat	moment vMF	mix .5	mix .9
.05	7.2×10^{-4}	3.6×10^{-3}	1.00	6.49
1.0	7.9×10^{-5}	1.9×10^{-2}	.87	3.69
1.6	4.6×10^{-5}	1.3×10^{-2}	.84	4.33
3.0	3.2×10^{-5}	9.8×10^{-3}	.80	1.41

Tail sensitivity of compatibility estimates. The mixture velocity $v_r = -\partial_r F_r / f_r$ can become very large in the low-density valley between its two modes. On identical samples, repeated estimates show that the largest 0.1% of draws account for 84–99% of $\mathbb{E}[V^2 w]$ for the mixtures (Table 27). When these large velocities dominate both numerator and denominator, $T_0^{(2)}$ is driven toward one and can vary substantially between repeats. The L^1 estimator is more stable, but is not immune to rare draws. The light-tailed vMF control is stable under both estimators. The shape residual uses the same second moments and inherits this sensitivity for mixtures: we interpret its mixture values as orders of magnitude, while the two-to-four-order gap from vMF supports the qualitative ordering.

Table 27: Compatibility estimators recomputed on identical samples, with two repeats per setting at $d = 128$. The last column is the fraction of $\mathbb{E}[V^2 w]$ contributed by the largest 0.1% of samples.

Kernel	r	$T_0^{(1)}$, repeats	$T_0^{(2)}$, repeats	Tail share
vMF	1.0	.019 / .019	.024 / .024	.001
Mix .5	1.0	.87 / .87	.99 / .99	.90–.94
Mix .9	.05	7.40 / 7.34	2.46 / 1.51	.84–.96
Mix .9	1.0	4.14 / 4.32	1.05 / 1.12	.97–.99

The conditional-law and intermediate-marginal checks intermediate marginals and final token frequencies. At 360 steps, the quantile ODE reaches KS at most 0.026 for Heat, vMF, and $\lambda = .5$, and 0.023–0.044 for $\lambda = .9$, against null floors of 0.004–0.014. The velocity integrator fails for the mixtures. The Heat score-ODE arm in the kernel study reaches $58.42 \pm 0.44\%$ (Table 24); the separate valid-transport study reports $54.5 \pm 3.2\%$ for Heat's quantile ODE (Table 28; the same group appears in Table 33). These estimates come from distinct studies and are therefore interpreted as separate measurements rather than a paired accuracy comparison. We also computed a 99.60–99.95% puzzle-level agreement between Heat paths.

Table 28: Hard-Sudoku accuracy (%) using each kernel's valid quantile ODE, mean $\pm$ sd over three training seeds for every kernel. SLERP uses its full path here; Table 33 also reports the truncated path used in Section 4.2. Posterior decoding gives NFE $= B+1$. We did not run SLERP at $B = 720$.

B	Heat	vMF	SLERP (full)	Mix .5	Mix .9
180	54.5 ± 3.2	54.7 ± 2.3	56.2 ± 2.2	53.3 ± 2.4	55.3 ± 2.3
720	54.4 ± 1.6	55.2 ± 2.1	–	53.2 ± 2.6	55.6 ± 2.2

Shape residual. At each radius, let $V = v_r(c)/(1-c^2)$, $S = a_r(c)/(d-1)$, and $w = 1 - c^2$. Weighted least squares gives $\alpha^* = -\mathbb{E}[VSw]/\mathbb{E}[V^2 w]$ and residual $\|\alpha^* V + S\|/\|\alpha^* V\|$. The residual is zero exactly when the local path is Heat after a time change. For SLERP, only the band-averaged residual is available; its radius-wise range and closest-Heat KS statistic are not reported in Table 29.

Table 29: Compatibility estimates across 13 radii from .05 to 5. Raw $T_0^{(2)}$ is the L^2 estimator with 50,000 draws per radius; Table 26 reports the corresponding L^1-based estimate. Mixture L^2 ratios and residuals are tail-sensitive order-of-magnitude estimates. SLERP's residual is a band average, not a range.

Kernel	d	Raw $T_0^{(2)}$ (L^2)	α^*	Residual	KS
Heat	128	$\leq .001$	1.000	$\leq .001$	$\leq .012$
Heat	768	$\leq .007$	1.000	$\leq .007$	$\leq .030$
moment vMF	128	.004–.025	.993–1.012	.001–.025	.0001–.017
moment vMF	768	.004–.012	.993–1.012	.000–.010	.0001–.022
info-vMF	768	.004–.45	.993–1.45	.000–.010	.0001–.022
SLERP (S-FLM path)	128	–	1.003	.047	–
mix .5	128	.99–1.22	.007–.57	2.0–16	.007–.45
mix .5	768	1.00–1.62	.000–.59	1.6–77	.007–.45
mix .9	128	1.00–2.48	.000–.26	9.2–98	.007–.95
mix .9	768	1.02–10.4	.005–1.78	2.4–44	.007–.95

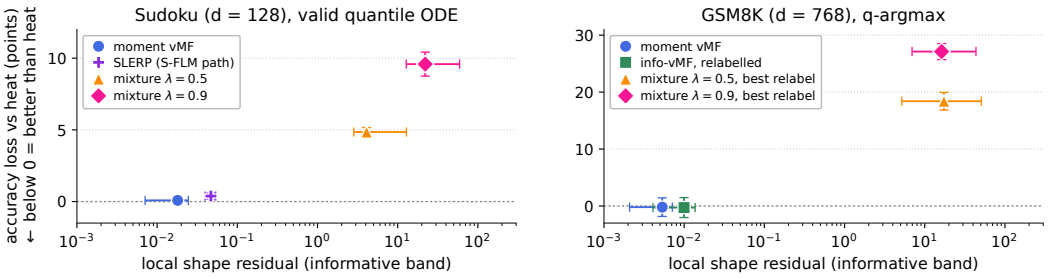

Figure 8: Mismatch cost and the local shape residual. Horizontal bars span the residual range over tested noise levels. Sudoku uses the valid quantile ODE at $B = 180$ with posterior decoding (181 NFEs); only the mixture points use information relabelling. vMF and SLERP use their native clocks. GSM8K uses q-argmax. The residual is local in r, so the figure shows ordering instead of a fitted relationship. Negative values mean better accuracy than Heat. Differences as small as the vMF/SLERP costs are not resolved with three training seeds; this does not apply to the larger mixture losses. Mixture residuals are tail-sensitive and indicate orders of magnitude. SLERP uses Heat-trained weights with the matched geodesic quantile ODE; its native exact velocity and radial transport agree numerically (Appendix E.4).

Figure 8 orders the measured mismatch costs by shape residual; Table 30 lists the plotted values.

Table 30: Kernel mismatch costs shown in Figure 8 on hard Sudoku. Positive values mean worse accuracy than Heat. Costs are paired mean ± sd over three Heat training seeds at $B = 180$ with posterior decoding (181 NFEs). Only mixtures use time relabelling; their residuals are order-of-magnitude estimates.

Kernel	Figure residual	Cost vs. Heat (points)
Moment-matched vMF	.017	$+.08 \pm .08$
SLERP (S-FLM path)	.047	$+.38 \pm .20$
Mixture $\lambda = .5$	5.5	$+4.8 \pm .3$
Mixture $\lambda = .9$	26.9	$+9.6 \pm .8$

Dynamics and decoding. The misspecified $\lambda = .9$ baseline fails for two reasons. Posterior decoding fixes the diffuse final component, while quantile transport fixes the path; Table 31 separates these effects.

Table 31: Hard-Sudoku accuracy (%) at a sampler budget of $B = 180$ network calls (Brownian ODE: Heun, 90 steps; quantile ODE: 180 steps), mean ± sd over three training seeds. Nearest decoding spends 180 NFEs; posterior decoding adds one call (181 NFEs).

Kernel	Dynamics	Nearest	Posterior
Mix .5	Brownian ODE	44.8 ± 1.4	55.0 ± 2.4
	quantile ODE	52.1 ± 2.3	53.3 ± 2.4
Mix .9	Brownian ODE	0.0 ± 0.0	15.3 ± 3.6
	quantile ODE	11.2 ± 0.7	55.3 ± 2.3

Threshold tuning and information grid. Thresholds are selected per kernel and budget on held-out puzzles. At $B = 180$ the chosen values are 0.999995 for Heat, 0.999999 for $\lambda = .5$, and 0.9999995 for $\lambda = .9$. The information grid blends measured information progress with log-radius progress at weight 0.15. At $B = 180$ (181 NFEs with posterior decoding), q-argmax reaches 77.9 ± 1.8, 83.9 ± 1.2, and 83.1 ± 1.9 for Heat, $\lambda = .5$, and $\lambda = .9$. Tuned gating reaches 82.4 ± 1.6, 76.7 ± 1.4, and 78.6 ± 2.0. These comparisons use five independently trained models per kernel, so the deviations measure training-seed variation. Table 32 reports the full gated-kernel comparison.

Table 32: Effect of the training kernel on tuned confidence-gated refinement on hard Sudoku. Entries are accuracy differences in percentage points relative to Heat on the information grid, mean ± standard deviation over five independent training seeds per kernel. Parentheses give the paired t statistic. SLERP is a separate three-training-seed comparison with paired spreads; see Table 33. Budgets are sampler calls B, with one further call for posterior decoding.

B	Moment-matched vMF	SLERP	Mix $\lambda = .5$	Mix $\lambda = .9$
16	$+0.3 \pm 0.8$	–	-4.7 ± 2.3 $(t = -4.5)$	-5.0 ± 4.1 $(t = -2.7)$
64	$+0.2 \pm 1.1$	$+1.1 \pm 3.6$	-5.2 ± 2.4 $(t = -4.8)$	-5.3 ± 3.6 $(t = -3.3)$
180	-0.1 ± 1.0	$+0.1 \pm 3.4$	-5.8 ± 1.8 $(t = -7.4)$	-3.8 ± 3.0 $(t = -2.8)$

Moment-matched vMF is also within ± 0.5 points of Heat at every tested budget on the log grid, including -0.2 ± 1.1 points at $B = 180$, and selects a similar confidence threshold. On the information grid, the $B = 16$ difference of $+0.3 \pm 0.8$ points gives an approximate two-sided 95% t interval of $[-0.7, 1.3]$ points. We therefore describe the two channels as practically equivalent within roughly one point at that budget, rather than as exactly equal. The same near-zero pattern holds at $B = 64$ and $B = 180$, with wider intervals. vMF's q-argmax result is 0.5–0.8 points above Heat, but is only borderline at $B = 64$ ($t = 2.6$) and weaker elsewhere. The evidence is insufficient to establish a vMF advantage. Together with Proposition 1, this shows that moment-matched vMF remains within roughly one point of Heat under gated refinement, unlike the more distant bimodal kernels.

Time-relabelled cross. After information relabelling, the valid quantile ODE loses 4.8 ± 0.3 points for $\lambda = .5$, slightly more than its 4.4 ± 0.6-point loss before relabelling. For $\lambda = .9$, relabelling reduces the loss from 14.6 ± 2.9 to 9.6 ± 0.8 points. For q-argmax, Heat weights with mixture re-noising, relabelling, and the mixture information grid reach 82.5 ± 1.3 at $\lambda = .5$, compared with 77.9 ± 1.8 for Heat and 83.9 ± 1.2 for the matched mixture.

E.3.1 Brownian score-substitution baseline

This baseline applies Brownian dynamics to substituted mixture scores and uses nearest-codeword decoding, a natural but invalid choice for a non-Brownian path. At $B = 180$, Heat-trained SDE/ODE under $\lambda = .9$ both solve 0.0% of puzzles; q-argmax reaches 12.5 ± 3.3 and gating 80.3 ± 1.9. Increasing Heun resolution from 180 to 11,520 NFEs does not repair the misspecified dynamics. We include these misspecified-dynamics runs as failure-mode baselines.

Reproducibility note. Evaluations use checkpoints in evaluation mode, count actual Transformer calls, keep stochastic noise outside the drift cap (i.e. a clamp on the norm of the deterministic drift (score or velocity) step in the SDE/ODE integrators.), reseed every sampler, and use a tuning set disjoint from training and evaluation. We also validated the conditional CDF derivative.

E.4 SLERP comparison

We compare full and truncated SLERP paths alongside Heat. Each run uses 20k updates, global batch size 256, the same eight-layer backbone as the other Sudoku arms, 2,000 evaluation puzzles, and posterior decoding. Results below use three independent training seeds. The full SLERP path matches the Heat information clock within $|\Delta u| \leq 0.016$. We also evaluate a truncated path ending at $r_{\min} = 0.942$, where SLERP's own information is exhausted. Using the S-FLM cutoff unchanged at $d = 128$ would leave 77% of the entropy unresolved. Table 33 gives the accuracy comparison for both paths.

Table 33: SLERP kernel comparison on hard Sudoku. Values are accuracies in percent, mean $\pm$ sd over three independent training seeds; parenthesized values are paired differences from Heat, also mean $\pm$ sd. All runs use the same eight-layer backbone, 20k updates, batch 256, 2,000 evaluation puzzles, and posterior decoding, so total NFE is $B + 1$. The truncated gating difference at $B = 180$ has paired $t = -2.3$ and is not resolved with three seeds.

Sampler	B	Heat	SLERP	SLERP truncated
ODE	16	53.1 ± 2.2	55.0 ± 1.6 $(+2.0 \pm 2.0)$	51.2 ± 2.7 (-1.8 ± 4.8)
ODE	180	54.5 ± 3.2	56.2 ± 2.2 $(+1.8 \pm 3.3)$	51.4 ± 1.9 (-3.1 ± 4.8)
q-argmax	64	65.3 ± 1.5	67.4 ± 1.3 $(+2.1 \pm 2.6)$	66.8 ± 1.7 $(+1.5 \pm 3.2)$
q-argmax	180	71.5 ± 2.5	73.1 ± 1.3 $(+1.6 \pm 2.5)$	71.9 ± 1.6 $(+0.3 \pm 3.9)$
Gating, tuned	64	78.3 ± 2.0	79.5 ± 2.1 $(+1.1 \pm 3.6)$	77.7 ± 0.9 (-0.7 ± 2.7)
Gating, tuned	180	81.7 ± 1.1	81.8 ± 3.0 $(+0.1 \pm 3.4)$	78.4 ± 1.4 (-3.2 ± 2.5)

Full-path SLERP is nominally ahead of Heat in all six cells. Across both paths, paired differences are at most 3.2 points in magnitude and are not resolved with three training seeds. Its information clock is close to Heat's, so it behaves like Heat and vMF rather than like the bimodal mixtures. The matched SLERP velocity agrees with S-FLM's exact velocity to 1.4×10^{-7} per step and 4×10^{-5} over a rollout. These checks use SLERP's own model. Separately, the $+0.38 \pm 0.20$-point cost in Table 30 comes from Heat-trained checkpoints 0–2 sampled with SLERP's matched geodesic quantile ODE, paired against the same checkpoints with Heat's valid quantile ODE. Both arms use $B = 180$ and posterior decoding (181 NFEs). This cross uses the full SLERP path without time relabelling: $\alpha^* = 1.003$ and its clock differs from Heat by at most 0.016. Moment-matched vMF likewise uses its native clock; only the mixture crosses use information relabelling.

