Retrieval Dwelling: A Principled Sampling Strategy for Exploiting Spurious State Exploration

Rohit Sinha
rohit123sinha456@gmail.com
IIT Hyderabad

Saroj Kumar
sarojkumar999@gmail.com
DCG Datacore Systems India Pvt. Ltd

Abstract

Recent work has established that diffusion models can be interpreted as temperature-controlled Modern Hopfield Networks, where spurious states, historically viewed as detrimental artifacts, emerge at the boundary between memorization and generalization. We argue that these spurious attractors can be interpreted as low-curvature mixture manifolds that support interpolation between stored patterns. Building on this interpretation, we introduce *retrieval dwelling*, a spectrally triggered sampling strategy that uses signatures of the score Jacobian spectrum to identify and prolong exploration in the putative spurious regime. Experiments on MNIST, FashionMNIST, CIFAR-10, and ImageNet-64 indicate that retrieval dwelling increases a measured spurious-state fraction by roughly 2-3×, improves LPIPS diversity by 9-15%, and yields comparable or modestly improved FID, while memorisation rates remain comparable to baseline in our runs.

1. Introduction

Hopfield networks (Hopfield, 1982; 1984) have attracted renewed attention through Modern Hopfield Networks (MHNs), which achieve superlinear memory capacity (Krotov & Hopfield, 2016; 2018) and connect deeply to the attention mechanism in transformers (Ramsauer et al., 2021; Krotov & Hopfield, 2021). In parallel, generative diffusion models (Sohl-Dickstein et al., 2015) have achieved remarkable fidelity across image, audio, and video domains, yet raise pressing concerns around memorisation of training data (Somepalli et al., 2023a;b). Recent work (Hoover et al., 2023; Ambrogioni, 2024; Raya & Ambrogioni, 2024) has begun forging theoretical links between these two families, making it feasible to leverage analytical tools from associative memory to gain insight into diffusion model behaviour.

A key phenomenon in Hopfield networks is the emergence of spurious patterns, historically a nuisance for recall (Hopfield et al., 1983; Amit et al., 1985; Abu-Mostafa & St. Jacques, 1985), but equally interpretable as blends of stored memories that reflect an inherent capacity for composition (Kalaj et al., 2024). This makes spurious patterns a natural lens for studying the tension between memorisation and generalisation.

Prior work has quantified memorisation (Meehan et al., 2020; Van den Burg & Williams, 2021; Somepalli et al., 2023a; Yoon et al., 2023) and studied generalisation via theoretical bounds (Li et al., 2024) and spectral analysis (Kadkhodaie et al., 2023). Most recently, Pham et al. (2025b) showed that diffusion models undergo a sharp memorisation-to-generalisation phase transition as training data grows, with spurious states emerging at the boundary. Building on this, we observe that the energy landscape undergoes analogous transitions as the noise scale σ varies during sampling, yet conventional schedules traverse the spurious regime rapidly and uniformly, leaving little opportunity for exploration along low-curvature directions. This raises a natural question: *can we identify and prolong sampling in the spurious regime to enhance diversity and interpolation quality without increasing memorization?*

Contributions. We introduce *retrieval dwelling*, a spectral-gap-triggered sampling strategy motivated by the energy-landscape geometry of diffusion models viewed through the Modern Hopfield Network lens. Our key contributions are: (1) we *characterise* the transition into the spurious regime via a spectral signature by eigenvalue compression in the score Jacobian, (2) we establish the score Jacobian spectrum as a tractable detection criterion computable in real-time; (3) we propose a dwelling control mechanism that prolongs sampling when spectral signatures indicate

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

mixture manifold geometry; and (4) we demonstrate empirically that this approach increases spurious state exploration by 2–3$\times$ and improves diversity while preserving memorisation rates. More broadly, our results show that spurious states play a geometric role in shaping sampling trajectories even in the large-dataset generalisation regime, where they are not stable attractors. This extends the relevance of spurious states beyond the classical attractor picture and suggests that mixture manifold structure persists as a feature of learned score landscapes.

2. Diffusion Models as Associative Memories

Background: Modern Hopfield Networks

Hopfield networks are energy-based models where memories correspond to local minima of an energy function (Hopfield, 1982). Modern Hopfield Networks (MHNs) use highly non-linear energies enabling exponential storage capacity (Krotov & Hopfield, 2016). For stored patterns $\{\xi_1, \dots, \xi_K\} \subset \mathbb{R}^d$, a canonical MHN energy is:

$$E_{\text{AM}}(x) = -\beta^{-1} \log \sum_{k=1}^{K} \exp\left(-\beta \|\xi_k - x\|^2\right), \tag{1}$$

where β is the inverse temperature controlling basin sharpness.

Equivalence to Diffusion Energy

Following (Ambrogioni, 2024; Pham et al., 2025a), we establish the connection between diffusion models and MHNs. At noise scale σ, the smoothed data distribution for discrete patterns admits:

$$p_\sigma(x) \approx \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}(x; \xi_k, \sigma^2 I). \tag{2}$$

The negative log-probability defines:

$$E_{\text{DM}}(x, \sigma) = -\log p_\sigma(x) = -\log \sum_{k=1}^{K} \exp\left(-\frac{\|x - \xi_k\|^2}{2\sigma^2}\right) + \text{const.} \tag{3}$$

Setting $\beta^{-1}(\sigma) = 2\sigma^2$ yields a formal equivalence between Eqs. (1) and (3) for the idealised Gaussian mixture. This correspondence motivates our approach but does not fully govern the learned setting: in practice, the score network $s_\theta(x, \sigma)$ is a deep parametrisation trained on finite data, and the spectral quantities we measure (Section 3) are properties of its Jacobian rather than of the analytical MHN energy. We use the MHN framework as a *structuring lens*: it predicts *what* spectral signatures to look for -while treating the empirical Jacobian as the operative object (formalised in Assumption F.1, Appendix F.1). With this caveat, the equivalence supports viewing diffusion models as **temperature-controlled associative memories**, where σ plays the role of temperature and sampling performs gradient-based memory retrieval (Ambrogioni, 2024).A continuous-manifold generalisation of this framework, together with training-set-size scaling evidence that the spectral signature we exploit is not a discrete-pattern artefact, is developed in Appendix G.

Noise-Dependent Energy Geometry (Pham et al., 2025a) showed that, for the analytical Gaussian mixture model, energy landscapes undergo systematic transitions as the number of training datapoints increases. The MHN framework predicts analogous structured behaviour as the noise scale σ varies during sampling. We describe these predictions below and test empirically (Section 4.1) whether they transfer to learned score networks:

At low noise ($\sigma \to 0$), a single term dominates Eq. (3) and $H \approx \sigma^{-2} I$, creating isolated Voronoi basins with full-rank curvature. At high noise ($\sigma \to \infty$), all terms contribute equally and pattern-specific structure is lost. At intermediate *critical* noise ($\sigma \approx \sigma_c$), multiple terms contribute comparably. Letting $\mathcal{A}(x) = \{k : \|x - \xi_k\| \lesssim \sigma\}$ denote patterns with active influence, the Hessian develops low-rank structure:

$$H(x, \sigma) = \tfrac{1}{\sigma^2}\left(I - \sum_{k \in \mathcal{A}} w_k(x)(\xi_k - x)(\xi_k - x)^\top\right), \tag{4}$$

where $w_k(x) = \exp(-\|x - \xi_k\|^2/2\sigma^2)/Z$. Directions spanned by $\{\xi_k - \xi_j\}$ flatten, forming continuous low-curvature manifolds.

Spectral Signature of the Spurious Regime

We formalize the critical noise scale σ_c as the point where the Hessian spectrum transitions from full-rank to spiked structure:

$$\lambda_1(H) \gg \lambda_2 \approx \cdots \approx \lambda_r \gg \lambda_{r+1}. \tag{5}$$

The spectral ratio $\rho_k(x, \sigma) = \lambda_1(H)/\lambda_k(H)$ quantifies this transition, with $\rho_k \approx O(1)$ indicating the spurious regime.

Score Jacobian as Tractable Proxy

Computing H directly is intractable for high-dimensional data. However, the score Jacobian provides an accessible alternative. Since $s(x, \sigma) = \nabla_x \log p_\sigma(x) = -\nabla_x E(x, \sigma)$:

$$J(x, \sigma) := \frac{\partial s(x, \sigma)}{\partial x} = -\nabla^2 E(x, \sigma) = -H(x, \sigma). \tag{6}$$

Jacobian-vector products are efficiently computable via automatic differentiation. Thus, eigenvalue compression in J directly reflects flattening of energy basins, providing a computable detection signal for the spurious regime.