E.5 Published Sudoku baselines

Table 34 reports the published comparison with autoregressive, discrete-diffusion, and continuous-diffusion Sudoku baselines. The SBD row uses the gated q-argmax result at $B = 180$ (181 total NFEs) from Table 19; its uncertainty is over five sampler seeds on the fixed checkpoint for each difficulty. The AR, MDLM, Duo, CANDI, FLM, and S-FLM values in the main comparison come

from their published evaluations (Deschenaux & Gulcehre, 2026; Sahoo et al., 2024; 2025; Pynadath et al., 2026a; Lee et al., 2026). FLM is from Lee et al. (2026). The FMLM+ row uses our five-seed evaluation of the released checkpoint in Table 22 (Agarwal et al., 2026).

Table 34: Sudoku exact-match accuracy (%) at $B = 180$ sampler-call budget. AR, MDLM, Duo, CANDI, FLM, and S-FLM are published values. FMLM+ and SBD use five sampler seeds on fixed checkpoints; gated SBD uses 181 calls. Bold/underline mark the best/second-best non-AR value per column.

Model	Easy	Medium	Hard
Autoregressive			
AR (sample)	13.9	5.1	0.6
AR (greedy)	14.6	5.1	1.0
Discrete diffusion			
MDLM	92.0	77.1	30.2
Duo	96.3	84.7	58.4
Continuous diffusion			
CANDI	79.3	45.9	16.7
FLM	94.2	82.7	44.5
S-FLM	81.5	50.6	14.0
+ $\alpha^\star(0.1)$	94.0	77.6	43.2
+ adaptive schedule	94.8	85.2	45.0
FMLM+ PR	$\underline{97.77 \pm .13}$	$\underline{92.99 \pm .33}$	$\underline{81.65 \pm .36}$
Ours	$\mathbf{99.11 \pm .12}$	$\mathbf{96.66 \pm .15}$	$\mathbf{83.36 \pm .42}$

E.6 GSM8K SETUP AND METRIC

We train on TinyGSM (Liu et al., 2023), which pairs synthetic math word problems with Python solutions generated by GPT-3.5. Evaluation uses all 1,319 problems in the GSM8K test split (Cobbe et al., 2021); we do not report TinyGSM test accuracy. We tokenize the problem and solution with the SmolLM tokenizer and use context length 512. The problem statement is kept clean during training and fixed during generation.

Executed final-answer accuracy. For each problem j, we generate one Python program and execute it to obtain $\widehat{a}_j$. Let a_j be the reference numerical answer and $b_j = 1$ only when execution returns the correct answer. The reported score is

$$\mathrm{Acc}_{\mathrm{GSM8K}} = \frac{100}{1319} \sum_{j=1}^{1319} b_j. \tag{57}$$

Programs that do not produce the correct answer are not counted as successful. We compare answers, not program text: different programs can solve the same problem. This is a single-generation accuracy, not pass@k or a best-of-multiple-generations score. Parentheses in Table 43 give $\sum_j b_j$, and percentages are rounded to two decimal places.

Baselines and comparability. Dense NFE sweeps use the released 250k-update checkpoints for MDLM, Duo, CANDI, FLM, S-FLM, and FMLM+ variants (Deschenaux & Gulcehre, 2026; Agarwal et al., 2026). This suite matches the TinyGSM training data, backbone scale (12 layers, hidden size 768, and 12 attention heads), and 250k model-update budget. Training hyperparameters, including the learning rate and noise schedule, remain method-specific; we do not claim an identical training recipe. The main controlled curves are MDLM, Duo, S-FLM, FMLM+ trained from scratch, and our Brownian model.

FMLM+ initialized from MDLM starts from a pretrained checkpoint, and its distilled variant uses a pretrained MDLM teacher (Agarwal et al., 2026). These curves appear in the enhanced comparison in Figure 9. The 250k updates describe training of the evaluated model and do not include the cost of producing its teacher or initialization. IDLM also uses pretrained MDLM or Duo teachers and initializes its student from them (Li et al., 2026). We retain these strong baselines as teacher-assisted comparisons, not as matched from-scratch or equal-total-compute runs.

IDLM and vMF results are reported values from their papers (Li et al., 2026; Chemseddine et al., 2026b), not evaluations of our models. In Figure 3, † marks the paper-reported vMF curve; IDLM appears only in Table 36. In our numerical tables, † instead indicates pretrained assistance and ‡ indicates self-conditioning with predictor-corrector sampling. We also report vMF PC without self-conditioning from Table 20 of Chemseddine et al. (2026b). Both PC curves use the best reported configuration at each NFE budget, with corrector calls included in the budget; they are not new evaluations under our sampler. Our model uses neither pretrained assistance nor self-conditioning. AR scores use 512 sequential passes. LatentMDM reports 20.8%, 33.8%, and 45.5% with two, four, and eight latent-segment rounds (Kim et al., 2026b). Because a segment round is not a matched NFE, we do not include these values in Table 36. A missing NFE measurement is marked NR; we do not interpolate missing accuracy values.

E.7 Full GSM8K comparison

Figure 9 and Table 35 extend the controlled comparison with teacher-initialized and self-conditioned models.

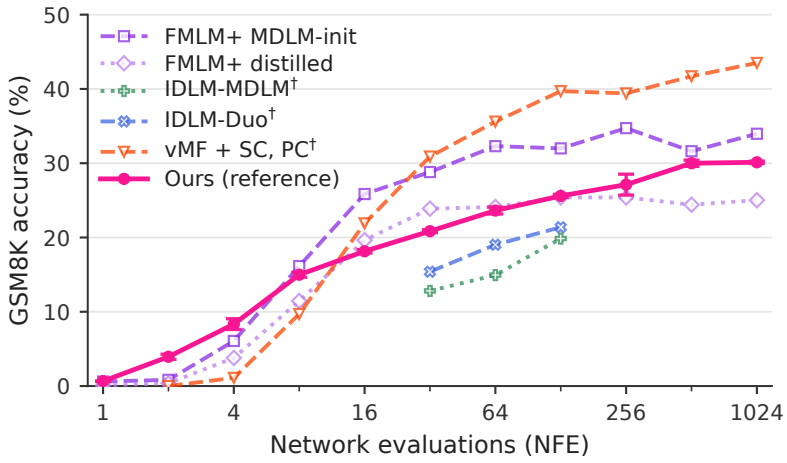

Figure 9: Enhanced GSM8K comparisons using teacher initialization or self-conditioning. Our model uses neither; the controlled comparison is in Figure 3.

Table 35 summarizes the full-budget comparison; Table 36 gives the broader baseline ladder and the three-seed SBD q-argmax values used in Figure 3. The matched-temperature subset is in Table 14, and the SBD memoryless curve in Table 15. We report the complete NFE ladder when a released checkpoint is available and use NR for values that were not reported.

Table 35: GSM8K accuracy (%) at the full reported budget. Diffusion/flow models use the reported 1,024-call budget; decoding regimes differ: MDLM/Duo use $T = 1$, SBD uses q-argmax, and S-FLM top-1 and FMLM+ use deterministic token selection. vMF uses its published PC setting. SBD is the mean over three inference seeds; AR uses 512 sequential passes. Bold/underline mark the best/second-best non-AR result. †: pretrained assistance; ‡: self-conditioning with predictor-corrector sampling. See Appendix E.6 for provenance and comparability.

Model	Accuracy
Autoregressive	
AR (sample)	53.9
AR (greedy)	63.3
Discrete diffusion	
MDLM	16.91
Duo	18.95
Continuous diffusion and flows	
CANDI	0.15
FLM	0.53
S-FLM (exact)	11.98
S-FLM (top-1)	16.53
FMLM+ (scratch)	20.32
FMLM+ (MDLM init.)†	<u>33.97</u>
FMLM+ (distilled)†	25.02
vMF (PC, no SC)	33.0
vMF+SC (PC)‡	**43.5**
Ours: q-argmax	30.12

Table 36: GSM8K final-answer accuracy (%) across network-evaluation budgets. Dense rows are evaluations of released 250k-update checkpoints, excluding teacher/init pretraining costs. vMF PC (with/without SC) and IDLM values are taken from their papers. SBD q-argmax reports a mean and a separate standard-deviation row over three inference seeds; other dense rows are single runs. †: pretrained teacher or initialization; ‡: self-conditioning and PC. NR means that a value is not reported. AR uses 512 sequential passes. Bold marks the best non-autoregressive result at each budget.

Model	1	2	4	8	16	32	64	128	256	512	1024
Autoregressive											
AR (sample)	NR	NR	NR	NR	NR	NR	NR	NR	NR	53.9	NR
AR (greedy)	NR	NR	NR	NR	NR	NR	NR	NR	NR	63.3	NR
Discrete diffusion and distillation											
MDLM	0.00	0.00	0.00	0.23	1.82	5.69	10.01	13.57	15.09	14.94	16.91
Duo	0.00	0.38	1.67	3.64	10.01	13.87	16.60	15.39	18.50	18.27	18.95
IDLM-MDLM†	NR	NR	NR	NR	NR	12.81	14.94	19.86	NR	NR	NR
IDLM-Duo†	NR	NR	NR	NR	NR	15.39	19.03	21.38	NR	NR	NR
Continuous diffusion and flows											
CANDI (lr 10^{-3})	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.08	0.23	0.15
FLM (default)	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.15	0.23	0.45	0.53
S-FLM (exact)	**0.83**	2.58	4.70	7.51	9.25	10.31	11.22	11.22	11.68	11.75	11.98
S-FLM (top-1)	**0.83**	3.34	8.04	11.22	15.47	16.45	16.07	16.07	16.60	16.45	16.53
FMLM+ (scratch)	0.15	0.15	3.87	8.19	14.48	19.26	20.92	21.68	20.62	21.99	20.32
FMLM+ (MDLM init.)†	0.61	0.83	6.07	**16.15**	**25.85**	28.81	32.30	31.99	34.72	31.61	33.97
FMLM+ (distilled)†	0.08	0.38	3.79	11.45	19.64	23.88	24.11	25.40	25.40	24.41	25.02
vMF (PC, no SC)	NR	0.0	0.7	6.2	15.5	21.8	26.9	27.4	31.8	32.4	33.0
vMF+SC (PC)‡	NR	0.0	1.1	9.7	21.9	**30.9**	**35.6**	**39.7**	**39.4**	**41.7**	**43.5**
Ours											
Ours: q-argmax, mean	0.66	**3.94**	**8.34**	14.99	18.15	20.85	23.63	25.60	27.12	30.00	30.12
Ours: standard deviation	0.04	0.35	0.72	0.34	0.23	0.20	0.46	0.22	1.41	0.42	0.16

E.8 S-FLM S-ARCH Q-ARGMAX TRANSFER

We also apply q-argmax to the released S-FLM S-arch checkpoint, using SLERP re-noising at the model's own time labels. The full NFE curve is shown in Figure 3; at higher budgets this refinement roughly doubles GSM8K accuracy relative to S-FLM top-1 decoding. S-FLM and SBD differ in schedule, truncation, embedding handling, and batch size, so this transfer does not isolate a kernel effect.

Table 37 gives the full values. This is one evaluation seed, so it shows transfer feasibility instead of seed-averaged comparative performance.

Table 37: GSM8K accuracy of q-argmax on the released S-FLM S-arch checkpoint, using SLERP re-noising at the model's own time labels. One evaluation seed.

NFE	1	2	4	8	16	32	64	128	256	512	1024
Accuracy (%)	0.8	3.3	8.2	14.9	20.5	23.2	26.2	28.7	29.3	30.5	31.2

E.9 GSM8K KERNEL-CONSISTENCY STUDY

We evaluate all 1,319 GSM8K test problems with EMA weights, context length 512, an entropy-spaced grid, and three sampling seeds. The primary readout is q-argmax with posterior decoding. Quantile-flow and Brownian ODE results are included as checks, although both continuous samplers are weak on this task.

Kernel cross. Moment-matched vMF remains close to Heat, while the mixture mismatch destroys program accuracy even under valid transport. The information-matched vMF is a pure time-change control: querying the Heat network at the equal-cosine time removes its low-budget cost (Table 38).

Table 38: GSM8K accuracy (%), mean $\pm$ standard deviation over three inference seeds. Each row names training weights and inference kernel. Budgets are sampler calls B; posterior decoding adds one call, giving 65 and 257 total NFEs.

Weights $\rightarrow$ kernel	q-argmax		quantile ODE	
	$B = 64$	$B = 256$	$B = 64$	$B = 256$
Heat $\rightarrow$ Heat	$23.63 \pm .46$	27.12 ± 1.41	$6.92 \pm .61$	$7.56 \pm .61$
Heat $\rightarrow$ moment vMF	$22.74 \pm .08$	$27.32 \pm .24$	$6.85 \pm .69$	$7.48 \pm .69$
Heat $\rightarrow$ mixture .5	$0.05 \pm .09$	$0.03 \pm .04$	$0.40 \pm .16$	$0.45 \pm .13$
Heat $\rightarrow$ mixture .9	0.00	0.00	0.00	0.00
Info-vMF $\rightarrow$ info-vMF	$23.05 \pm .85$	$27.24 \pm .76$	$7.15 \pm .16$	$7.43 \pm .27$
Heat $\rightarrow$ info-vMF	$19.81 \pm .42$	$26.59 \pm .42$	$5.94 \pm .36$	$6.85 \pm .56$
Heat $\rightarrow$ info-vMF, relabelled	$23.10 \pm .27$	$27.39 \pm .42$	$6.87 \pm .78$	$7.43 \pm .67$

Kernel and training comparison. Table 39 compares Heat-trained weights sampled with Heat against information-vMF-trained weights sampled with their own kernel. It changes the training channel and inference kernel while leaving the Heat table resolution unchanged. At $B = 32$ the paired difference is -1.67 ± 0.92 points ($t = -3.1$, $p \simeq 0.09$); at $B = 256$ it is $+0.13$ points.

Denoising probe and relabel maps. We build the relabel maps on 512 TinyGSM validation examples that are disjoint from both the GSM8K test set and the 128-example mechanism probe. For each mixture time, the map chooses the monotone Heat-network time label with the highest clean-token accuracy. This gives an upper bound on what a scalar time change can repair; the probe measurements are in Table 40.

The best-case map partly repairs the .5 mixture but cannot repair .9. At $B = 256$ with posterior decoding (257 NFEs), q-argmax moves from 0.03% to $8.72 \pm 0.60\%$ for .5, while .9 stays at 0.00%; the Heat reference is 27.12%. The mixture tables match their target means, the conditional-law and intermediate-marginal checks pass at $d = 768$, and generated sequences show the same posterior-level corruption.

Table 39: GSM8K kernel/training comparison: Heat-trained Heat inference versus information-vMF-trained information-vMF inference. Values are means over three inference seeds; differences are info-vMF minus Heat. Budgets are sampler calls B, before a separate posterior decode.

B	Heat $\to$ Heat	info-vMF $\to$ info-vMF	Difference
1	0.66	0.96	$+0.30$
2	3.94	4.20	$+0.25$
4	8.34	8.82	$+0.48$
8	14.99	14.25	-0.73
16	18.15	18.12	-0.03
32	20.85	19.18	-1.67 ± 0.92 ($t = -3.1$, $p \simeq 0.09$)
64	23.63	23.05	-0.58
256	27.12	27.24	$+0.13$

Table 40: Code-token probe: one-step top-1 accuracy of the Heat posterior on the 9,345 program tokens in 128 TinyGSM validation answers, excluding PAD. Unqualified columns use the raw time label; "best" chooses the best tested Heat-network time label.

r	Heat	moment vMF	mix .5	mix .5, best	mix .9	mix .9, best
.2	1.000	1.000	.996	.999	.828	.971
.5	.999	.999	.994	.998	.620	.945
1.0	.996	.996	.949	.988	.381	.880
1.6	.984	.983	.860	.960	.213	.758
2.5	.936	.936	.814	.883	.153	.481

Why program accuracy amplifies small token errors. The full-answer probe uses the same 128 TinyGSM validation answers, mixture corruption, and raw time labels as the code-token probe, but counts the full answer region (code plus PAD, about 459 positions per answer). PAD accounts for about 80% of these positions and is predicted correctly more than 99% of the time, so these per-token accuracies are not directly comparable with Table 40. In the full-answer probe, Heat at $r = .5$ has full-answer-region token accuracy .9999 and exact sequence accuracy .969. The .5 mixture at $r = 1.0$ has per-token accuracy .9917 but only .250 exact accuracy, while the .9 mixture at $r = .1$ has per-token accuracy .9792 and zero exact accuracy. The mixture-to-Heat success ratio by program length is reported in Table 41.

Table 41: GSM8K success by generated-program length at $B = 256$ with posterior decoding (257 NFEs). The same 1,319 problems define both columns; values average three inference seeds.

Length range (mean)	n	Heat	mix .5, relabelled	ratio
67–108 (96)	259	44.3	21.0	.47
108–125 (116)	258	40.8	13.3	.33
125–142 (132)	265	26.7	6.2	.23
142–164 (152)	266	17.8	3.5	.20
164–290 (189)	271	7.3	.25	.03

The success ratio decreases monotonically with length, with correlation -0.96 between length and log ratio. Here the *independent-token model* assumes that the mismatch independently preserves each generated token with a length-independent probability p; for a program of length L, it therefore predicts a mixture-to-Heat success ratio proportional to p^L. The per-token probability in this diagnostic is measured over the full-answer probe's full answer region, including PAD, instead of the code-token probe's code-only mask. The observed decay is steeper, consistent with context-mediated error propagation. The exponential curve in Figure 2 is a descriptive fit; it is not a causal test of this explanation.

Matched fine-tuning. We fine-tune the same Heat weights for 25k updates on either Heat or the .5 mixture, then sample each model with its own kernel. Table 42 reports the results; each arm uses one training seed, and the standard deviation is over three inference seeds.

This shows that restoring train–inference kernel consistency can recover most of the mismatch loss, but does not determine the performance of a mixture model trained from scratch. The nearest-

Table 42: Matched GSM8K fine-tunes (%), mean ± sd over three inference seeds, with one training seed per arm. The main comparison uses posterior decoding at $B = 256$ (257 NFEs); nearest-codeword controls use B total NFEs. The posterior-decoded paired mixture-minus-Heat difference is -1.09 ± 1.80 points.

Posterior decoding, EMA	$B = 256$ (257 NFEs)
Heat fine-tune → Heat	$27.32 \pm .52$
Mix .5 fine-tune → mix .5	26.23 ± 1.30

Nearest-codeword decoding	64 NFEs	256 NFEs
Heat fine-tune → Heat, EMA	23.83 ± 1.00	$27.32 \pm .52$
Mix .5 fine-tune → mix .5, EMA	$21.10 \pm .49$	26.03 ± 1.22
Heat fine-tune → Heat, no EMA	$21.53 \pm .27$	$26.00 \pm .50$
Mix .5 fine-tune → mix .5, no EMA	$19.16 \pm .24$	23.02 ± 1.37

decoded EMA comparison gives a paired gap of -1.29 ± 1.71 points at 256 NFEs. With posterior decoding, the gap is -1.09 ± 1.80 at 257 NFEs; the three seed differences are $+0.5$, -3.0, and -0.8 points. The residual gap is not resolved by these three inference seeds.

E.10 GSM8K CHECKPOINT SWEEP

Figure 10 summarizes the checkpoint comparison; Table 43 reports all available NFE measurements for the five checkpoints used in our GSM8K ablation. Each percentage is computed on the 1,319-example GSM8K test set; the number of correct programs is shown in parentheses. Entropy sampling is strongest from 16 to 256 NFEs, whereas the online adaptive configuration is strongest at 1 to 8 and 512 to 1,024 NFEs.

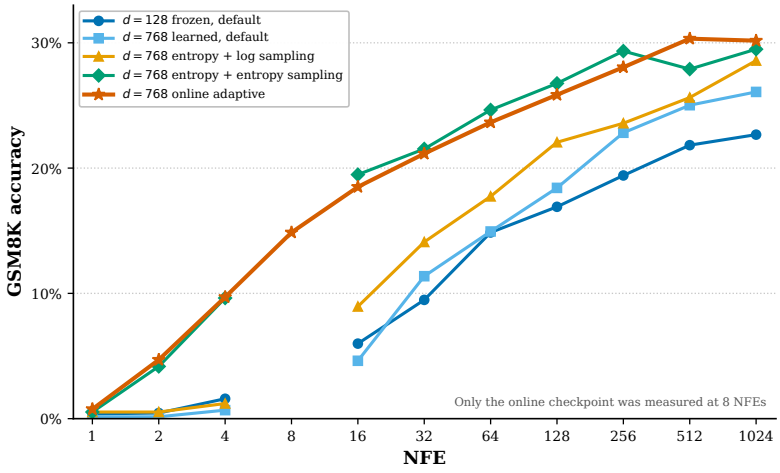

Figure 10: Our GSM8K checkpoint and sampling ablations. All configurations use plain q-argmax without Gumbel perturbations, score updates, self-conditioning, or pretrained assistance. Missing NFE measurements are omitted; Table 43 gives the numerical results.

Full-test continuous-sampler sweep. Table 44 reports ODE and Heun SDE accuracy on all 1,319 GSM8K test questions. ODE is stronger at 32 and 64 NFEs, while Heun SDE is stronger from 128 NFEs onward. Both achieve their highest reported accuracy at 256 NFEs, with ODE matching that value at 512 NFEs. Increasing the budget beyond 256 gives no further gain in this sweep. These full-test measurements are separate from the generated-trajectory diagnostics in Table 48, which use 64 examples per condition and separately record a final diagnostic call.

Table 43: GSM8K final-answer accuracy (%) across NFE budgets for our TinyGSM-trained checkpoints. Parentheses give the number of correct examples out of 1,319. Bold marks the best result at each budget; NR means not reported.

NFE	$d = 128$ frozen default	$d = 768$ learned default	$d = 768$ entropy + log sampling	$d = 768$ entropy + entropy sampling	$d = 768$ online adaptive
1	0.38 (5)	0.15 (2)	0.53 (7)	0.53 (7)	**0.76 (10)**
2	0.45 (6)	0.15 (2)	0.53 (7)	4.17 (55)	**4.70 (62)**
4	1.59 (21)	0.68 (9)	1.21 (16)	9.63 (127)	**9.70 (128)**
8	NR	NR	NR	NR	**14.86**
16	5.99 (79)	4.62 (61)	8.95 (118)	**19.48 (257)**	18.50 (244)
32	9.48 (125)	11.37 (150)	14.10 (186)	**21.53 (284)**	21.15 (279)
64	14.86 (196)	14.94 (197)	17.74 (234)	**24.64 (325)**	23.65 (312)
128	16.91 (223)	18.42 (243)	22.06 (291)	**26.76 (353)**	25.85 (341)
256	19.41 (256)	22.82 (301)	23.58 (311)	**29.34 (387)**	28.05 (370)
512	21.83 (288)	25.02 (330)	25.63 (338)	27.90 (368)	**30.33 (400)**
1024	22.67 (299)	26.08 (344)	28.58 (377)	29.49 (389)	**30.17 (398)**

Table 44: GSM8K executed-answer accuracy (%) for continuous samplers on the full 1,319-question test set. Budgets are reported NFEs. Bold marks the better sampler at each budget. This table is separate from the checkpoint and generated-trajectory ablations.

NFE	ODE	Heun SDE
32	**3.64**	0.00
64	**7.81**	7.20
128	7.96	**10.08**
256	8.34	**12.21**
512	8.34	**11.60**
1024	8.26	**11.98**

E.11 OpenWebText evaluation

Appendix G gives unedited generated excerpts at four target entropy levels and six NFE budgets.

Generation settings. Our reported OWT model uses a freely learned $d = 768$ codebook and the online adaptive training schedule, with 1M updates. For each method and NFE budget, we sweep the same 15 whole-trajectory temperatures, $T \in \{0.50, 0.55, \ldots, 1.20\}$. Apart from temperature, each method's sampler settings remain fixed throughout the sweep. Temperature is applied at every sampling step: to the categorical transition for MDLM and Duo, and to the posterior weights used by S-FLM and our mixture score. We use $\mathrm{NFE} \in \{32, 64, 128, 256, 512, 1024\}$. Our model and the MDLM, Duo, and S-FLM models shown as temperature sweeps use 1M training updates. The main paper shows 1,024 NFEs; Figure 11 gives all six budgets. The FLM, FMLM, and FMLM+ markers are our evaluations of released checkpoints with their native samplers, not paper-reported numbers or temperature sweeps. We do not treat them as matched-training-budget comparisons. For FMLM+, we retain four posterior-refinement rounds, $p = 0.999$, fresh noise, and an equal step budget per refinement round. Only the 128-NFE point uses the authors' released OWT configuration: 32 inner flow steps per round, for $32 \times 4 = 128$ evaluations. At the other budgets, † marks our budget-matched extrapolation of the inner step count to $\mathrm{NFE}/4$, including 256×4 at 1,024 NFEs. These values correspond to directly measured operating points, with perplexity and entropy evaluated at those settings.

Hybrid sampler. For our $d = 768$ model, we use an entropy grid and initialize from the uniform distribution. Above $r = 2.064065$, we use stochastic Heun score updates. Below this switch, k controls the q-projection frequency: $k = 1$ projects at every step, $k = 2$ every second step, and $k = 4$ every fourth step. Intervening steps use continuous stochastic score/SDE updates. At a fixed NFE budget, increasing k therefore allocates fewer updates to q-projection and more to the score sampler. Here k is a cadence, not top-k posterior truncation. The default is $k = 1$; at 1,024 NFEs we also evaluate $k = 2$ and $k = 4$ with the same temperature sweep and all other sampler settings fixed.

In the Gumbel configuration, q-projection uses the perturbed posterior without the additional commitment gate (the confidence gating defined in Section 3). One fixed Gumbel noise tensor is reused at every q-projection. Its effective strength is

$$\gamma(r) = 0.25\,\frac{\log(r/r_{\min})}{\log(r_{\max}/r_{\min})}. \tag{58}$$

Thus 0.25 is a base multiplier, not a final temperature: the strength is 0.25 at $r_{\max}$ and decreases toward zero at $r_{\min}$. Since q-projection begins below the switch, its first effective strength is already smaller than 0.25. Whole-trajectory temperature is swept while this schedule and the other settings remain fixed. Budgets are expressed in network evaluations, not in grid points; each Heun update requires two evaluations.

Generative perplexity. We use the sample-level quality and diversity metrics of partition generative modeling (Deschenaux et al., 2026b). The main temperature sweeps and isolated reference points use $N = 512$ sequences per setting. The plotted $k = 2$ and $k = 4$ variants instead use 64 sequences, with 32 at each of seeds 1 and 2; the plotted $k = 1$ curve uses 512, with 64 at each of seeds 1 to 8. We also evaluate $k = 1$ with the same 64-sample protocol as $k = 2/4$ for the controlled temperature comparison below. All sequences have length 1,024 and are scored with the same pretrained GPT-2 Large reference model and revision (Radford et al., 2019), using its tokenizer. For a sample $\mathbf{x}$, we compute

$$\mathrm{PPL}_{\mathrm{gen}}(\mathbf{x}) = \exp\left(-\frac{1}{m}\sum_{i=1}^{m}\log p_{\mathrm{GPT2\text{-}L}}(x_i \mid x_{<i})\right), \tag{59}$$

where m is the number of scored tokens before the first end-of-text marker. Tokens at and after that marker are excluded from the score. This measures the reference model likelihood assigned to generated text. Lower values indicate text that the reference model finds more likely, but can also result from repetitive outputs.