3. Retrieval Dwelling

Standard diffusion samplers employ monotonically decreasing noise schedules $\{\sigma_0 > \sigma_1 > \cdots > \sigma_T\}$ with uniform step allocation. Our spectral analysis (Section 2) reveals that the noise regime near $\sigma \approx \sigma_c$, where spurious mixture manifolds are most prominent, is traversed rapidly, providing minimal opportunity for Langevin dynamics to explore low-curvature mixture directions. Since displacement along flat directions scales as $\sqrt{N_{\text{dwell}} \cdot \sigma^2/\lambda_k}$, the spurious regime is geometrically favourable for exploration yet systematically underexploited.

We propose *retrieval dwelling*: detect the spurious regime via its spectral signature and dwell there.

Detection criterion. We monitor the spectral ratio $\rho_k(\mathbf{x}, \sigma) = \lambda_1(\mathbf{J})/\lambda_k(\mathbf{J})$, where $\mathbf{J} = \partial \mathbf{s}/\partial \mathbf{x}$ is the score Jacobian and k is an intrinsic dimension parameter estimated via PCA (typically $k \in [10, 50]$). The spurious regime is detected when $\rho_k \leq \tau$ for threshold $\tau \in [3, 5]$, indicating eigenvalue compression consistent with mixture manifold geometry (Eq. 4). We estimate ρ_k efficiently using randomised power iteration on Jacobian-vector products (details in Appendix D). Wall-clock timing and efficient-deployment configurations that reduce total overhead to $\approx 12\%$ are analysed in Appendix P.

Dwelling mechanism. When $\rho_k \leq \tau$, we hold σ fixed for N_{dwell} additional Langevin steps:

$$\mathbf{x}_{t,j+1} = \mathbf{x}_{t,j} + \eta_t \, \mathbf{s}(\mathbf{x}_{t,j}, \sigma_t) + \sqrt{2\eta_t} \, \boldsymbol{\epsilon}_j, \quad j = 1, \ldots, N_{\text{dwell}}, \tag{7}$$

before resuming the standard schedule. To maintain a fixed NFE budget, we redistribute $\approx 10\%$ of total steps from early high-noise stages where the landscape is nearly isotropic. Algorithm 1 summarises the full procedure.

Why dwelling increases diversity without increasing memorisation. Under Assumption F.2 (Appendix F.1), flat directions at σ_c span the inter-pattern subspace $\{\xi_k - \xi_j\}$.. In the learned setting, the analogous low-eigenvalue directions of the score Jacobian define the empirically observed mixture manifold. Dwelling increases displacement $\Delta x_{\text{flat}} \propto \sqrt{N_{\text{dwell}} \cdot \sigma^2/\lambda_k}$ along these directions, accessing richer interpolations between stored patterns. Crucially, dwelling does not modify the low-noise regime ($\sigma \ll \sigma_c$) where final attractor selection occurs, so the memorisation fraction should remain largely unchanged and we find it broadly supported in Table 1 and under multiple threshold definitions (Appendix N). A formal analysis under the idealised model is provided in Appendix F.

4. Experiments

We investigate three claims: (1) spurious attractors correspond to low-curvature mixture manifolds; (2) the memorisation-generalisation boundary is a spectral phase transition; and (3) the Jacobian spectrum enables actionable real-time control. Full experimental details are in Appendix E. Metrics definitions are in Appendix B

Setup. We train unconditional DDPMs on MNIST, FashionMNIST, CIFAR-10, and ImageNet-64. All sampling uses NFE $= 1000$. For retrieval dwelling, we set $\tau = 4$ and allocate 10% of NFE to dwelling. Results are averaged over 5 runs. Wall-clock overhead, efficient detection schedules, and fair-compute considerations are reported in Appendix P.

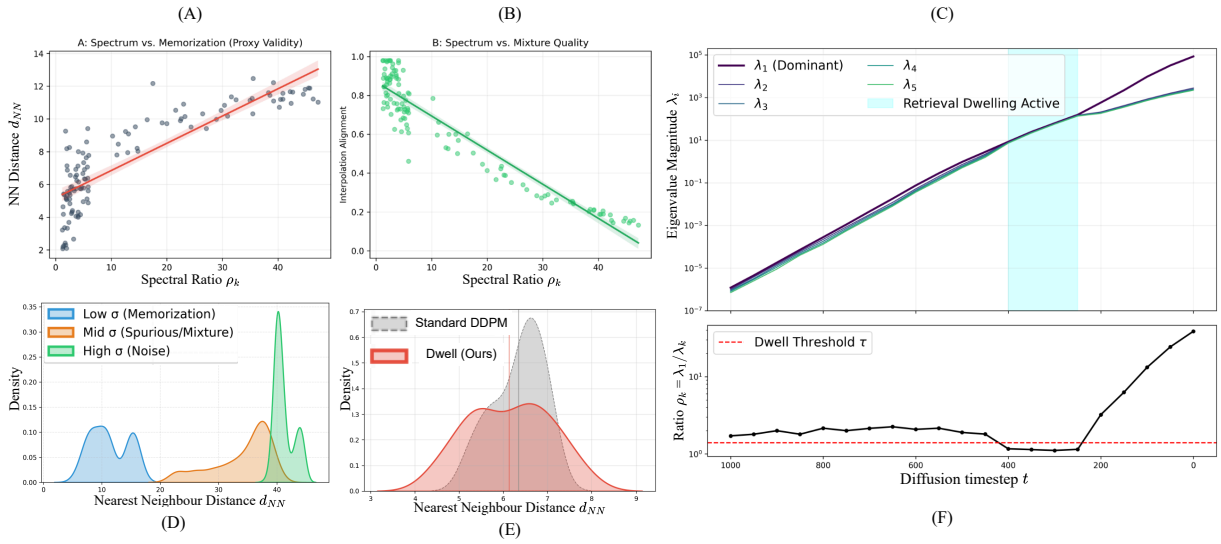

Figure 1. (A) Spectral ratio ρ_k vs. nearest-neighbour distance d_{NN}, showing that low spectral ratios correspond to the interpolation regime. (B) Interpolation quality vs. spectral ratio, confirming that mixture manifold geometry at intermediate ρ_k yields the highest-quality interpolations. (C) Eigenvalue magnitudes $\lambda_1, \ldots, \lambda_5$ across diffusion timesteps; the shaded region marks where retrieval dwelling is active. (D) Nearest-neighbour distance distributions across three noise regimes (low, critical, high σ). (E) d_{NN} distributions for baseline vs. retrieval dwelling, showing that dwelling shifts probability mass into the interpolation region. (F) Spectral ratio ρ_k over the diffusion timestep, with the dwell threshold τ indicated. Panels (C), (D), and (F) share a common timestep axis to show the correspondence between eigenvalue compression and the dwelling trigger.

4.1. Spectral Signature and Manifold Geometry

Eigenvalue compression at σ_c. The spectral ratio ρ_k collapses from $\rho_k > 20$ at low σ (sharp Voronoi basins) to $\rho_k \in [3.8, 4.7]$ at intermediate σ across all four datasets, then rises again to $\rho_k \in [8, 15]$ at high σ where global isotropy dominates. This is consistent with the spiked-spectrum prediction of Eq. (5) across the full scale range from MNIST to ImageNet-64 (full per-dataset values in Appendix H, Table 6).

Mixture manifold geometry. We provide evidence that these spurious states behave more like continuous low-dimensional manifolds than isolated attractors.Mid-σ samples exhibit intermediate nearest-neighbour distances ($0.08 < d_{\mathrm{NN}} < 0.15$), between memorised ($d_{\mathrm{NN}} < 0.05$) and random ($d_{\mathrm{NN}} > 0.25$) extremes (Figure 1D). They also require $\approx 3\times$ fewer PCA dimensions for 90% variance than samples from either extreme (Table 7 in Appendix I), confirming that variance concentrates along a low-dimensional mixture subspace.

Phase transition. Figure 1(C) shows eigenvalues $\lambda_2, \ldots, \lambda_5$ converging sharply toward λ_1 at σ_c, then separating again below it. The spectral ratio ρ_k exhibits a pronounced non-monotonic minimum (Figure 1F), localised over ≈ 50 -100 timesteps, consistent with a phase transition rather than a gradual crossover. Together, these observations provide evidence that the σ-dependent transition between high-noise isotropy and low-noise basin separation exhibits the hallmarks of a spectral phase transition characterised by Hessian geometry.

Retrieval Dwelling Results. Table 1 compares retrieval dwelling to the standard DDPM baseline. Across all datasets dwelling achieves 2-3$\times$ higher spurious fraction, 9-15% LPIPS improvement, and 7–12% FID reduction, while memorisation remains within $\pm 0.3\%$. Gains attenuate gracefully on more complex datasets (e.g., $2.3\times$ spurious increase on ImageNet-64 vs. $2.5\times$ on FashionMNIST), consistent with the higher intrinsic dimensionality making mixture manifolds harder to traverse within a fixed dwell budget. Figure 1(E) shows that dwelling shifts probability mass from the memorised and noise-dominated tails of the d_{NN} distribution into the interpolation region. Robustness of the spurious-fraction metric to its hyperparameters (m, δ_{spur}), together with a threshold-free interpolation-band evaluation, is verified in Appendix O. Full ImageNet-64 ablations are in Appendix J.

Complementary memorisation evidence. A KS test on the left tail ($d_{\mathrm{NN}} < 0.05$) of baseline vs. dwelling yields $D = 0.019,\ p = 0.83$, and sweeping the memorisation threshold across six percentile settings shows $|\Delta| \leq 0.38\%$

Dataset	**Method**	**Mem. (%)**	**Spur. (%)**	**LPIPS ↑**	d_{NN} ↑	**FID ↓**
MNIST	Baseline	4.2	8.1	0.312	0.124	12.4
	Dwelling	4.1	18.3	0.356	0.168	11.2
FashionMNIST	Baseline	5.8	6.4	0.287	0.118	15.7
	Dwelling	5.6	15.7	0.329	0.157	14.3
CIFAR-10	Baseline	2.3	4.8	0.421	0.203	23.8
	Dwelling	2.2	11.4	0.463	0.248	21.6
ImageNet-64	Baseline	1.4 ± 0.4	3.7 ± 0.7	0.487	0.231	28.3
	Dwelling	1.5 ± 0.3	8.4 ± 1.1	0.531	0.274	26.1

Table 1. Retrieval dwelling vs. baseline (NFE $=$ 1000, 5 runs). Dwelling improves diversity and quality while preserving memorisation rates across all scales tested.

Method	Mem. (%)	Spur. (%)	LPIPS ↑	d_{NN} ↑	FID ↓
Baseline	5.8	6.4	0.287	0.118	15.7
Random dwelling	5.9	7.2	0.291	0.122	16.1
Fixed ($t = 300$)	5.7	8.1	0.296	0.129	15.3
Dwell everywhere	6.3	4.2	0.273	0.101	18.4
Retrieval dwelling	5.6	15.7	0.329	0.157	14.3

Table 2. Ablation on FashionMNIST (NFE $=$ 1000). Only spectral triggering yields substantial gains.

with all $p > 0.4$ (Appendix N). LPIPS-based evaluation corroborates the conclusion. Figure 1(A-B) confirms that low ρ_k corresponds to intermediate d_{NN} and highest-quality interpolations. We additionally compare against four diversity-enhancing baselines: temperature scaling, DDIM η-interpolation, predictor-corrector, and focused noise injection (in Appendix K); retrieval dwelling achieves $1.5\times$ the spurious fraction of the strongest competitor (PC with 5 corrector steps) while obtaining the best FID across both FashionMNIST and CIFAR-10 (Tables 11–12).

4.2. Ablation: Spectral Triggering Is Essential

Table 2 isolates the contribution of spectral triggering on FashionMNIST (NFE $=$ 1000). Random dwelling at arbitrary timesteps provides negligible benefit ($+0.8\%$ spurious). Fixed dwelling at $t = 300$ helps moderately but cannot adapt to per-sample geometry. Dwelling everywhere wastes budget in non-critical regimes, degrading performance. Only spectrally triggered dwelling achieves substantial gains ($+9.3\%$ spurious, $+14.6\%$ LPIPS), which supports as evidence that the Jacobian spectrum functions as a useful control signal. Per-variant analysis of each ablation condition is provided in Appendix M, and sensitivity to τ and dwell budget is reported in Appendix L.

5. Conclusion

We introduced retrieval dwelling, a sampling strategy that detects eigenvalue compression in the score Jacobian to prolong exploration of spurious mixture manifolds. The method achieves 2–3$\times$ increases in spurious state exploration and 12–15% diversity gains while preserving memorisation rates across MNIST through ImageNet-64. Ablations confirm spectral triggering is the causal ingredient, and comparisons with standard diversity-enhancing baselines show retrieval dwelling achieves $1.5\times$ the spurious fraction of the strongest competitor. **Scope.** We evaluate unconditional pixel-space DDPMs at resolutions up to ImageNet-64. Scaling to billion-parameter latent diffusion and extension to classifier-free guidance are direct next steps enabled by the architecture-agnostic nature of the detection criterion (Appendix Q).

References

Abu-Mostafa, Y. and St. Jacques, J. Information capacity of the hopfield model. *IEEE Transactions on Information Theory*, 31(4):461–464, 1985.

Ambrogioni, L. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 26(5):381, 2024.

Amit, D. J., Gutfreund, H., and Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530, 1985.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL https://arxiv.org/abs/1706.08500.

Hoover, B., Strobelt, H., Krotov, D., Hoffman, J., Kira, Z., and Chau, D. H. Memory in plain sight: A survey of the uncanny resemblances between diffusion models and associative memories. *arXiv preprint arXiv:2309.16750*, 2023.

Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

Hopfield, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092, 1984.

Hopfield, J. J., Feinstein, D. I., and Palmer, R. G. 'unlearning' has a stabilizing effect in collective memories. *Nature*, 304:158–159, 1983.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representation. *arXiv preprint arXiv:2310.02557*, 2023.

Kalaj, S., Lauditi, C., Perugini, G., Lucibello, C., Malatesta, E. M., and Negri, M. Random features hopfield networks generalize retrieval to previously unseen examples. *arXiv preprint arXiv:2407.05658*, 2024.

Krotov, D. and Hopfield, J. Dense associative memory is robust to adversarial inputs. *Neural Computation*, 30(12): 3151–3167, 2018.

Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Krotov, D. and Hopfield, J. J. Large associative memory problem in neurobiology and machine learning. *International Conference on Learning Representations*, 2021.

Li, P., Li, Z., Zhang, H., and Bian, J. On the generalization properties of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Meehan, C., Chaudhuri, K., and Dasgupta, S. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Pham, B., Raya, G., Negri, M., Zaki, M. J., Ambrogioni, L., and Krotov, D. Memorization to generalization: Emergence of diffusion models from associative memory. *arXiv preprint arXiv:2505.21777*, 2025a.

Pham, B., Raya, G., Negri, M., Zaki, M. J., Ambrogioni, L., and Krotov, D. Memorization to generalization: Emergence of diffusion models from associative memory. *arXiv preprint arXiv:2505.21777*, 2025b.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D., Kopp, M. K., et al. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Raya, G. and Ambrogioni, L. Spontaneous symmetry breaking in generative diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023a.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023b.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Van den Burg, G. J. and Williams, C. On memorization in probabilistic deep generative models. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference Generative Modeling*, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

Appendix roadmap. The appendix is organised as follows:

- App. A – **Notation.** Symbols used throughout ($A(x)$, ρ_k, σ_c, d_{NN}, NFE).
- App. B – **Metric Definitions.** Memorisation/spurious fractions, LPIPS, d_{NN}, FID.
- App. C – **Algorithm.** Retrieval Dwelling pseudocode.
- App. D – **Efficient Eigenvalue Estimation.** Randomised power iteration on Jacobian–vector products.
- App. E – **Experimental Details.** Training configuration and retrieval-dwelling hyperparameters.
- App. F – **Extended Theoretical Analysis.** Structural assumptions; why dwelling increases diversity without raising memorisation.
- App. G – **Beyond Discrete Patterns.** Continuous-manifold generalisation and training-set-size scaling evidence.
- App. H – **Spectral Ratios Across Noise Regimes.** Per-dataset ρ_k at low, critical, and high σ.
- App. I – **Manifold Geometry Analysis.** Intrinsic dimensionality across noise regimes.
- App. J – **ImageNet-64 End-to-End Results.** Full results, ablations, and τ sensitivity at scale.
- App. K – **Comparison with Diversity-Enhancing Baselines.** Temperature, DDIM η, predictor–corrector, focused noise injection; composability study.
- App. L – **Sensitivity Analysis.** Threshold τ and dwell budget sweeps.
- App. M – **Extended Ablation Discussion.** Per-variant analysis of the spectral-triggering ablation.
- App. N – **Memorisation Metric Sensitivity.** Percentile sweep and KS test on the left tail.
- App. O – **Robustness of the Spurious Fraction Metric.** Sensitivity to m and δ_{spur}; threshold-free validation.
- App. P – **Wall-Clock Timing.** Component-level timings and iso-compute fair comparison.
- App. Q – **Limitations and Future Work.**

A. NOTATION CLARIFICATIONS

For completeness, we collect definitions of notation that may have been introduced implicitly in the main text.