Unigram entropy. Within each sample, let $c(v, \mathbf{x})$ count occurrences of token v and $L_{\mathbf{x}}$ be the number of tokens used in the entropy calculation. The empirical frequencies are $\widehat{p}_{\mathbf{x}}(v) = c(v, \mathbf{x})/L_{\mathbf{x}}$. We compute

$$H(\mathbf{x}) = -\sum_{v\in\mathcal{V}}\widehat{p}_{\mathbf{x}}(v)\log\widehat{p}_{\mathbf{x}}(v). \tag{60}$$

We take $0\log 0 = 0$ and use natural logarithms, so entropy is measured in nats. Repeated use of a small set of tokens gives low entropy; a broader token distribution gives higher entropy. This is a within-sample token diversity measure, not the entropy of the learned posterior or a measure of semantic diversity across the batch.

Averaging and frontier interpretation. Each temperature contributes one point obtained by averaging the two metrics separately across samples:

$$\overline{H} = \frac{1}{N}\sum_{n=1}^{N}H(\mathbf{x}^{(n)}), \qquad \overline{\mathrm{PPL}}_{\mathrm{gen}} = \frac{1}{N}\sum_{n=1}^{N}\mathrm{PPL}_{\mathrm{gen}}(\mathbf{x}^{(n)}). \tag{61}$$

We average per-sample perplexities after exponentiation; this differs from exponentiating the pooled mean negative log-likelihood. We likewise average per-sample unigram entropies rather than pooling token counts across samples.

The plotted curves retain only measured Pareto-nondominated points: a point is excluded if another measured point has at least as much entropy and no larger perplexity, with one strict improvement. We connect the retained points without fitting a curve. The temperature-matched ablation uses all 15 settings before this filtering.

For each NFE budget, the temperature sweep gives a quality-diversity curve. At a shared entropy, lower generative perplexity is better. Perplexity alone can favor collapsed samples, while entropy alone does not guarantee coherent text; the paired metrics expose this trade-off. We compare only the entropy range reached by both methods and do not extrapolate a curve beyond its measured range. The sweep gives an empirical frontier under the stated sampler settings, without asserting global optimality across all possible samplers.

Results across sampling budgets. Figure 11 gives the complete NFE sweep. At 128 NFEs, our default $k = 1$ sampler has lower perplexity than S-FLM at low and moderate entropy, but the advantage disappears at the high-entropy end. At 1,024 NFEs, it has lower perplexity throughout their shared entropy range. MDLM and Duo generally achieve lower perplexity at matched entropy. Increasing NFE mainly improves our low- and moderate-entropy performance, while the default sampler's high-entropy coverage remains limited. The isolated FLM, FMLM, and FMLM+ reference points are not full temperature sweeps, so we do not infer their frontiers from these markers.

Effect of q-update frequency. The controlled k ablation uses 64 samples per temperature and variant, with 32 samples at each of seeds 1 and 2. It holds the model, nominal NFE budget, noise switch, grid construction rule, Gumbel schedule, whole-trajectory temperature, and reference-model revision fixed, changing the allocation between q-updates and continuous stochastic score updates below the switch. Entropy and perplexity both increase from $k = 1$ to $k = 2$ to $k = 4$ at every tested temperature. This operating-point shift should be distinguished from a comparison at matched entropy, which can pair different temperatures: in the common entropy range, $k = 1$ generally has lower perplexity than $k = 2$ and $k = 4$. Larger k extends the attained entropy range, so no $k = 1$ comparison is available beyond its measured endpoint. Budgets refer to network calls, not grid points: $k = 1$ uses 443 two-call score updates and 138 one-call q-updates, $k = 2$ uses 479 and 65, and $k = 4$ uses 496 and 32. The realised NFEs are therefore 1,024, 1,023, and 1,024, respectively; $k = 2$ leaves one evaluation unused because of the two-call Heun updates. Table 45 gives representative temperatures from the matched-64-sample data, separate from the 512-sample $k = 1$ curve in Figure 11. These aggregate means do not provide error bars or a statistical significance test.

Table 45: Temperature-matched OWT k ablation. Each entry uses 64 samples (two seeds, 32 each) at the same nominal 1,024-NFE budget; realised NFEs are 1,024 / 1,023 / 1,024 for $k = 1/2/4$. H is mean unigram entropy in nats and PPL is mean generative perplexity.

	$k = 1$		$k = 2$		$k = 4$	
T	H	PPL	H	PPL	H	PPL
0.80	4.56	16.32	4.61	20.82	4.78	26.13
1.00	5.18	35.12	5.24	43.58	5.36	59.42
1.20	5.36	65.10	5.57	136.87	5.75	230.57

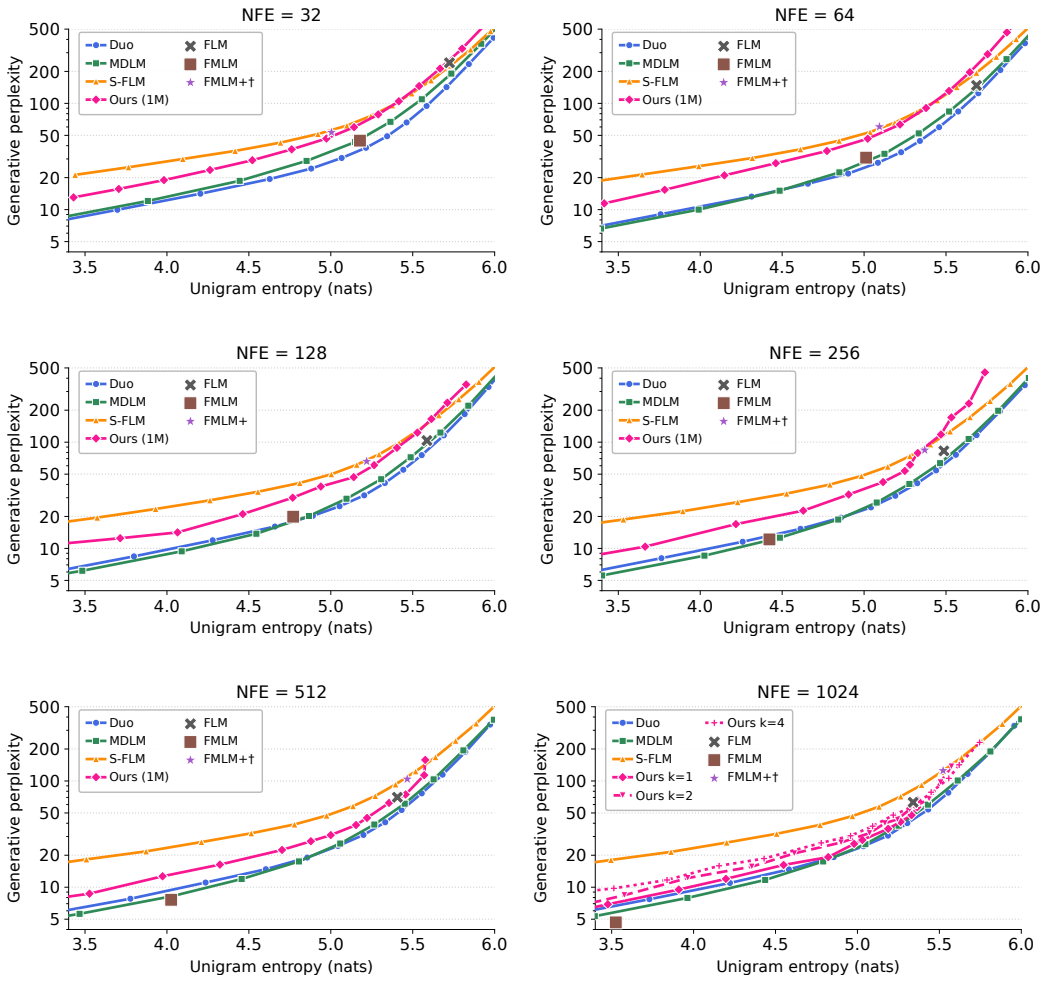

Figure 11: OpenWebText frontiers at all six budgets. Our 1M-update model uses $k = 1$ unless noted. Curves show measured Pareto-filtered points from 15 temperatures, with 512 samples per point except $k = 2/4$ (64). Reference markers use 512 samples; $k = 2$ realises 1,023 evaluations. FMLM+ † denotes our budget-matched extrapolation of the inner flow-step count with four refinement rounds (256×4 at 1,024); only $32 \times 4 = 128$ is the released OWT configuration. Lower perplexity is better at matched entropy.

Evolution during training. Figure 12 compares our OWT frontiers after 50k, 100k, 500k, and 1M training updates at 256 and 1,024 NFEs. Pynadath et al. (2026b) report that 50k-update frontiers can already closely resemble those at 1M updates for MDLM, Duo, and CANDI. Our early checkpoints also produce a clear quality-diversity trade-off, but later training lowers perplexity at matched entropy, particularly in the intermediate- and high-entropy range. Most of the visible improvement occurs by 500k updates, while the 500k and 1M curves are close in both panels.

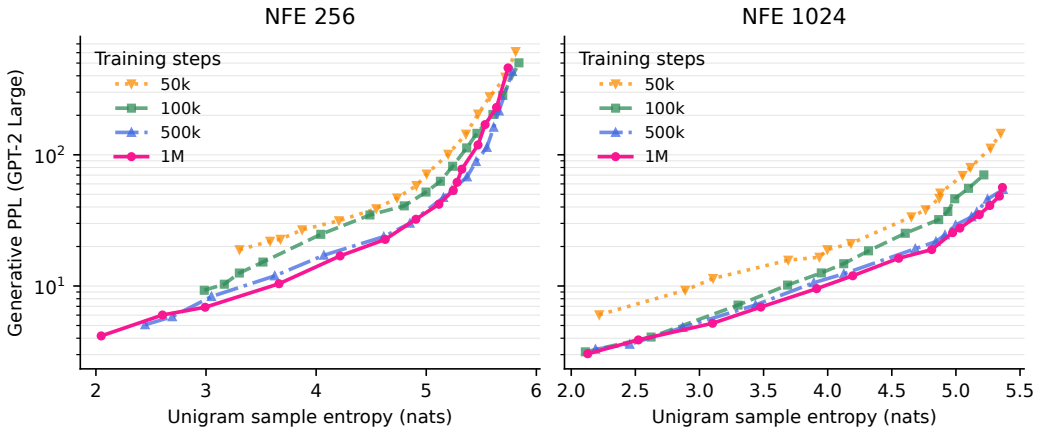

Figure 12: Evolution of the OWT generative frontier during training. Curves compare checkpoints at 50k, 100k, 500k, and 1M updates, with 256 NFEs (left) and 1,024 NFEs (right). This 128-sample checkpoint diagnostic uses GPT-2 Large to measure generative perplexity; lower perplexity is better at matched unigram entropy. The 500k and 1M curves are close, although both improve over the early checkpoints across much of the shared entropy range. Sample counts differ from the main 512-sample frontier evaluation.

E.12 Discrete commitment across tasks

We distinguish posterior diagnostics on forward-noised reference examples from measurements along actual generation trajectories. The former test denoising states obtained from known clean sequences; the latter include states induced by the sampler's own predictions. Their statistics need not agree. These additional diagnostic runs complement the main benchmark results and the OWT temperature sweeps.

Diagnostic setup. Table 46 identifies the checkpoints and grids. Language checkpoints use exponential-moving-average (EMA) weights. Generated-trajectory diagnostics use 64 examples per condition and sampling seeds 1 and 2. Forward readiness uses 128 validation examples, two noise seeds, and 41 radii; no forward GSM8K probe is run. OWT score temperatures are $T \in \{0.60, 0.75, 0.90, 1.05, 1.20\}$.

Table 46: Setup for the additional commitment diagnostics. These runs are separate from the matched-NFE Sudoku accuracy sweep.

Task	Checkpoint	Trajectory settings
Sudoku easy/medium	8-layer DiT, embedding 512; $d = 768$, learned-online, 20k updates	$T = 1$, log-radius grid
TinyGSM / GSM8K	Online 250k checkpoint, EMA	Length 512, $T = 1$, entropy grid
OWT	Free-online 1M checkpoint, EMA	Length 1,024, entropy grid; score-temperature sweep

Posterior diagnostics. At position ℓ, write $q_k^\ell(v) = q_\theta^\ell(v \mid \mathbf{z}_k, r_k)$. The token-posterior entropy, confidence indicator, and adjacent-state MAP stability indicator are

$$\begin{aligned} H_q^\ell(k) &= -\sum_v q_k^\ell(v) \log q_k^\ell(v), \\ C_{q,c}^\ell(k) &= \mathbf{1}\left[\max_v q_k^\ell(v) > c\right], \\ S_q^\ell(k) &= \mathbf{1}\left[\arg\max_v q_k^\ell(v) = \arg\max_v q_{k+1}^\ell(v)\right]. \end{aligned} \tag{62}$$

Entropy uses natural logarithms. We record confidence thresholds $c \in \{0.9, 0.99, 0.999\}$ and show $c = 0.99$ in Figure 13. Confidence and stability rates summarize the corresponding indicators over the evaluated positions and examples. Stability measures agreement between token hypotheses, not agreement with the correct answer. Posterior entropy also differs from the unigram entropy of generated text (Equation 60). Prompt positions are excluded throughout. The primary masks contain unknown Sudoku digits, reference non-PAD TinyGSM answer positions, all GSM8K answer slots including PAD, and all OWT positions. A separate GSM8K diagnostic restricts to predicted non-PAD positions. Because the masks differ, their aggregate confidence statistics answer different questions. Adjacent-state stability compares predictions at the sampler's actual successive radii. The q-argmax and Heun grids have different radius increments at the matched language-call budget; both radii are recorded. Their stability values are therefore not directly matched over a common noise interval. Comparing stability across samplers therefore requires matched radius increments.