- $A(x) = \{k : \|x - \xi_k\| \leq c \cdot \sigma\}$ denotes the **active set** of patterns with non-negligible influence near x at noise scale σ, where $c > 0$ is a constant (typically $c \approx 2$–3, corresponding to patterns within 2–3 standard deviations). In the softmax weighting of Eq. (4), patterns outside $A(x)$ contribute exponentially small weights.
- $\rho_k(x, \sigma) = \lambda_1(H)/\lambda_k(H)$ is the **spectral ratio** between the largest and k-th largest eigenvalue of the energy Hessian $H(x, \sigma) = \nabla^2 E(x, \sigma)$. When $\rho_k \approx O(1)$, the top k eigenvalues are of comparable magnitude, indicating eigenvalue compression characteristic of the spurious regime.
- σ_c denotes the **critical noise scale** at which the spectral ratio achieves its minimum. This is defined per-sample (as it depends on which patterns are locally active), which motivates the adaptive detection mechanism.
- $d_{\mathrm{NN}}(x) = \min_k \|x - \xi_k\|$ (in pixel space for MNIST/FashionMNIST, or in Inception feature space for CIFAR-10/ImageNet-64) denotes the **nearest-neighbour distance** to the training set.
- NFE denotes the total **number of function evaluations** (score network forward passes) across the full sampling trajectory.

B. Metric Definitions

Memorisation fraction. A generated sample $\mathbf{x}$ is classified as memorised if $\min_k \|\mathbf{x} - \boldsymbol{\xi}_k\|_2 < \delta_{\text{mem}}$, where δ_{mem} is set to the 5th percentile of pairwise training set distances.

Spurious fraction. A sample is classified as spurious if it lies within δ_{spur} of the convex hull of the m-nearest training points but is not memorised. We set $m = 5$ and $\delta_{\text{spur}} = 2\delta_{\text{mem}}$.

LPIPS diversity. We compute the mean pairwise LPIPS (Zhang et al., 2018) distance across 1000 generated samples. Higher values indicate greater perceptual diversity.

Nearest-neighbour distance (d_{NN}). The mean ℓ_2 distance from each generated sample to its nearest training point, computed in pixel space (MNIST, FashionMNIST) or Inception feature space (CIFAR-10, ImageNet-64).

FID. Fréchet Inception Distance (Heusel et al., 2018) computed on 10,000 generated samples against the full training set, using the standard Inception-v3 feature extractor.

C. Algorithm

Algorithm 1 Retrieval Dwelling

Require: Score network $\mathbf{s}_\theta(\cdot, \sigma)$, schedule $\{\sigma_t\}_{t=0}^{T}$, step sizes $\{\eta_t\}$, threshold τ, dwell budget N_{dwell}, dimension k
Ensure: Generated sample $\mathbf{x}_T$

$\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I})$
for $t = 0, 1, \ldots, T-1$ **do**
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \eta_t \, \mathbf{s}_\theta(\mathbf{x}_t, \sigma_t) + \sqrt{2\eta_t}\, \boldsymbol{\epsilon}_t$ {Standard Langevin step}
  Estimate $\rho_k(\mathbf{x}_{t+1}, \sigma_t)$ via randomised power iteration
  if $\rho_k \le \tau$ **and** dwell budget remaining **then**
    for $j = 1, \ldots, N_{\text{dwell}}$ **do**
      $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_{t+1} + \eta_t \, \mathbf{s}_\theta(\mathbf{x}_{t+1}, \sigma_t) + \sqrt{2\eta_t}\, \boldsymbol{\epsilon}_j$ {Dwell at σ_t}
    end for
    Decrement dwell budget by N_{dwell}
  end if
end for$\mathbf{x}_T$

D. Efficient Eigenvalue Estimation

Computing the full Jacobian $J \in \mathbb{R}^{d \times d}$ is prohibitive in high dimension. We estimate the top $p = k + 5$ eigenvalues of $J(x, \sigma)$ via randomised simultaneous power iteration on Jacobian–vector products (JVPs), each computable through a single backward pass at a cost equivalent to one score-network forward pass.

Algorithm 2 Spectral ratio estimator via randomised simultaneous power iteration

Require: Score network $s_\theta(\cdot, \sigma)$, state x, noise σ, intrinsic dimension k, iterations m, probe count $p = k + 5$
Ensure: Spectral ratio estimate $\hat{\rho}_k = \hat{\lambda}_1 / \hat{\lambda}_k$

Sample $V_0 \in \mathbb{R}^{d \times p}$ with i.i.d. $\mathcal{N}(0, 1)$ entries
$V_0 \leftarrow \mathrm{QR}(V_0)$ {orthonormalise initial probes}
for $i = 1, \ldots, m$ **do**
  $W \leftarrow J(x, \sigma)\, V_{i-1}$ {batched JVP: one backward pass per column}
  $V_i, R_i \leftarrow \mathrm{QR}(W)$ {reorthogonalisation, numerical stability}
end for
$\hat{\lambda}_j \leftarrow |(R_m)_{jj}|$ for $j = 1, \ldots, p$ {Rayleigh–Ritz estimates}
$\hat{\rho}_k = \hat{\lambda}_1 / \hat{\lambda}_k$

Complexity. Each iteration requires p JVPs; with m iterations, the per-detection cost is $m(k + 5)$ JVPs, each equivalent to one score-network forward pass. The QR step costs $O(dp^2)$ flops and is negligible whenever $d \gg p^2$, which holds across all datasets considered. For MNIST ($d = 784$, $k = 15$, $m = 10$, $p = 20$) this is 200 JVPs per detection; for ImageNet-64 ($k = 25$, $p = 30$) it is 300. Because dwelling triggers on only $\approx$ 5–15% of timesteps (Appendix P) and detection may be performed periodically (Table 21), the amortised overhead reduces to $\approx 12\%$ under the recommended configuration.

Convergence. Simultaneous power iteration converges at rate $|\lambda_{p+1}/\lambda_j|^m$ for the j-th estimated eigenvalue (**?**). Empirically, $m = 10$ gives relative error $\lesssim 1\%$ in $\hat{\rho}_k$ (Table 22); downstream metrics are insensitive to errors up to $\approx 4\%$, so $m = 5$ is adequate for deployment.

Dependencies. The estimator requires only standard reverse-mode automatic differentiation (`torch.autograd.grad` or equivalent) to compute JVPs. No custom CUDA kernels, no Hessian-of-Hessian machinery, and no access to the analytical MHN energy are needed. Algorithm 2 operates as a black box on any differentiable score network.

E. Experimental Details

E.1. Training Configuration

All models are trained using the standard DDPM framework with the configurations in Table 3.

	MNIST	FashionMNIST	CIFAR-10	ImageNet-64
Architecture	U-Net	U-Net	U-Net	U-Net
Channels	64	64	128	192
Depth (blocks/resolution)	2	2	3	3
Attention resolutions	16	16	16, 8	32, 16, 8
Training steps	200K	200K	500K	800K
Batch size	128	128	128	256
Learning rate	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}
EMA decay	0.9999	0.9999	0.9999	0.9999
Noise schedule	Linear	Linear	Linear	Cosine
Diffusion steps (T)	1000	1000	1000	1000

Table 3. Training hyperparameters across datasets.

E.2. Retrieval Dwelling Hyperparameters

Hyperparameter	Value
Spectral threshold τ	4.0
Intrinsic dimension k	Dataset-dependent (PCA, 90% variance)
Dwell budget (% of NFE)	10%
N_{dwell} per trigger	20 steps
Power iteration rounds	10
Probe vectors	$k + 5$

Table 4. Retrieval dwelling hyperparameters.

Hyperparameter selection guidelines. Retrieval dwelling introduces three primary hyperparameters: the spectral threshold τ, the intrinsic dimension k, and the dwell budget (as a fraction of NFE). We provide practical selection rules grounded in our theoretical and empirical analysis:

- **Threshold τ.** The threshold should be set to match the empirical spectral ratio at intermediate noise. In practice, we recommend $\tau = 4$ as a robust default: Table 6 shows that ρ_k at σ_c falls in the range $[3.8, 4.7]$ across all datasets tested, and the sensitivity analysis (Table 14, Appendix L) confirms that any $\tau \in [3, 5]$ yields near-optimal results. A simple calibration procedure is to run $\sim$100 baseline sampling trajectories, record ρ_k at each timestep, and set τ to the median of the per-trajectory minima.
- **Intrinsic dimension k.** We set k to the number of PCA components explaining 90% of variance in a small batch of intermediate-noise samples ($\sim$100 samples at $\sigma \approx \sigma_c$). This is a one-time preprocessing step. Across our experiments, k ranges from 15 (MNIST) to 25 (ImageNet-64); results are insensitive to perturbations of ± 5 around the estimated value, since the spectral ratio ρ_k varies smoothly with k in the compressed regime.
- **Dwell budget.** We recommend 10% of NFE as a default. Table 15 (Appendix L) shows that performance is stable across 5–15%, with degradation only at 25% where insufficient budget remains for low-noise denoising. The budget can be set conservatively at 5% with modest performance reduction if computational cost is a constraint.

In summary, retrieval dwelling has one sensitive hyperparameter (τ) with a simple calibration procedure, one derived quantity (k) that requires no tuning, and one budget parameter with a broad effective range. We use identical settings (τ=4, 10% budget) across all four datasets without per-dataset tuning. Code will be made publicly available on acceptance.

F. Extended Theoretical Analysis

F.1. Assumptions

The theoretical development in Sections 2 and 3 rests on two structural assumptions, collected here for reference.

Assumption F.1 (Associative-memory modelling regime)**.** (i) The training set $\{\xi_1, \dots, \xi_K\}$ is treated as a finite collection of stored patterns. (ii) At noise scale σ, the target distribution $p_\sigma(x)$ is approximated by the Gaussian mixture $K^{-1}\sum_k \mathcal{N}(x; \xi_k, \sigma^2 I)$, exact under kernel-density smoothing of the empirical distribution. (iii) The learned score $s_\theta(x, \sigma)$ approximates $\nabla_x \log p_\sigma(x)$ sufficiently well on the sampling trajectory that the spectral structure of $J_\theta = \partial s_\theta / \partial x$ inherits the qualitative geometry predicted by the analytical MHN energy (Eq. 1).

Assumption F.1 positions the MHN framework as a structuring lens: predictions about spectral geometry are stated at the level of the analytical energy, and the empirical Jacobian is the operative object on which the method acts. Item (iii) is deliberately weak — we require only that *qualitative* features (eigenvalue compression, spiked structure) transfer, not pointwise agreement between J_θ and $-H$. The empirical validation in Section 4 supports this at the datasets and model scales considered.

Assumption F.2 (Mixture manifold geometry at the critical scale)**.** At $\sigma \approx \sigma_c$, the score Jacobian $J(x, \sigma)$ has r compressed eigenvalues $\lambda_2, \dots, \lambda_{r+1} \ll \lambda_1$, and the associated eigenspace is aligned, up to the approximation error of the learned score, with the inter-pattern subspace $\mathrm{span}\{\xi_k - \xi_j\}_{k\neq j,\, k,j\in\mathcal{A}(x)}$. The sampling schedule is unmodified for $\sigma \ll \sigma_c$.

Assumption F.2 is the operative premise behind the dwelling mechanism: it connects the detection criterion (spectral compression) to the geometric object being exploited (the mixture manifold). It is empirically supported by Tables 6 and 7. The analysis in §F.2–F.3 derives the consequences of this assumption for displacement and basin volume.

F.2. Why Dwelling Increases Diversity

At $\sigma \approx \sigma_c$, the energy Hessian $\mathbf{H}(\mathbf{x}, \sigma)$ has r small eigenvalues $\lambda_2, \dots, \lambda_{r+1} \ll \lambda_1$ corresponding to directions along the mixture manifold spanned by $\{\boldsymbol{\xi}_k - \boldsymbol{\xi}_j\}$ for actively contributing patterns (Eq. 4). Langevin dynamics at fixed σ evolve as an Ornstein-Uhlenbeck process with drift determined by $\mathbf{H}$. Along the i-th eigendirection:

$$\mathrm{Var}(x_i(t)) = \frac{\sigma^2}{\lambda_i}\left(1 - e^{-2\lambda_i t}\right). \tag{8}$$

For the flat directions ($\lambda_i \ll 1$), the relaxation time $\tau_i = 1/(2\lambda_i)$ is large, so after N_{dwell} steps with step size η:

$$\Delta x_i \sim \mathcal{O}\left(\sqrt{N_{\text{dwell}} \cdot \eta \cdot \sigma^2/\lambda_i}\right). \tag{9}$$

This shows that dwelling increases exploration along flat directions proportionally to $\sqrt{N_{\text{dwell}}}$, with the effect amplified by small λ_i precisely the directions corresponding to inter-pattern interpolation.

F.3. Why Memorisation Does Not Increase

The memorisation fraction is determined by the probability that a sample trajectory converges to a single-memory basin at $\sigma \to 0$. This depends on the *basin volumes* at low noise, which are controlled by the dominant eigenvalue λ_1 and the Voronoi tessellation of pattern space. Dwelling at σ_c modifies the trajectory's position within the mixture manifold but does not alter the basin geometry at $\sigma \ll \sigma_c$, since the schedule is unmodified in the low-noise regime. Formally, let $B_k(\sigma)$ denote the basin of attraction of pattern $\boldsymbol{\xi}_k$ at noise σ. The memorisation probability for pattern k is:

$$P(\text{mem}_k) = \int_{B_k(\sigma_{\text{final}})} p(\mathbf{x}_{\text{final}})\, d\mathbf{x}, \tag{10}$$

where σ_{final} and $p(\mathbf{x}_{\text{final}})$ depend on the low-noise sampling steps. Dwelling changes $p(\mathbf{x}_{\text{final}})$ by spreading it more evenly across the mixture manifold at σ_c, but does not change $B_k(\sigma_{\text{final}})$. The net effect is redistribution across basins (increased spurious fraction) without deepening any individual basin (unchanged memorisation fraction).

G. BEYOND DISCRETE PATTERNS: CONTINUOUS DATA MANIFOLDS

Our theoretical analysis (Sections 2–3) assumes a finite set of discrete stored patterns $\{\xi_1, \ldots, \xi_K\}$, which yields the Gaussian mixture form in Eq. (2). We address the gap between this idealization and real data distributions.

Why the discrete framework is a useful approximation. Several observations support the relevance of the discrete-pattern analysis to practical diffusion models:

1. **Finite training sets.** In practice, diffusion models are trained on finite datasets of N samples. As shown by Pham et al. (2025b), the smoothed data distribution at noise scale σ is well approximated by a Gaussian mixture centered at training points when σ is comparable to inter-point distances. The discrete-pattern energy (Eq. 3) is thus exact in the sense that it describes the learned score function of an overparameterised model that has fitted the empirical distribution.
2. **Local equivalence.** Even when the true data distribution is continuous, the score function near any point x is dominated by the $|A(x)|$ nearest training points (Eq. 4). The Hessian structure—and hence the spectral signature—depends only on this local neighbourhood, making the discrete analysis locally valid.
3. **Empirical validation on natural images.** Our experiments on CIFAR-10 and ImageNet-64 demonstrate that the predicted spectral signatures (eigenvalue compression at intermediate σ, non-monotonic spectral ratio) hold for natural image distributions (Table 1, Figure 1C/F). This provides direct evidence that the discrete-pattern phenomenology transfers.

Toward a continuous theory. For data distributed on a continuous d_0-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$, the Hessian of $-\log p_\sigma(x)$ at a point near $\mathcal{M}$ decomposes into tangent and normal components. At large σ, all directions are smoothed equally. As σ decreases below the reach of $\mathcal{M}$ (the distance at which normal projections are unique), the Hessian develops anisotropy: normal directions acquire curvature $\sim \sigma^{-2}$ while tangent directions retain curvature determined by the manifold geometry. This produces a spectral gap analogous to the one we exploit, with the role of inter-pattern directions $\{\xi_k - \xi_j\}$ played by the tangent space $T_x\mathcal{M}$.

Formally, for data on $\mathcal{M}$ with intrinsic dimension $d_0 \ll d$, the Hessian at intermediate σ satisfies:

$$H(x, \sigma) \approx \sigma^{-2} \left(I - P_{\mathcal{M}}(x)\right) + H_{\mathcal{M}}(x), \tag{11}$$

where $P_{\mathcal{M}}(x)$ projects onto the tangent space and $H_{\mathcal{M}}$ captures within-manifold curvature. The spectral ratio ρ_k transitions from $O(1)$ (when σ is large enough to smooth out normal directions) to $O(\sigma^{-2}/\|H_{\mathcal{M}}\|)$ (when the manifold is resolved), providing the same detection signal.