Diagnostic call accounting. Language trajectories use 256 sampling calls: 256 q-argmax steps, 128 two-call Heun SDE steps, or 111 Heun steps plus 34 q-argmax steps for the native OWT hybrid ($2 \times 111 + 34 = 256$). Each run also records one separate final diagnostic call, giving 257 network evaluations in total. The reasoning comparison below is therefore matched in both sampling and total call counts. Sudoku trajectory diagnostics instead use the same 90-step noise grid for both samplers, requiring 90 q-argmax calls versus 180 Heun calls. These diagnostics compare posterior behavior along a shared grid; matched-NFE accuracy is reported separately in Table 19.

Forward-noised reference examples. Table 47 reports diagnostics near $r = 3.04$. Sudoku posteriors are almost fully resolved, TinyGSM is intermediate, and OWT retains much greater uncertainty. These results describe readiness for a single projection; repeated-projection behavior is examined below.

Table 47: Posterior diagnostics on forward-noised reference examples near $r = 3.04$. Entropy is in nats; confidence and adjacent-noise MAP stability are percentages. Measurements are rounded.

Task	Posterior entropy	$\max q > 0.99$	MAP stability
Sudoku easy	≈ 0	100	100
Sudoku medium	≈ 0	99.99	100
TinyGSM	0.20	79.4	97.2
OWT	5.61	0.51	55.3

Generated trajectories. The fixed-checkpoint Sudoku sweep in Table 19 shows that inference procedure changes accuracy at matched NFE. On hard puzzles, increasing the budget from 16 to 180 NFEs gives gains of 10.74 points for q-argmax and 8.56 points for gated q-argmax. Nevertheless, trajectory diagnostics near $r = 3.15$ find that approximately 98% to 99.5% of unknown Sudoku cells have $\max q > 0.99$ for both q-argmax and Heun SDE, with almost perfectly stable adjacent-state MAP predictions. These readiness traces use easy and medium puzzles with a learned $d = 768$, 20k-update checkpoint; the accuracy sweep uses the frozen $d = 128$ checkpoint. At these measured noise levels on the shared 90-step grid, the SDE is not simply uncertain while q-argmax is sharp. Because this diagnostic uses a different checkpoint and unequal call counts, it cannot explain the matched-NFE accuracy gap by itself. The samplers act differently on their predictions: q-argmax repeatedly selects a codeword and re-noises it, whereas the SDE evolves a continuous state with the soft mixture score.

Table 48 gives the separate reasoning and OWT comparisons at 256 sampling calls plus one diagnostic call. For TinyGSM, q-argmax has lower posterior entropy, a larger high-confidence fraction,

and higher accuracy; its accuracy advantage also appears on GSM8K. The pattern is consistent with a refinement effect, although confidence does not directly measure content-token correctness. TinyGSM's primary statistics exclude reference PAD. GSM8K's PAD-inclusive confidence must instead be interpreted alongside its separate predicted-non-PAD diagnostic.

Table 48: Additional generated-trajectory diagnostics with 256 sampling calls plus one separate final diagnostic call (257 total evaluations). Accuracy and high-confidence fractions are percentages; both entropy metrics use nats but measure different quantities. These runs are separate from the main checkpoint sweeps.

Task	Metric	q-argmax	Heun SDE
TinyGSM	Posterior entropy	0.12	0.32
	Positions with $\max q > 0.99$	83.6	70.9
	Executed-answer accuracy	39.1	20.3
GSM8K	Executed-answer accuracy	24.2	12.5
OWT	Sample unigram entropy	≈ 0.014	5.31

For OWT, pure q-argmax can become stable and confident while producing almost constant text. Its near-zero sample unigram entropy is evidence of within-sample token-diversity collapse, not successful generation. The hybrid's continuous stochastic phase provides an alternative to early repeated MAP commitment. The late-stage k ablation in Table 45 separately shows that reducing projection frequency increases both entropy and perplexity at fixed temperature. The ablation therefore establishes only the trade-off among the tested projection frequencies.

Irreversible freezing and padding. A separate GSM8K diagnostic uses a 512-token context and 1,024 NFEs, covering all 1,319 questions with sampling seeds 3 and 4. Naive permanent confidence freezing solves 0 of 1,319 problems. It is distinct from the 64-example trajectory diagnostics above, although it uses the same context length as the main reasoning experiments. One possible failure mechanism is premature PAD commitment: confident PAD predictions can freeze positions that later need to become content, removing the revision channel. Sudoku has no corresponding variable-length output decision: every solution cell must contain a digit.

Interpretation. Our working explanation separates stable token hypotheses from the desirability of mode-seeking. Constrained evaluation rewards recovery of a correct answer, whereas unconditional text evaluation also requires diversity. High confidence can accompany an incorrect token, premature PAD, or repetitive output, making it insufficient for irreversible commitment.

E.13 Learned codebook geometry

Checkpoint and controls. This is a separate OWT diagnostic with $d = 128$, not the $d = 768$, 1M-update model used for the generative frontier. We analyze the free learned codebook at 150k updates, with anchor weight zero and separation weight one. Noise levels are drawn log-uniformly for the first 100k updates and from the entropy-based schedule from 100k to 150k. The controls are a frozen max-min packing and a random spherical codebook at the same dimension. We compute all diagnostics with the same evaluation code.

Geometry and alignment metrics. For each unit codeword, the nearest-neighbor angle is

$$\theta_v = \arccos\left(\max_{w \neq v} \langle \mathbf{e}_v, \mathbf{e}_w \rangle\right). \tag{63}$$

We summarize its median, first percentile, and minimum over the vocabulary. Effective rank is the participation ratio of the centered covariance spectrum: if λ_i are its eigenvalues,

$$r_{\text{eff}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}. \tag{64}$$

Centering matters: at 150k, the uncentered value is 44.04, compared with 50.38 after centering (50.4 when rounded).

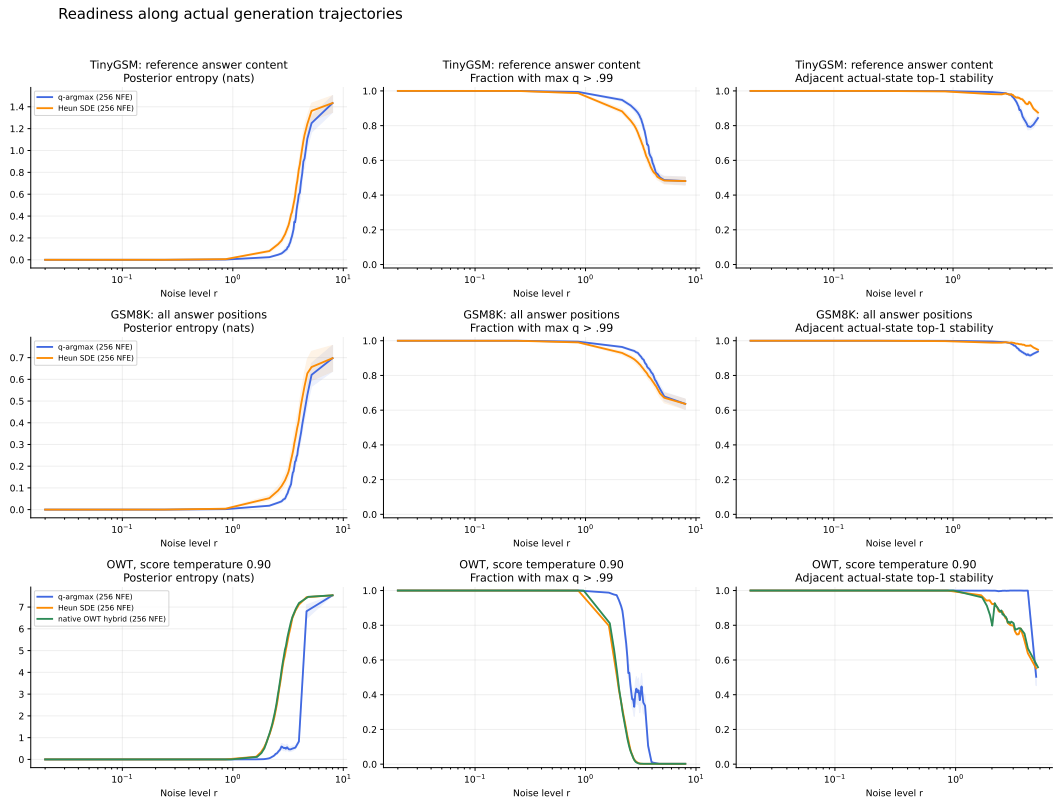

Figure 13: Posterior readiness along generated language trajectories. Rows: TinyGSM reference non-PAD answer positions, GSM8K answer slots including PAD, and OWT at score temperature $T = 0.90$. Columns: posterior entropy, the fraction with $\max q > 0.99$, and adjacent-state MAP stability. Blue: q-argmax; orange: Heun SDE; green: the native OWT hybrid. Each method uses 256 sampling calls plus one separate final diagnostic call. The horizontal axis is noise level r; generation proceeds from right to left. Stability uses different adjacent-radius increments across samplers and is not a common-interval comparison. GSM8K confidence includes PAD and should not be read as content-only certainty.

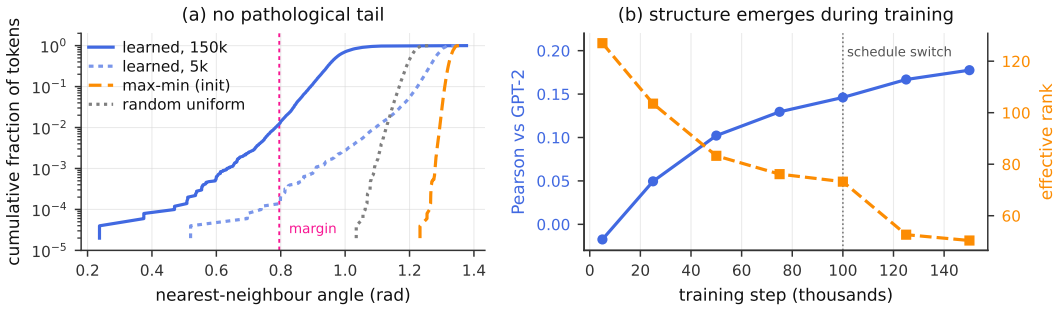

Figure 14: Geometry of the learned spherical codebook ($d = 128$). **(a)** Empirical CDF of nearest-neighbor (NN) angles at 5k and 150k updates, compared with frozen max-min and random uniform controls. Only 1.2% fall below the separation margin $\arccos(0.7) \simeq 0.7954$ radians. **(b)** GPT-2 cosine alignment increases while centered effective rank (participation ratio) decreases. The dashed line marks the log-to-entropy training-schedule switch at 100k; changes after it do not isolate the schedule's effect. This is a single-run diagnostic.

Table 49 reports the codebook diagnostics. The primary Pearson alignment metric correlates codebook and GPT-2 pairwise cosine similarities on two million uniformly sampled pairs over the full vocabulary. Only Pearson uses this pair sample. Top-10 overlap is evaluated over all 50,257 tokens shared with GPT-2, without token sampling: it is the fraction whose nearest codebook neighbor also appears among their ten nearest GPT-2 neighbors, using cosine similarity. Leading-space agreement measures whether a token and its nearest codebook neighbor share the same leading-space form. This token-form diagnostic gives 55.0% for both controls; semantic agreement is evaluated separately, and no analytical chance level is assumed.

Table 49: OWT codebook diagnostics ($d = 128$, 150k updates). Angles are in radians; rank is centered. Above-margin counts use NN cosine > 0.7, or angle $< \arccos(0.7) \simeq 0.7954$. Top-10 overlap uses all 50,257 shared tokens.

Metric	Learned	Max-min	Random
NN angle: median	0.970	1.327	1.202
NN angle: first percentile	0.784	1.294	1.128
NN angle: minimum	0.237	1.232	1.035
Codewords with NN cosine > 0.7	624	0	0
Centered effective rank	50.4	127.9	127.7
Mean pairwise cosine	0.041	0.00005	-0.00006
Pearson alignment with GPT-2	0.176	0.0003	0.0013
NN in GPT-2 top ten (%)	21.4	0.01	0.02
Leading-space agreement (%)	95.3	55.0	55.0

Organization and separation. The codebook becomes denser and anisotropic, while most codewords remain well separated: the first-percentile nearest-neighbor angle is 0.784 radians. Only 624 of 50,258 codewords (1.2%) have NN cosine above 0.7, or angle below $\arccos(0.7) \simeq 0.7954$ radians. This is the positive-argument region of the softplus regularizer in Equation 44, not the fraction with any nonzero loss: softplus also gives nonzero penalties below the margin. Median angular displacement from the initial max-min anchor is 63.7 degrees. Greater displacement is associated with smaller final nearest-neighbor angles (Spearman correlation -0.764). Frequency rank, with larger ranks denoting rarer tokens, correlates positively with nearest-neighbor angle (0.735): frequent tokens occupy denser regions. The above-margin codewords have median frequency rank 1,015, compared with 25,128 over the full vocabulary. These correlations suggest denser codebook allocation around frequent tokens, but are observational.

Emergence during training. Table 50 separates the training trajectory from the endpoint comparison. Between 5k and 100k, before the noise-draw change, effective rank decreases from 127.0 to 73.2 while Pearson alignment rises from -0.017 to 0.146 and top-10 overlap from 0.8% to 21.1%. At 5k, alignment and token-form diagnostics remain much closer to the controls than at later checkpoints. Further changes after 100k confound additional training with the switch to entropy-based noise draws. The pre-switch trajectory already shows progressive organization.

Table 50: Codebook organization through training in one OWT run. The noise draw changes from log to entropy at 100k; comparisons across that boundary do not isolate the effect of additional training. Angles are in radians, and overlap and form agreement are percentages.

Updates	NN median	Eff. rank	Pearson	GPT-2 top-10 (%)	Space agreement (%)
5k	1.279	127.0	-0.017	0.8	57.8
25k	1.160	103.5	0.050	6.7	79.1
50k	1.082	83.3	0.102	14.5	85.2
75k	1.055	76.2	0.130	18.9	86.0
100k	1.044	73.2	0.146	21.1	86.2
125k	0.970	52.6	0.167	21.3	94.8
150k	0.970	50.4	0.176	21.4	95.3

Sensitivity to pair sampling. Pearson alignment depends on which token pairs are evaluated (Table 51). We use uniform sampling over the full vocabulary as the primary reported metric. Frequency-weighted and restricted-vocabulary results serve as descriptive sensitivity checks alongside the primary headline metrics. The rise in alignment among frequent tokens is consistent with the density analysis.