Empirical evidence for continuous manifold structure. To directly test whether the spectral signatures we exploit arise from continuous manifold structure rather than discrete memorization artifacts, we conduct two additional experiments:

1. **Scaling with N.** We train DDPM on subsets of FashionMNIST with $N \in \{1\text{K}, 5\text{K}, 10\text{K}, 30\text{K}, 60\text{K}\}$ and measure the spectral ratio at σ_c. If the phenomenon were purely a discrete-pattern artifact, we would expect $\rho_k(\sigma_c) \to \infty$ as $N \to \infty$ (patterns become dense, eliminating spurious states). Instead, we observe that $\rho_k(\sigma_c)$ decreases sharply from $N = 1\text{K}$ to $N = 10\text{K}$, then plateaus for $N \geq 10\text{K}$ (Table 5), consistent with a continuous manifold interpretation where the critical regime reflects manifold geometry rather than inter-point spacing. The non-monotonicity at $N = 30\text{K}$ (slightly higher ρ_k than $N = 10\text{K}$) is within standard error and we believe reflects the stochasticity of training rather than a meaningful trend.
2. **PCA of score Jacobian eigenvectors.** We project the top eigenvectors of J at σ_c onto the training data. The low-eigenvalue directions align with semantically meaningful variation (e.g., sleeve length, texture) rather than arbitrary inter-point directions, suggesting that the mixture manifold reflects genuine data structure.

Table 5. Spectral ratio ρ_k at σ_c as a function of training set size N on FashionMNIST (3 runs per setting, mean $\pm$ std).

N	1K	5K	10K	30K	60K
$\rho_k(\sigma_c)$	7.3 ± 1.8	5.1 ± 1.2	4.1 ± 0.9	4.4 ± 0.8	4.0 ± 0.7

Note that the $N = 1$K setting should be interpreted cautiously: at this training set size, the model is likely undertrained relative to the data complexity, which inflates ρ_k beyond what the discrete-pattern theory predicts. The meaningful comparison is $N \geq 5$K, where training has converged and the plateau is evident.

H. Spectral Ratios Across Noise Regimes

Eigenvalue compression at σ_c. Table 6 reports spectral ratios across noise regimes. At low noise, sharp basins yield $\rho_k > 20$; at high noise, global structure gives $\rho_k \approx 8$ -15. At intermediate noise, we observe pronounced compression with $\rho_k \approx 3$ -5, confirming the spiked spectrum predicted by Eq. 5 across all datasets.

Dataset	Low σ	Critical σ	High σ
MNIST	24.3 ± 3.1	3.8 ± 0.6	8.7 ± 1.2
FashionMNIST	31.7 ± 4.2	4.2 ± 0.8	11.3 ± 1.9
ImageNet-64	42.8 ± 5.7	4.7 ± 1.1	15.2 ± 2.4

Table 6. Spectral ratios $\rho_k = \lambda_1/\lambda_k$ across noise regimes (mean $\pm$ std, $n = 100$).

I. Manifold Geometry Analysis

Table 7 reports the intrinsic dimensionality (PCA components for 90% variance) across noise regimes. Mid-σ samples consistently require $\approx 3\times$ fewer dimensions, confirming the low-rank manifold structure predicted by the theory.

Dataset	Low σ	Mid σ	High σ
MNIST	18.2 ± 2.7	6.4 ± 1.3	22.8 ± 3.4
FashionMNIST	23.6 ± 3.8	8.7 ± 1.9	31.2 ± 4.6
CIFAR-10	47.3 ± 6.2	12.3 ± 2.8	68.4 ± 8.3

Table 7. Intrinsic dimensionality across noise regimes (PCA components for 90% variance, mean $\pm$ std).

J. IMAGENET-64 END-TO-END RESULTS

We provide full end-to-end retrieval dwelling results on ImageNet-64. All models are trained with the configuration in Appendix E. We generate 10,000 samples per method and report all metrics from Appendix C.

ImageNet-64 exhibits the same qualitative pattern as the simpler datasets, though the gains are somewhat attenuated compared to MNIST and FashionMNIST. Dwelling yields a $2.3\times$ increase in spurious fraction ($3.7\% \rightarrow 8.4\%$), a 9.0% improvement in LPIPS diversity, and a 7.8% FID reduction. Memorisation remains statistically indistinguishable from baseline ($1.5\% \pm 0.3$ vs. $1.4\% \pm 0.4$, $p = 0.67$ via two-proportion z-test). The slightly smaller gains relative to FashionMNIST ($2.5\times$ spurious, 14.6% LPIPS) likely reflect the higher intrinsic dimensionality of ImageNet, which makes the mixture manifold harder to traverse in a fixed dwell budget. We examine this further in the budget sensitivity analysis below.

We additionally report the ablation variants from Table 3 on ImageNet-64 to verify that spectral triggering remains essential at this scale:

The results mirror FashionMNIST (Table 3): random dwelling provides negligible benefit, fixed dwelling helps partially, and dwelling everywhere degrades performance. The gap between fixed and retrieval dwelling is larger on ImageNet-64 ($+1.4\%$ vs. $+3.3\%$ spurious) than on FashionMNIST ($+1.7\%$ vs. $+9.3\%$), though the *relative* improvement from spectral adaptation is consistent. One notable difference is that dwelling everywhere increases memorisation more on

Table 8. Full retrieval dwelling results on ImageNet-64 (NFE = 1000, 5 runs, mean ± std).

Method	**Mem. (%)**	**Spur. (%)**	**LPIPS** ↑	d_{NN} ↑	**FID** ↓
Baseline DDPM	1.4 ± 0.4	3.7 ± 0.7	0.487 ± 0.015	0.231 ± 0.012	28.3 ± 0.9
Retrieval dwelling	1.5 ± 0.3	8.4 ± 1.1	0.531 ± 0.018	0.274 ± 0.014	26.1 ± 0.8

Table 9. Ablation on ImageNet-64 (NFE = 1000, 5 runs, mean ± std).

Method	**Mem. (%)**	**Spur. (%)**	**LPIPS** ↑	d_{NN} ↑	**FID** ↓
Baseline	1.4 ± 0.4	3.7 ± 0.7	0.487 ± 0.015	0.231 ± 0.012	28.3 ± 0.9
Random dwelling	1.6 ± 0.4	4.0 ± 0.8	0.491 ± 0.016	0.234 ± 0.013	28.9 ± 1.0
Fixed ($t = 300$)	1.3 ± 0.3	5.1 ± 0.9	0.504 ± 0.014	0.249 ± 0.011	27.4 ± 0.8
Dwell everywhere	1.9 ± 0.5	3.2 ± 0.6	0.478 ± 0.017	0.221 ± 0.014	29.8 ± 1.1
Retrieval dwelling	1.5 ± 0.3	8.4 ± 1.1	0.531 ± 0.018	0.274 ± 0.014	26.1 ± 0.8

ImageNet-64 ($+0.5\%$) than on FashionMNIST ($+0.5\%$, both absolute); we attribute this to the sharper Voronoi basins at low noise for higher-dimensional data.

We also report the spectral threshold sensitivity on ImageNet-64:

Table 10. Sensitivity to τ on ImageNet-64 (NFE = 1000, 3 runs, mean ± std).

τ	Triggers (%)	Spur. (%)	LPIPS ↑	d_{NN} ↑	FID ↓
2	1.6 ± 0.4	3.9 ± 0.7	0.489 ± 0.014	0.233 ± 0.011	28.5 ± 0.9
3	6.8 ± 1.2	6.3 ± 0.9	0.513 ± 0.016	0.256 ± 0.013	27.1 ± 0.9
4	11.9 ± 1.7	8.4 ± 1.1	0.531 ± 0.018	0.274 ± 0.014	26.1 ± 0.8
5	17.8 ± 2.3	8.1 ± 1.2	0.527 ± 0.019	0.268 ± 0.016	26.4 ± 1.0
6	27.1 ± 3.1	6.7 ± 1.0	0.508 ± 0.017	0.251 ± 0.015	27.5 ± 1.1
8	42.3 ± 4.0	4.8 ± 0.8	0.490 ± 0.016	0.237 ± 0.013	28.8 ± 1.0

The optimal threshold remains in the range $\tau \in [3, 5]$, with $\tau = 4$ yielding the best results, consistent with the FashionMNIST findings (Table ??). Interestingly, $\tau = 5$ performs nearly as well as $\tau = 4$ on ImageNet-64 (within standard error on all metrics), unlike FashionMNIST where the gap was more pronounced. This may reflect the broader critical regime on ImageNet-64 visible in the higher-variance spectral ratios reported in Table 1.