Table 51: Sensitivity of GPT-2 pairwise cosine alignment to pair sampling for the 150k learned codebook. Uniform full-vocabulary sampling is the primary metric; restricted subsets are diagnostic comparisons.

Pair sampling	Pearson
Uniform, full vocabulary	0.176
Frequency-weighted by unigram prior	0.210
Uniform, 20,000 most frequent tokens	0.277
Uniform, 8,000 most frequent tokens	0.382
Uniform, 2,000 most frequent tokens	0.463

This single-run analysis documents organization of the learned codebook; its effect on generation quality remains unisolated. Figure 14 presents the empirical NN-angle CDFs and the alignment/rank trajectory.

E.14 Numerical validation of the heat kernel

These diagnostics evaluate the conditional corruption channel independently of a vocabulary or trained denoiser. They assess numerical consistency and the asymptotic analysis, while downstream language-model accuracy is evaluated separately.

Clock and semigroup. For $d \in \{32, 128, 768\}$, we estimate the retained mean cosine from forward draws and compare it with e^{-r}. Relative errors are reported only when the target is at least ten Monte Carlo standard errors above zero; otherwise the small denominator makes relative error misleading. This is a consistency check because the clock is defined by $r = -\log \rho$.

For the semigroup diagnostic, we compare a direct draw at $r_1 + r_2$ with two successive corruptions at r_1 and r_2. Each draw set contains 400k samples. We compare the empirical distributions of the final cosine relative to the initial codeword using the two-sample Kolmogorov-Smirnov distance,

$$D_{\mathrm{KS}} = \sup_c \left| \widehat{F}_{\mathrm{direct}}(c) - \widehat{F}_{\mathrm{sequential}}(c) \right|. \tag{65}$$

A null control compares two independent direct draw sets at the same radius and sample count. This estimates the finite-sample floor: a small distance is not evidence of exact agreement. Across the three dimensions and seven tested radius pairs, the distance is at most 0.023, but the worst discrepancies exceed the null floor. We therefore distinguish the exact analytic heat semigroup from its approximate numerical sampler.

Moment-matched heat-vMF comparison. We use the validated Heat tables described in Appendix D, at $d \in \{16, 32, 64, 128, 256, 512, 768, 1024\}$. At each operating point, the vMF concentration is chosen to match the heat law's mean cosine, using deterministic quadrature on the same cosine grid as the error integrals. This avoids using a Bessel-ratio approximation outside its stable range. The coordinate densities include the spherical volume factor $(1 - c^2)^{(d-3)/2}$ and are normalized before integration.

Let $f_H(c)$ and $f_V(c)$ be these cosine densities and $a_H(c) = \partial_c \log K^H_{d,r}(c)$. We measure the RMS relative radial-score error and the conditional channel KL:

$$R_{\mathrm{score}} = \left[\int_{-1}^{1} f_H(c) \left(\frac{a_H(c)}{\kappa^\star} - 1 \right)^2 \mathrm{d}c \right]^{1/2}, \tag{66}$$

$$D_{\mathrm{KL}}(H \| V) = \int_{-1}^{1} f_H(c) \log \frac{f_H(c)}{f_V(c)} \, \mathrm{d}c. \tag{67}$$

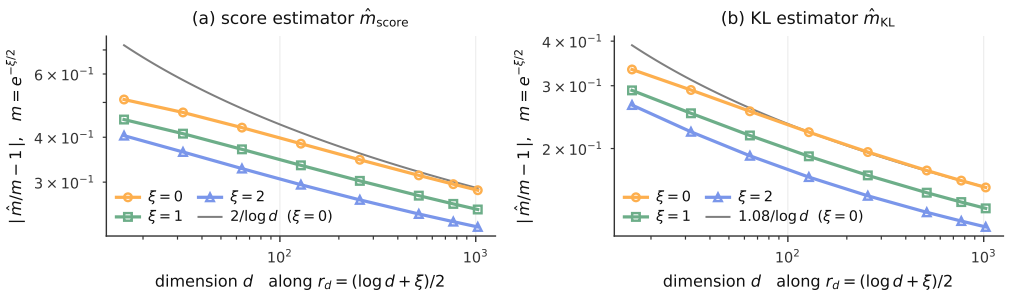

Figure 15: Numerical validation of Proposition 1 along $r_d = (\log d + \xi)/2$ for $\xi \in \{0, 1, 2\}$ and $d = 16$ to 1024. The score- and KL-based coefficient estimates approach the unfitted target $m = e^{-\xi/2}$; gray curves show $1/\log d$ guides.

The main scaling sweep follows $r_d = (\log d + \xi)/2$ for $\xi \in \{0, 1, 2\}$ and compares $\widehat{m}_{\text{score}} = R_{\text{score}}/t_d$ and $\widehat{m}_{\text{KL}} = (4D_{\text{KL}}/t_d^2)^{1/4}$ with $m = e^{-\xi/2}$. RMS is not divided by the Gaussian mean-absolute conversion factor; that factor applies only when starting from a mean-absolute error. We also report the fixed-r diagnostic separately because it is a different asymptotic regime. These are codebook-free conditional errors, not posterior KL under a learned vocabulary.

Figure 15 gives the numerical validation of Proposition 1: across eight dimensions and three critical-noise paths, both coefficient estimates approach the predicted target, with finite-dimensional discrepancies shrinking slowly.

F Technical estimates for the heat-vMF proof

This appendix supplies the analytic details used in Appendix C. We keep the same notation and fix ξ throughout. The bounds below justify differentiation and the likelihood Taylor expansion; the numerical sweep is not used in the proof.

F.1 Coordinate moments and concentration

In the dimension-free clock $r = (d-1)t/2$, the rescaled cosine obeys

$$\mathrm{d}U_r = -U_r\,\mathrm{d}r + \sqrt{\frac{2(d-U_r^2)}{d-1}}\,\mathrm{d}B_r, \qquad U_0 = \sqrt{d}. \tag{68}$$

At $r = r_d$, variation of constants gives

$$U_{r_d} = m + \int_0^{r_d} e^{-(r_d-s)}\sqrt{\frac{2(d-U_s^2)}{d-1}}\,\mathrm{d}B_s. \tag{69}$$

The martingale has quadratic variation at most $d/(d-1)$, so

$$\Pr_H(|U-m| \ge x) \le 2\exp\left[-\frac{d-1}{2d}x^2\right]. \tag{70}$$

The generator applied to u^j also gives the exact recursion

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}[U_r^j] = -\frac{j(d+j-2)}{d-1}\mathbb{E}[U_r^j] + \frac{j(j-1)d}{d-1}\mathbb{E}[U_r^{j-2}]. \tag{71}$$

Induction in j yields bounded moments of every fixed order. Solving the recursion explicitly for $j \le 4$ also gives

$$\mathrm{Var}_H[\mathrm{He}_2(U-m)] = 2 + O(t_d + d^{-1}). \tag{72}$$

The vMF coordinate density is uniformly strongly log-concave. Indeed, its negative log-density has

$$W_d''(u) = \frac{(d-3)(d+u^2)}{(d-u^2)^2} \ge \frac{d-3}{d}, \tag{73}$$

and its mean is m by construction. It therefore has a Gaussian tail of the same form as Equation 70, up to absolute constants.

F.2 Proof of the heat-Poisson estimate

Formally expanding the operator in Equation 14 to first order gives

$$\log \frac{K^H_{d,r_d}(u/\sqrt{d})}{P_{q_d,d}(u)} = -\frac{m^2 t_d}{2}\,\mathrm{He}_2(u-m) + \text{remainder}, \tag{74}$$

because direct differentiation of Equation 12 yields, on the central scale,

$$\frac{D(D-1)P_{q_d,d}(u)}{P_{q_d,d}(u)} = m^2\,\mathrm{He}_2(u-m) + O\!\left(\frac{1+|u|^4}{d}\right). \tag{75}$$

Here and below $\mathrm{He}_2(x) = x^2 - 1$.

The remainders in Lemma 1 also satisfy the following bound for some fixed polynomial exponent k and every sufficiently large fixed A, uniformly on $|u-m| \le A\sqrt{\log d}$, with

$$|R^H_d(u)| + |S^H_d(u)| \le C(t_d^2 + t_d/d)(1+|u|^k). \tag{76}$$

Proof of Lemma 1. Define the rescaled Poisson kernel

$$\Pi_d(z,u) := P_{z/\sqrt{d},d}(u) = \left(1 - \frac{z^2}{d}\right)\left(1 - \frac{2zu}{d} + \frac{z^2}{d}\right)^{-d/2}. \tag{77}$$

Thus $P_{q_d,d}(u) = \Pi_d(m,u)$. If $Z \sim \mathcal{N}(0,1)$ and $W_d = t_d/2 + i\sqrt{t_d}Z$, then

$$\mathbb{E}[e^{nW_d}] = e^{-n(n-1)t_d/2}. \tag{78}$$

Termwise expectation in the absolutely convergent zonal expansion therefore turns Equation 14 into the exact identity

$$K^H_{d,r_d}(u/\sqrt{d}) = \mathbb{E}\big[\Pi_d(me^{W_d}, u)\big]. \tag{79}$$

Indeed, $|q_d e^{W_d}| = q_d e^{t_d/2} < 1$ for all sufficiently large d.

We next control the analytic Taylor remainder. Put

$$f_{d,u}(s) = \frac{\Pi_d(me^s, u)}{\Pi_d(m,u)}. \tag{80}$$

For every fixed A, direct differentiation of Equation 77 gives fixed integers K and constants C_A such that, for $j \le 1$, $\ell \le 4$, and $|u-m| \le A\sqrt{\log d}$,

$$\big|\partial_u^j \partial_s^\ell f_{d,u}(\theta W_d)\big| \le C_A(1+|u|)^K \exp\big\{C_A t_d(1+|u|)(1+Z^2)\big\}, \qquad 0 \le \theta \le 1. \tag{81}$$

To verify this bound, note that $|me^{\theta W_d}| = O(1)$ and the two denominators in Equation 77 are bounded away from zero. Its logarithmic derivatives are therefore polynomial in u. For the ratio itself, taking real parts and using $1 - \cos(\theta\sqrt{t_d}Z) \le \theta^2 t_d Z^2/2$ gives the exponential factor in Equation 81. Since $t_d(1+|u|) = o(1)$ on this window, the right-hand side has a uniformly bounded Gaussian expectation after multiplication by $|W_d|^4$.

Taylor expansion through degree three, followed by expectation, uses

$$\mathbb{E}W_d = \frac{t_d}{2}, \qquad \mathbb{E}W_d^2 = -t_d + \frac{t_d^2}{4}, \qquad \mathbb{E}W_d^3 = -\frac{3t_d^2}{2} + \frac{t_d^3}{8}, \tag{82}$$

and $\mathbb{E}|W_d|^4 = O(t_d^2)$. Equation 81 then yields, also after one u derivative,

$$\frac{K^H_{d,r_d}(u/\sqrt{d})}{P_{q_d,d}(u)} = 1 - \frac{t_d}{2}B_d(u) + O\big(t_d^2(1+|u|)^K\big), \tag{83}$$

where

$$B_d(u) = \frac{[D(D-1)\Pi_d](m,u)}{\Pi_d(m,u)}, \qquad D = z\partial_z. \tag{84}$$

For clarity, the first-order coefficient follows because $f''_{d,u}(0) - f'_{d,u}(0) = B_d(u)$; all second- and third-moment contributions not shown explicitly are $O(t_d^2)$.

The coefficient is elementary. From

$$D \log \Pi_d(z,u) = \frac{dz(u-z)}{d-2zu+z^2} - \frac{2z^2}{d-z^2} \tag{85}$$

and $B_d = D^2 \log \Pi_d - D \log \Pi_d + (D \log \Pi_d)^2$ at $z = m$, expansion of the two denominators gives, for $j \le 1$,

$$\partial_u^j \big[B_d(u) - m^2 \operatorname{He}_2(u-m) \big] = O\!\left(\frac{1+|u|^4}{d} \right). \tag{86}$$

On the central window, $t_d B_d(u) = o(1)$. Applying $\log(1+x)$ to Equation 83, differentiating once, and using Equation 86 proves Equations 15, 16, and the pointwise bound in Equation 76.

It remains to pass from the central window to the heat-weighted norms. The moment recursion in Equation 71 gives $\sup_d \mathbb{E}_H |U|^p < \infty$ for every fixed p, and Equation 70 makes the complement of $|U-m| \le A\sqrt{\log d}$ smaller than any prescribed polynomial order when A is chosen large enough. We only need polynomial global bounds for the remainders and the radial score. We record them next.

Let $K_t(\mathbf{z}, \mathbf{e})$ denote the spherical heat kernel relative to uniform measure, with generator $\frac{1}{2}\Delta_{\mathbb{S}^{d-1}}$. Positivity of the spectral coefficients and $|P_{n,d}(c)| \le 1$ give

$$K_{t_d}(\mathbf{z}, \mathbf{e}) \le K_{t_d}(\mathbf{e}, \mathbf{e}) \le P_{q_d,d}(\sqrt{d}) = \exp(O(\sqrt{d})). \tag{87}$$

The Li-Yau parabolic Harnack inequality on the sphere (Li & Yau, 1986) gives

$$K_{t_d/2}(\mathbf{e}, \mathbf{e}) \le 2^{(d-1)/2} \exp\!\left(\frac{\operatorname{dist}(\mathbf{z}, \mathbf{e})^2}{t_d} \right) K_{t_d}(\mathbf{z}, \mathbf{e}). \tag{88}$$

Since $K_{t_d/2}(\mathbf{e}, \mathbf{e}) \ge 1$ and the spherical distance is at most π, Equations 87 and 88, together with the explicit Poisson formula, imply

$$\left| \log \frac{K^H_{d,r_d}(u/\sqrt{d})}{P_{q_d,d}(u)} \right| = O(d) \qquad (|u| < \sqrt{d}). \tag{89}$$

Thus the density remainder is polynomially bounded on its full support, and the sub-Gaussian tail upgrades its central L^2 and L^3 estimates to Equations 17.

For the differentiated tail, apply Hamilton's gradient estimate to $K_{t_d/2+s}(\mathbf{z}, \mathbf{e})$ for $0 \le s \le t_d/2$ (Kotschwar, 2007). Its spacetime maximum is $M_d = K_{t_d/2}(\mathbf{e}, \mathbf{e})$, and

$$M_d \le P_{\sqrt{q_d},d}(\sqrt{d}) = \exp(O(d^{3/4})). \tag{90}$$

After rescaling from generator Δ to $\frac{1}{2}\Delta$, Hamilton's estimate and Equation 88 give

$$|\nabla \log K_{t_d}(\mathbf{z}, \mathbf{e})|^2 \le \frac{C}{t_d} \log \frac{M_d}{K_{t_d}(\mathbf{z}, \mathbf{e})} \le \frac{Cd}{t_d}. \tag{91}$$

For $c = \langle \mathbf{z}, \mathbf{e} \rangle$ and $u = \sqrt{d}\, c$, radiality gives almost everywhere

$$|\partial_u \log K_{t_d}|^2 = \frac{|\nabla \log K_{t_d}|^2}{d(1-c^2)}. \tag{92}$$

On $|c| \le 1/2$, this is $O(t_d^{-1})$, so the sub-Gaussian tail controls the part outside the central window. On $|c| > 1/2$, using Equation 87 and the uniform-coordinate density gives

$$\mathbb{E}_H \big[|\partial_u \log K_{t_d}|^2 \mathbf{1}_{\{|C|>1/2\}} \big] \le \frac{C}{t_d} e^{C\sqrt{d}} \int_{|c|>1/2} \frac{\nu_d(\mathrm{d}c)}{1-c^2} \tag{93}$$

$$\le \operatorname{poly}(d) e^{C\sqrt{d}} \left(\frac{3}{4} \right)^{(d-5)/2} = o(d^{-B}) \tag{94}$$

for every fixed B, where ν_d is the uniform coordinate law of C. Finally, $\partial_u \log P_{q_d,d}(u) = O(1)$ globally, and the leading linear term in Equation 16 has the same sub-Gaussian tail. This proves Equation 18. Centering R_d^H changes its L^p norm by at most a factor of two, completing the proof. □

F.3 Explicit Poisson-vMF calibration

This part of the comparison is explicit. Let μ_d be the uniform coordinate law

$$\mu_d(\mathrm{d}u) = c_d \left(1 - \frac{u^2}{d}\right)^{(d-3)/2} \mathbf{1}_{\{|u|<\sqrt{d}\}}\, \mathrm{d}u, \tag{95}$$

and define its log-partition function

$$\psi_d(\lambda) = \log \int e^{\lambda u}\, \mu_d(\mathrm{d}u). \tag{96}$$

The exact even moments

$$\mathbb{E}_{\mu_d}[U^{2j}] = d^j \frac{(1/2)_j}{(d/2)_j} \tag{97}$$

give, uniformly for λ in compact sets,

$$\psi_d(\lambda) = \frac{\lambda^2}{2} - \frac{\lambda^4}{4d} + O(d^{-2}), \qquad \psi_d'(\lambda) = \lambda - \frac{\lambda^3}{d} + O(d^{-2}). \tag{98}$$

Writing $\lambda_d^\star = \kappa_d^\star/\sqrt{d}$, moment matching is $\psi_d'(\lambda_d^\star) = m$. Hence

$$\lambda_d^\star = m + \frac{m^3}{d} + O(d^{-2}), \qquad \kappa_d^\star = m\sqrt{d} + \frac{m^3}{\sqrt{d}} + O(d^{-3/2}). \tag{99}$$

For $q_d = m/\sqrt{d}$, direct expansion of Equation 12 gives

$$\log P_{q_d,d}(u) = mu - \frac{m^2}{2} + \frac{-m^2 + m^4/4 - m^3 u + m^2 u^2}{d} + O\!\left(\frac{1+|u|^3}{d^2}\right), \tag{100}$$

$$\log K^V_{d,\kappa_d^\star}(u/\sqrt{d}) = mu - \frac{m^2}{2} + \frac{m^3 u - 3m^4/4}{d} + O\!\left(\frac{1+|u|}{d^2}\right). \tag{101}$$

Subtracting and differentiating proves Equations 19 and 20. The deliberately loose polynomial envelopes in Equations 19 and 20 are convenient on the growing central region. On the full support, the explicit formulas give a bound polynomial in d for the log-ratio and its derivative. Splitting at $|u-m| = A\sqrt{\log d}$ and using the heat-law sub-Gaussian estimate proved in Equation 70 therefore upgrades the displayed remainders to $O(d^{-2})$ in the fixed $L^p(H)$ norms used below, after increasing A if necessary. No uniform Bessel asymptotic is required for this Poisson-vMF step.

F.4 Exponentially weighted likelihood tails

For completeness, the exponentially weighted likelihood tail can be controlled without assuming exponential moments of the residual. If $X_d = L_d - \mathbb{E}_H L_d$ and $G_d = \{|U-m| \le A\sqrt{\log d}\}$, then the exact change of measure gives

$$\mathbb{E}_H[e^{-X_d}\mathbf{1}_{G_d^c}] = e^{D_{\mathrm{KL}}(H\|V)} \Pr_V(G_d^c). \tag{102}$$

The prefactor is bounded. Spectral orthogonality and Equation 13 imply

$$\int (K^H_{d,r_d})^2\, \mathrm{d}\mu_d \le \sum_{n\ge 0} h_{n,d} q_d^{2n} = P_{q_d^2,d}(\sqrt{d}) = O(1). \tag{103}$$

Hence $\mathbb{E}_H \log K^H \le O(1)$, while $-\mathbb{E}_H \log K^V = -\kappa_d^\star \mathbb{E}_H C + \psi_d(\lambda_d^\star) = O(1)$. Thus $D_{\mathrm{KL}}(H\|V) = O(1)$. Equations 70, 73, and 102 make the exponentially weighted complementary term smaller than any required polynomial order by choosing A sufficiently large. In addition, Equation 22 gives $\|X_d\|_{L^3(H)} = O(t_d)$; Hölder's inequality and Equation 70 then control the linear and quadratic complementary terms. On G_d, the central expansions in Lemma 1 and Equation 19 give $\sup_{G_d} |X_d| = o(1)$, while the global L^3 bound controls the cubic Taylor remainder. Thus this argument requires no additional tail assumption.

G OpenWebText generated samples

We show the first 140 words of sample index 0 at each saved point of our 1M-update OWT model, without quality ranking or text edits. Temperatures select the nearest saved nondominated point to each target entropy. Reported H and GenPPL describe the source point, not the passage; different NFEs are separate generations, not revisions.

`[U+FFFD]` marks a decoder replacement character already in the output. These passages are not verified factual statements.

G.1 Target entropy $H = 4.5$ nats

Ours (1M), budget: 32 NFE $H_{\text{target}} = 4.5,\ k = 1$

$T = 0.70$; $H = 4.5213$ nats; GenPPL $= 29.17$; 32 calls; $n = 512$.

<|endoftext|> it."

At the conference, the designation attracted the support of the state's development player, Gov. Peter L. Turner, who said the state would be looking for a jack presence.

Mobhan said the move was part of the state's "undefastic vision" for the state. In that description, he noted discussions between the town of Baltimore Smith and the surrounding community the need to find local funding for the development.

He also referred to the cap of the amount that an entity can fund Pro. in the time they go to the state to serve on the U.S. Coast.

Mobhan said the had changed in the past.

Games and the end of its game

Burson said that some of the Moody's schools, and the future of the state's history, will be under pressure to fund the program.

Andy Jensen, the longtime

[Excerpt ends here.]

Ours (1M), budget: 64 NFE $H_{\text{target}} = 4.5,\ k = 1$

$T = 0.75$; $H = 4.4612$ nats; GenPPL $= 27.33$; 64 calls; $n = 512$.

<|endoftext|> it."

At the time, though, there was an army of loud tour musicians — "on the building, trying to get it from 6 a.m. to 3 a.m.," Dorander said.

"We don't know if we can keep to with it. Why do the club?"

One or two sessions discussing the tour with the rest of the announcement.

"When they say it was possibly terrorism, it's an invasion, which is a big deal. They just can't battle the street," she said.

Pam also wondered what happened the third day shaking the conductor's hand.

"A pile of thought it had to do with the rest of the band and the sound of it," she said.

"You don't know where people are going,' you say, 'We don't want a reason or how the rest of this happened.'

"It took us half a week

[Excerpt ends here.]

Ours (1M), budget: 128 NFE $H_{\text{target}} = 4.5,\ k = 1$

$T = 0.75$; $H = 4.4625$ nats; GenPPL $= 21.06$; 127 calls; $n = 512$.
<|endoftext|>.
"Most people said this happen at the end, but many don't want to get up the road … trying to get it at 3 a.m. to 7 a.m.," Mathic said.
"We don't know if it will happen, but then we will have to stay."
Meanwhile, buses were shutting around last year before the river, but it would be a longer time to go back, especially in the frozen foothills of the park.
The trio expects it will remain a project.
On Friday afternoon, after a chat, a friend of business, Shirley, made a formal opinion on the tunnel's construction.
"The road and the main technique, and the bridge itself, that's important for us," she said.
"You don't complain about it. You don't talk to church. We don't have homeless people a lot, but [under the road], we don't have
[Excerpt ends here.]

Ours (1M), budget: 256 NFE $H_{\text{target}} = 4.5,\ k = 1$

$T = 0.75$; $H = 4.6289$ nats; GenPPL $= 22.67$; 256 calls; $n = 512$.
<|endoftext|>.
600,000.
And the error rate is very bad, but we'll have to go to the Sun to know what a lot will go on yet.
And that:
Two Devapse from Earth's Sun
From the Sun eye that will begin, and the planets will begin to return to Earth, in a minute, then in two and a half of days.
This will be a pretty day, it will be to see the Sun, one of the Spicon. In two days it will see the Sun.
They will make it to the Sun orbit before view of the entire moon, in the great northern part of the Moon.
From there, they will see a moon will unfold between their galactic side and the northern side, much as in the northern part, and east.
The moon will be pointed to see the
[Excerpt ends here.]

Ours (1M), budget: 512 NFE $H_{\text{target}} = 4.5,\ k = 1$

$T = 0.75$; $H = 4.3232$ nats; GenPPL $= 16.33$; 512 calls; $n = 512$.
<|endoftext|> of 10,000,000.
Therefore the creation of the three months of creation, with that day of earth, will be constituted into account of its age, while, on the earth, It will consist of 10,000 years and shall be six days.
And the whole generation which will be composed of the ages, at a time of 70,000 and 50,000 days according to the names of one.
I can say that this generation the ages will be divided to the earth, lasting by 10,000 days, and sixty-three in a longer time.
That will be the whole period of the whole life of the earth seven, in the whole life of all the stars. It is, that the age between two years, all the earth, and the year thirty, can be as many as 34,000 years.
This earth will be formed in the
[Excerpt ends here.]

Ours (1M), budget: 1,024 NFE $H_{\text{target}} = 4.5,\ k = 1$

$T = 0.80$; $H = 4.5494$ nats; GenPPL $= 16.20$; 1,024 calls; $n = 512$.
<|endoftext|> day of October 11, 1955.
For the specific quantity of the operation of this period, which is not to say it is to be given to the manufacturer, or any of the services, may have been called the preceding medium, in provision to the company.
So there is an additional factor which is that the operation of the operation, the, after the date of October 1, 1955, prior to the beginning of the operation, appears on the ground as to the quantity of the plant, and the quantity, producing the water produced in this necessary quantity, and does not affect the relation between the water company and the plant, and in the important quantity of water which was produced in the plant where it was produced. Then there is a second factor, but one must hypothesise that there are
[Excerpt ends here.]

G.2 TARGET ENTROPY $H = 5.0$ NATS

Ours (1M), budget: 32 NFE $H_{\text{target}} = 5.0,\ k = 1$

$T = 0.80$; $H = 4.9733$ nats; GenPPL $= 46.68$; 32 calls; $n = 512$.

<|endoftext|> it."

Not only that, the move shows the support of the men's golf team. Rep. Peter L. Schumer, D-N.C., voted for the drawats.

Hobna said the move was part of the state's "irquastic project" for several reasons. In this case, he noted discussions with a group of state athletes and the community and the need to find local benefits for the team.

He also outlined in in the bonds the amount that an organization can pay. and in the time they get to the team they play on the U.S. Coast.

Hobna said this went out in the past.

Games into the General Master' game

Bash said that the growth and growth of interest in soccer, and the future of the state's pro league will be sorely positioned to support the change.

He said, the next group of legislators

[Excerpt ends here.]

Ours (1M), budget: 64 NFE $H_{\text{target}} = 5.0,\ k = 1$

$T = 0.85$; $H = 5.0222$ nats; GenPPL $= 46.51$; 64 calls; $n = 512$.

<|endoftext|> there."

At the Sunday, Saturday, there is an army of street music bands — "finding the building, trying to get it around 6 a.m. to 3 a.m.," Mathas said.

"We don't know if we can keep away with kids. We love the club."

Students and musician sessions break up a rally in the downtown, Friday.

Last July, Progressive announced it was interested in the Queen's Centre Project, which is a landmark venture. They were participating in what will tour the stations, offices and transport rooms.

Corza questioned what happened the teens an affirmation of Ontario's needs.

"These sorts of things have nothing to do with the values of the province and the values of all," she said.

"You don't know where people are going," she added, "We don't want to have a stage choir coming to this festival."

Korice

[Excerpt ends here.]

Ours (1M), budget: 128 NFE $H_{\text{target}} = 5.0,\ k = 1$

$T = 0.85$; $H = 4.9393$ nats; GenPPL $= 38.23$; 127 calls; $n = 512$.

<|endoftext|>.

"We have developed our attitude in the past, but we can't get our player up the road — trying to see what the model is — and tell them what the idea is," Casinson said.

"We don't know if it will happen, but certainly we will have to stay."