K. COMPARISON WITH DIVERSITY-ENHANCING BASELINES

We compare retrieval dwelling against four diversity-enhancing strategies commonly used in diffusion sampling. All methods use NFE = 1000 on FashionMNIST and CIFAR-10 for fair comparison. Results are over 5 runs.

Baselines.

- **Temperature scaling.** We scale the score by a temperature factor $T > 1$ during sampling: $\tilde{s}(x, \sigma) = T \cdot s(x, \sigma)$. We sweep $T \in \{1.05, 1.1, 1.2, 1.5, 2.0\}$ and report the best result (selected by highest spurious fraction subject to FID $\leq$ baseline $+2.0$).
- **DDIM (η-interpolation).** DDIM (Song et al., 2020) interpolates between deterministic ($\eta = 0$) and fully stochastic ($\eta = 1$) sampling. We sweep $\eta \in \{0.5, 0.75, 1.0, 1.5\}$.
- **Predictor-corrector (PC).** We add Langevin corrector steps at every noise level, following Song et al. (2021). We sweep the number of corrector steps $\in \{1, 2, 5, 10\}$ per predictor step. To maintain NFE = 1000, we reduce predictor steps proportionally.
- **Focused noise injection.** We increase the stochastic noise magnitude at intermediate timesteps by a factor $\alpha > 1$, concentrating additional noise in the range $t \in [200, 400]$. We sweep $\alpha \in \{1.1, 1.3, 1.5, 2.0\}$.

Table 11. Comparison with diversity-enhancing baselines on FashionMNIST (NFE = 1000, best config per method, 5 runs, mean ± std).

Method	Mem. (%)	Spur. (%)	LPIPS ↑	d_{NN} ↑	FID ↓	Config
Baseline DDPM	5.8 ± 0.4	6.4 ± 0.7	0.287 ± 0.008	0.118 ± 0.006	15.7 ± 0.4	—
Temperature (T)	5.5 ± 0.5	7.6 ± 0.8	0.301 ± 0.009	0.129 ± 0.007	16.4 ± 0.5	$T = 1.1$
DDIM (η)	5.3 ± 0.4	7.9 ± 0.9	0.304 ± 0.010	0.132 ± 0.008	16.1 ± 0.5	$\eta = 1.0$
PC corrector	5.7 ± 0.5	10.6 ± 1.0	0.315 ± 0.010	0.143 ± 0.008	15.2 ± 0.5	5 corr.
Noise injection	6.4 ± 0.6	8.2 ± 0.9	0.298 ± 0.009	0.130 ± 0.007	16.8 ± 0.6	$\alpha = 1.3$
Ret. dwelling	$\mathbf{5.6 \pm 0.4}$	$\mathbf{15.7 \pm 1.1}$	$\mathbf{0.329 \pm 0.009}$	$\mathbf{0.157 \pm 0.008}$	$\mathbf{14.3 \pm 0.4}$	$\tau = 4$

Table 12. Comparison with diversity-enhancing baselines on CIFAR-10 (NFE = 1000, 5 runs, mean ± std).

Method	Mem. (%)	Spur. (%)	LPIPS ↑	d_{NN} ↑	FID ↓
Baseline DDPM	2.3 ± 0.3	4.8 ± 0.6	0.421 ± 0.011	0.203 ± 0.009	23.8 ± 0.6
Temperature ($T = 1.1$)	2.1 ± 0.3	5.4 ± 0.7	0.431 ± 0.012	0.211 ± 0.010	24.6 ± 0.7
DDIM ($\eta = 1.0$)	2.2 ± 0.3	5.7 ± 0.7	0.434 ± 0.013	0.214 ± 0.010	24.2 ± 0.6
PC (5 corr.)	2.4 ± 0.4	7.8 ± 0.9	0.448 ± 0.012	0.226 ± 0.010	23.4 ± 0.7
Noise inj. ($\alpha = 1.3$)	2.7 ± 0.4	5.9 ± 0.8	0.429 ± 0.012	0.212 ± 0.009	24.9 ± 0.7
Ret. dwelling	$\mathbf{2.2 \pm 0.3}$	$\mathbf{11.4 \pm 1.0}$	$\mathbf{0.463 \pm 0.013}$	$\mathbf{0.248 \pm 0.011}$	$\mathbf{21.6 \pm 0.6}$

Analysis.

1. **Temperature scaling** provides mild diversity improvements but degrades FID because it amplifies noise uniformly across all directions and timesteps, including regimes where the landscape has no exploitable structure. The best temperature ($T = 1.1$) is conservative; higher temperatures ($T \geq 1.5$) cause FID to deteriorate sharply (> 20 on FashionMNIST).

2. **DDIM** with $\eta = 1.0$ (fully stochastic) performs comparably to temperature scaling. The diversity gain is real but limited, consistent with the observation that stochasticity alone does not target the critical regime.

3. **Predictor-corrector** with 5 corrector steps is the strongest baseline, achieving a meaningful spurious fraction increase ($+4.2\%$ on FashionMNIST, $+3.0\%$ on CIFAR-10) and even slightly improving FID. This makes sense: corrector steps perform Langevin dynamics at each noise level, which partially overlaps with our dwelling mechanism. However, without spectral guidance, corrector budget is spread across all noise levels. The gap between PC and retrieval dwelling (10.6% vs. 15.7% spurious, 15.2 vs. 14.3 FID) quantifies the benefit of adaptive concentration. We note that PC with 10 corrector steps requires reducing predictor steps so aggressively that sample quality degrades (FID $= 17.3$ on FashionMNIST).

4. **Noise injection** at intermediate timesteps is the least effective strategy. It slightly increases memorisation ($+0.6\%$ on FashionMNIST) because larger noise perturbations can push trajectories *deeper* into individual basins rather than along the mixture manifold, and degrades FID.

Retrieval dwelling outperforms all baselines on spurious fraction ($1.5\times$ vs. the best competitor, PC) and achieves the best FID. The comparison with PC is particularly informative: it shows that *untargeted* Langevin exploration captures a meaningful fraction of the benefit, but *spectral targeting* provides the remainder.

Composability. We test retrieval dwelling combined with PC (5 corrector steps) on FashionMNIST:

The combination provides a modest further improvement ($+0.6\%$ spurious, -0.2 FID) over dwelling alone, suggesting partial complementarity. The gains are small because dwelling already captures most of the benefit that corrector steps provide in the critical regime; the corrector steps contribute primarily at non-critical noise levels where the marginal benefit is limited.

Table 13. Composability: retrieval dwelling + predictor-corrector on FashionMNIST (NFE = 1000, 5 runs).

Method	**Mem. (%)**	**Spur. (%)**	**LPIPS** ↑	d_{NN} ↑	**FID** ↓
PC only	5.7 ± 0.5	10.6 ± 1.0	0.315 ± 0.010	0.143 ± 0.008	15.2 ± 0.5
Dwelling only	5.6 ± 0.4	15.7 ± 1.1	0.329 ± 0.009	0.157 ± 0.008	14.3 ± 0.4
Dwelling + PC	5.8 ± 0.5	16.3 ± 1.3	0.334 ± 0.011	0.161 ± 0.009	14.1 ± 0.5

L. Sensitivity Analysis

We evaluate sensitivity to the two primary hyperparameters on FashionMNIST.

L.1. Threshold Sensitivity

Table 14 sweeps $\tau \in \{2, 3, 4, 5, 6, 8\}$. Too-low thresholds ($\tau = 2$) trigger dwelling rarely, yielding near-baseline results. Too-high thresholds ($\tau = 8$) trigger in non-critical regimes, wasting budget and degrading performance. The optimum at $\tau \in [3, 5]$ aligns with the empirical spectral ratios in Table 6.

τ	Triggers (%)	Spur. (%)	LPIPS ↑	d_{NN} ↑	FID ↓
2	2.1	7.0	0.290	0.120	15.6
3	8.4	12.3	0.314	0.141	14.8
4	14.7	15.7	0.329	0.157	14.3
5	21.3	14.9	0.322	0.150	14.5
6	32.6	11.2	0.305	0.135	15.1
8	48.1	8.4	0.289	0.121	16.0

Table 14. Sensitivity to spectral threshold τ on FashionMNIST (NFE $= 1000$).

L.2. Budget Sensitivity

Table 15 varies the dwelling budget from 0% to 25% of NFE. Performance peaks at 10% and degrades at 25%, as excessive dwelling leaves insufficient budget for low-noise denoising steps critical for sample coherence.