The are expected to set together a club in the Territory, but it would be a long way to go forward, locking down the finals snout against the Malans, although the club expects it might be a way.

At home Camp, in the season, a group of about 30 people, played the game at the small city's expense.

"We think about how the model works and for the club, that's important for us," Burke said.

"You don't know about it. You don't know about family. We didn't want a family

[Excerpt ends here.]

Ours (1M), budget: 256 NFE $H_{\text{target}} = 5.0,\ k = 1$

$T = 0.80$; $H = 4.9059$ nats; GenPPL $= 32.15$; 256 calls; $n = 512$.

<|endoftext|> in the long run, too.

Certainly the end round is very bad, but we'll have to go to the US to see if a lot will go on yet.<|endoftext|>Posted on by: Director Tao Lau- in Shanbe's Congress

A new training mission that starts today, with the crew in position to go to China, from Chinese coast, 2 or 5 times a day at home.

This will be on Thursday as the crew will begin to climb the main surface station to the International Space Station. The 400 men will then take a prototype.

They will communicate the signal the satellite system receives instead of the a metre, in this case the path of the boom. From there, they will be getting a testator between their balloon information and the water minimum as well – as they are working, and

[Excerpt ends here.]

Ours (1M), budget: 512 NFE $H_{\text{target}} = 5.0,\ k = 1$

$T = 0.90$; $H = 4.9996$ nats; GenPPL $= 30.93$; 512 calls; $n = 512$.

<|endoftext|> Thursday, October 7, 2017.

But the main mission achieved an almost exclusively visual point for that day. The mission will be planning to launch once a year, and by Wednesday the Lunar and Mars Mission rockets are expected to send operations for at least six weeks.

The number of spacecraft that have been launched on the moon rests on a scale between 17,000 and 8,000 days, with all the other ones.

Officials say the mission took the 4.6 miles long, 61-mile trip from Mars launched in 1976, about 250 miles away from home to Earth.

That will propel the closest mission to the icy view of Mars that has sailed through its previous journey of 32 years. The rover orbits its core about 1,300 miles away from the sun, and the trajectory of space can be as low as

[Excerpt ends here.]

Ours (1M), budget: 1,024 NFE $H_{\text{target}} = 5.0,\ k = 1$

$T = 0.90$; $H = 4.9800$ nats; GenPPL $= 25.60$; 1,024 calls; $n = 512$.

<|endoftext|>, on March 11, 2017.

In the global age of waste waste problems, waste, a method similar to the one used by U.S. homes and homes in Wilmington, Michigan, would have been a great way to try and respond to that problem. But there are multiple alternative advantages that would be a solution to the problem. American Waste Inc. argued, in April, that it could reduce the amount of waste waste and reduce waste levels of as much as 1.4 billion pounds. This would allow waste to be used in buildings or other areas that could be identified from waste.

"In waste, there is a tremendous amount of waste that is injection into the waste that you put in. So it creates a very simple environment, and it also means that there are very acidic volumes of the waste needed

[Excerpt ends here.]

G.3 Target entropy $H = 5.25$ nats

Ours (1M), budget: 32 NFE $H_{\text{target}} = 5.25,\ k = 1$

$T = 0.90$; $H = 5.2894$ nats; GenPPL $= 78.70$; 32 calls; $n = 512$.

<|endoftext|> it."

On Monday afternoon, Christie and appearing at a rally called d's on Pittsburgh. Rep. Peter L. Mallan (R) & Tenn., met with a Hall of business owners Larry Klolffer.

Bosan called fostering the "irquastic model" for competing people. Earlier this year, he would work with a group community committee to discuss the problem and the need to boost efforts.<|endoftext|>CAP CEO Aie Hansen highlighted an email carrying in the Federal 25 process that an attorney might speak with other local organizations if they were to the committee to serve on the B.C. Greens.

Mosbee said the group he was involved with other members had bought into itself.

"I kind of thought about the talent and amount of involvement we need, and the overall impact on Fair's (conference) leadership, plan to tackle this program. I believe we have to recruit

[Excerpt ends here.]

Ours (1M), budget: 64 NFE $H_{\text{target}} = 5.25,\ k = 1$

$T = 0.90$; $H = 5.2192$ nats; GenPPL $= 63.26$; 64 calls; $n = 512$.

<|endoftext|> communities."

At the weekend, residents and vendors — an event of town service weekly — "that the mayor is going to prioritize police from 6 a.m. to 3 a.m.," Rars said.

Rars said the council "probiously will stand along with people to antagonite the city."

Books and business libraries have up a quarterly in the city, weeks.

Last week, Progressive announced it was investing in the Queen's Centre Team, working with the SpotlightSciters to bring in what will become the stations, businesses and transit services.

Rater also questioned what caused the dramatic continued rise of Winnipeg's needs.

"These sorts of things have nothing to do with the here of the few and the future of life," she said.

"We don't know where parents are going," she added, "We don't want to have a football bus coming to this city."

[Excerpt ends here.]

Ours (1M), budget: 128 NFE $H_{\text{target}} = 5.25,\ k = 1$

$T = 0.95$; $H = 5.2645$ nats; GenPPL $= 60.77$; 127 calls; $n = 512$.

<|endoftext|>[U+FFFD]

The list of players with an emphasis is limited. "We can't do a lot of the field science processing to guess what a game is, and then learn what the game is," Matto said. "But, because we understand what it is, it will kind of make you down right back to games."

The coaches need to test building a game in every river, but it will take a longer time to collect player data. In the process, none of the competitors will fit. The team expects it will be able to move those young players and their leaders in a row, a group of roughly 17,000 players to play at the Wildcats' 14th games.

"We like thinking about the simulation experience and the simulation problem, that's important for our coaches," Matto said. "You don't know about it. You don't

[Excerpt ends here.]

Ours (1M), budget: 256 NFE $H_{\text{target}} = 5.25,\ k = 1$

$T = 0.90$; $H = 5.2485$ nats; GenPPL $= 53.62$; 256 calls; $n = 512$.

<|endoftext|> Canada until March 1, 2016.

But the exact assessment is unclear: 'We'll have to go to military experts to know what a solution is,' Army spokesman Col. Gen. John Dosman, told reporters' event.

Dosman cautioned that the importance of deploying the troops would be a problem to an officer who drinks 10, 25 or four days a day in Canada.

He also said this fall the usual plan would be to expand the soldiers base battalion to 10,000 people. But two troops would not meet that option.

That would give the Canadian National Armed Forces an advance of about 10,000 from the Canadian side of the base base. From there, that would mean connecting three training zones between the base base and the infantry reserve, each of which has about 80,000 people.

The detachment could require Wednesday to train

[Excerpt ends here.]

Ours (1M), budget: 512 NFE $H_{\text{target}} = 5.25,\ k = 1$

$T = 1.00$; $H = 5.2211$ nats; GenPPL $= 44.93$; 512 calls; $n = 512$.

<|endoftext|> results of March 1, 2011.

In August, NASA launched the JENN-2 spacecraft, the third spacecraft the agency has built to undertake such a mission, and by July the agency's 12STAR missions will deploy 50 satellite, communications and EVICs. They will provide a limited spacecraft which will enable the agency to transmit data about a mission of 17 neosetary objects, including Jupiter, to the future solar mission.

Jawn was collecting at the 4.6MEs at the west point of the joint probe in 2005. Just two years later, NASA's only space mission is working directly toward the OELISIII mission, a spacecraft that has been at its best stage for seven years. The telescope has an estimated 440 kilometers between nearby objects V and C rockets. At the top of the spacecraft will work an estimated 1,100 passengers, according to the joint

[Excerpt ends here.]

Ours (1M), budget: 1,024 NFE $H_{\text{target}} = 5.25,\ k = 1$

$T = 1.05$; $H = 5.2632$ nats; GenPPL $= 41.24$; 1,024 calls; $n = 512$.

<|endoftext|> price of $0.79. The approval for the drug is an indication that Shada Pharma, already under way for the drug announced by U.S. and European regulators in January, might come ahead with an outlier clinical drug, despite the fact that B. vulgari is used in mice, to produce clinical testing and discovery trial. Shada has inserted DIMA, a test for proteins found in human cells, bones and other tissues consisting of as many as 1.4 billion genes. It has been used to identify spread of different cells in many different samples, in this case meat, liver, brain and liver, and in a specific source of tissue that is involved in the human genome. "DIMA is a common molecule that occurs in the body, but it is involved in lung and in muscle cord cells and when grown in muscle

[Excerpt ends here.]

G.4 Target entropy $H = 5.5$ nats

At the 1,024-NFE budget, the default $k = 1$ sweep does not reach this entropy target. Its $k = 2$ example below is a separate sampler configuration and uses 64 source samples rather than 512.

Ours (1M), budget: 32 NFE $H_{\text{target}} = 5.5,\ k = 1$

$T = 1.00$; $H = 5.5393$ nats; GenPPL $= 145.56$; 32 calls; $n = 512$.

<|endoftext|> her."

ITITIONATION: dinner and showing at a rally called yaleshell on Sept. 24. - surprised to get what K&a.S. did with a Paganiku: ACT organizer Jeol Perkins.

Oerr says the reservation of the Kinderhaster is tough for numerous people. In this post, he gives examples with two community and local cultures discuss the problem with the horning bird rescue.<|endoftext|>ORITA, Md. (AP) — The Fed's Education Department is working with other local organizations so they can test the money they save on law enforcement residents and their businesses.

Director Jason Taylor said the group of around 5,000 members have lost their wounds.

"A lot of thought about the skills and amount of information we need, and the biggest danger from Capital Education center (ETA) program is if you don't define where you are or recruit any group of people with

[Excerpt ends here.]

Ours (1M), budget: 64 NFE $H_{\text{target}} = 5.5,\ k = 1$

$T = 1.00$; $H = 5.5185$ nats; GenPPL $= 131.02$; 64 calls; $n = 512$.

<|endoftext|>`[U+FFFD]`

A number of mayors told Providence,C., meeting at an event of town management Wednesday that "the executive director is going to define what B.B.C. would be like. Nothing," Dorell said.

Quence said the committee "propartiously will continue to identify the streets that be of residents."

Education and job libraries have become increasingly focus in the area, too.

In July, Richmond said it was interested in the District's Centre Act, which a same-party "strong coalition" will begin on infrastructure, businesses and transit services — a tactic Manager Patricia Patterson described as signaled the broad continued status of Clark's body.

"These sorts of decisions have nothing to do with the best public facilities . . . and the highest city services," she said.

"You don't know exactly who people are," she added, "you don't want to have a moment to go to [the

[Excerpt ends here.]

Ours (1M), budget: 128 NFE $H_{\text{target}} = 5.5,\ k = 1$

$T = 1.05$; $H = 5.5284$ nats; GenPPL $= 123.09$; 127 calls; $n = 512$.

<|endoftext|>`[U+FFFD]`

The list of people with positive stature is limited. "We can't do a lot of research and then try to guess what a gene most people do before they know the person is," Paulas said. "But then, to understand what it is, it is kind of becoming just silent out in a situation."

So, people may look like a god once every century, but it would take a long time to recover their DNA. But the looming potential field of aging people looks complex. The team hopes it might be better to move these young people from their neighbors in a college, a group of maybe college individuals, and to collect data on others' feelings and preferences.

"We can learn about the intro arts and the technical community, that feed into using data for their insight," Paulas says. "You

[Excerpt ends here.]

Ours (1M), budget: 256 NFE $H_{\text{target}} = 5.5,\ k = 1$

$T = 1.05$; $H = 5.4690$ nats; GenPPL $= 118.06$; 256 calls; $n = 512$.

<|endoftext|> service until March 15, 2016.

And the sad thing is that Yahoo doesn't have to have to limit it, with never wanting to be so trusted and not yet.<|endoftext|>Posted on:: Joe Tao McDonald/WassTV's Flickr

-- PT is Radio One LIVE. View the form in here and listen to this title stream from 11,0595, for a moment of change.

FRE announced this week the radio station will be equipped with the AM surface monitor for fast resonation speeds. With two days already end, the OEMIADRAN satellites will test the radio components for the specific purpose of designing a field designed for the best synchronization of weather weather studies. They will begin with a satellite measuring two satellites added into their own definitions, an ocean receiving magnetom objects as high as 15,000 miles.

The station will provide the radio stations and

[Excerpt ends here.]

Ours (1M), budget: 512 NFE $H_{\text{target}} = 5.5,\ k = 1$

$T = 1.10$; $H = 5.4565$ nats; GenPPL $= 74.59$; 512 calls; $n = 512$.

<|endoftext|>, in particular26,27.

With the single thermore molecule BALIS2, a two-dimensional experimental model was developed to design such a compound. In October 1979 the German Regiwald University AMIRDEER/AMICE collaboration Otto A. Fagokamp conducted a demonstration of concrete, the mechanism for converting acid from a reaction that has six hydrogen molecules combined, namely the chemical unit (like carbon release buckhanine, for example), that store and store in lighting, helium-like material for helium molecules in contact.25 Six years later when Heinrich Erbmann developed the model for the AMIRDEER/AMICE collaboration, he resulted in his most experimental experiment for screening,28 which was the first chemical material for an meturous ion compound. The first product of the new approach was this zinc atom, which was being optimized for the "clinging electrode", but not yet bound to the interpean crystals.29 Example of the compound 87236-814 which

[Excerpt ends here.]

Ours (1M), budget: 1,024 NFE $H_{\text{target}} = 5.5,\ k = 2$

$T = 1.15$; $H = 5.5074$ nats; GenPPL $= 97.11$; 1,023 calls; $n = 64$.

<|endoftext|> proteins are amassing among genes. This findings provide a potential multam of B1-related gene expression indicating that the two homotypes are T, prima and prima6. There are also several subtypes, resulting in each B1 gene sequence. T, prima3,4 and 2.5 are identical to TGP. Aimensional OR of B1 gene genes identified in these results reveal cross-forming between LAT and F to protein regulator as well as regulator and allinetics. These results indicate a widely shared search for T or B1 gene variants. 3. Functional findings These findings include the presence of three B1-specific T intermedi groups that were identified in the same template that were selected for proteins belonging to a single-element gene. The results of this duplication are described in laboratory and outside sequence findings g. Stice coformation is more prone to adstian deformation alteration than in activated

[Excerpt ends here.]