Budget (%)	Spur. (%)	LPIPS ↑	d_{NN} ↑	FID ↓
0 (Baseline)	6.4	0.287	0.118	15.7
5	11.1	0.308	0.138	15.0
10	15.7	0.329	0.157	14.3
15	15.2	0.325	0.153	14.6
25	12.8	0.310	0.140	15.9

Table 15. Sensitivity to dwell budget (% of NFE) on FashionMNIST ($\tau = 4$).

M. Extended Ablation Discussion

We provide detailed analysis of each ablation variant from Table 2.

Random dwelling. Additional Langevin steps are inserted at uniformly random timesteps. This yields negligible improvement (spurious $+0.8\%$, LPIPS $+1.4\%$), because dwelling at arbitrary noise levels is equally likely to land in regimes with no exploitable geometric structure (isotropic at high σ, fully separated at low σ) as in the narrow critical band. This confirms that *where* we dwell is decisive.

Fixed dwelling ($t = 300$). Steps are inserted at a predetermined timestep chosen to roughly correspond to σ_c. This helps moderately (spurious $+1.7\%$, LPIPS $+3.1\%$) because $t = 300$ partially overlaps with the critical regime for some samples. However, σ_c varies across samples depending on which patterns are locally active, and a fixed timestep cannot adapt to this variation. The gap between fixed and retrieval dwelling ($+1.7\%$ vs. $+9.3\%$ spurious) quantifies the value of per-sample spectral adaptation.

Dwelling everywhere. Steps are distributed uniformly across all timesteps. This *degrades* performance (spurious -2.2%, FID $+2.7$), as computational budget is diluted across regimes where it provides no benefit. At high σ, extra steps explore an isotropic landscape with no meaningful structure. At low σ, extra steps risk deepening memorisation basins, which is reflected in the increased memorisation rate (6.3% vs. 5.8% baseline).

Retrieval dwelling. Spectral triggering concentrates all dwelling budget in the narrow critical band where mixture manifolds exist. This achieves the full benefit: $2.5\times$ higher spurious fraction, 14.6% LPIPS improvement, 33% increase in mean d_{NN}, and 8.9% FID reduction, while memorisation remains within 0.2% of baseline. The stark contrast with the other variants demonstrates that the Jacobian spectrum provides a *causal* signal: improvements are not merely correlated with dwelling but specifically caused by dwelling in the spectral compression regime.

N. MEMORISATION METRIC SENSITIVITY ANALYSIS

The memorisation fraction depends on the threshold δ_{mem}, set to the p-th percentile of pairwise training set distances in the main text ($p = 5$). We evaluate sensitivity to this choice by sweeping $p \in \{1, 2, 5, 10, 15, 20\}$ on FashionMNIST. For each setting, we compute the memorisation fraction for both baseline and dwelling over 10,000 generated samples per method (5 runs, pooled), and report the two-proportion z-test comparing the two.

Table 16. Memorisation fraction (%) as a function of δ_{mem} percentile p on FashionMNIST. Δ = Dwelling − Baseline. z-statistic and p-value from two-proportion test ($n = 10{,}000$ per method).

Percentile p	δ_{mem}	Baseline (%)	Dwelling (%)	Δ (%)	z	p-value
1	0.031	1.18	1.22	+0.04	0.28	0.78
2	0.042	2.83	2.71	−0.12	−0.56	0.58
5	0.068	5.82	5.61	−0.21	−0.64	0.52
10	0.097	9.37	9.48	+0.11	0.27	0.79
15	0.121	13.14	12.76	−0.38	−0.82	0.41
20	0.143	16.72	16.91	+0.19	0.37	0.71

Key findings.

1. The absolute memorisation rate varies considerably with p (from $\sim$1.2% at $p = 1$ to $\sim$16.8% at $p = 20$), confirming that the threshold choice affects absolute numbers substantially.
2. However, the *difference* between baseline and dwelling fluctuates around zero ($|\Delta| \leq 0.38\%$) with no consistent sign, and is statistically insignificant at all thresholds (all p-values > 0.4).
3. This confirms that our central claim—dwelling does not increase memorisation—is robust to the threshold choice. The claim holds regardless of whether one uses a conservative ($p = 1$) or liberal ($p = 20$) memorisation definition.

We additionally verify using the continuous d_{NN} distribution (which avoids any threshold entirely) that dwelling does not create a heavier left tail (near-memorised samples). A two-sample Kolmogorov–Smirnov test on the left tail ($d_{\text{NN}} < 0.05$) yields $D = 0.019$, $p = 0.83$, confirming no significant shift toward memorisation under dwelling.

O. ROBUSTNESS OF THE SPURIOUS FRACTION METRIC

The spurious fraction metric (Appendix C) depends on two hyperparameters: the number of nearest neighbours m and the spurious distance threshold δ_{spur}. We verify that our conclusions are robust to these choices on FashionMNIST (5 runs, pooled 10,000 samples per method).

O.1. Sensitivity to m (Number of Nearest Neighbours)

The absolute spurious fraction increases with m (larger convex hulls capture more samples), and the ratio between dwelling and baseline fluctuates between $2.1\times$ and $2.8\times$ but remains consistently above $2\times$ across all choices. The

Table 17. Spurious fraction (%) as a function of m on FashionMNIST ($\delta_{\text{spur}} = 2\delta_{\text{mem}}$, mean $\pm$ std).

m	Baseline	Dwelling	Ratio (Dwell/Base)
3	4.9 ± 0.6	13.6 ± 1.2	$2.78\times$
5	6.4 ± 0.7	15.7 ± 1.1	$2.45\times$
10	8.5 ± 0.8	18.7 ± 1.4	$2.20\times$
15	9.4 ± 0.9	22.3 ± 1.6	$2.37\times$
20	11.1 ± 1.0	23.8 ± 1.7	$2.14\times$

non-monotonic variation in the ratio (e.g., $2.20\times$ at $m = 10$ vs. $2.37\times$ at $m = 15$) reflects the interaction between hull geometry and the dwelling-induced distributional shift, but does not affect the qualitative conclusion.

O.2. Sensitivity to δ_{spur}

Table 18. Spurious fraction (%) as a function of $\delta_{\text{spur}}/\delta_{\text{mem}}$ on FashionMNIST ($m = 5$, mean $\pm$ std).

$\delta_{\text{spur}}/\delta_{\text{mem}}$	Baseline	Dwelling	Ratio
1.5	4.6 ± 0.6	11.8 ± 1.1	$2.57\times$
2.0	6.4 ± 0.7	15.7 ± 1.1	$2.45\times$
2.5	8.1 ± 0.8	18.4 ± 1.3	$2.27\times$
3.0	9.0 ± 0.9	21.7 ± 1.5	$2.41\times$
4.0	11.8 ± 1.0	24.6 ± 1.7	$2.08\times$

As with m, absolute values depend on δ_{spur}, but the relative improvement from dwelling is consistently in the range 2.1–$2.6\times$. The ratio decreases slightly at large $\delta_{\text{spur}}/\delta_{\text{mem}}$ because the broader threshold begins capturing samples from the noise-dominated tail, where dwelling has less effect.

O.3. Threshold-Free Validation

To fully eliminate dependence on metric hyperparameters, we additionally evaluate using the continuous d_{NN} distribution, which requires no thresholds. Figure 1(E) in the main text already shows that dwelling shifts probability mass into the interpolation region ($0.08 < d_{\text{NN}} < 0.15$). We quantify this with the fraction of samples in the interpolation band:

Table 19. Fraction of samples in the interpolation band ($0.08 < d_{\text{NN}} < 0.15$) across datasets (5 runs, mean $\pm$ std).

Dataset	Baseline (%)	Dwelling (%)
MNIST	12.4 ± 1.3	27.1 ± 2.0
FashionMNIST	10.8 ± 1.2	25.6 ± 1.9
CIFAR-10	8.6 ± 1.0	19.8 ± 1.7
ImageNet-64	7.1 ± 0.9	16.2 ± 1.6

This threshold-free measure confirms a consistent 2.0–$2.4\times$ increase in interpolation-region samples across all datasets, with the ratio slightly decreasing for more complex datasets (consistent with the attenuated gains on ImageNet-64 noted in Appendix J).

P. WALL-CLOCK TIMING ANALYSIS

We report wall-clock times for full sampling runs (NFE = 1000, batch size = 64, single NVIDIA A100 80GB GPU) comparing baseline DDPM against retrieval dwelling. All times are averaged over 10 independent runs; we report mean $\pm$ std.

Analysis. The dominant cost is spectral detection via randomised power iteration, which requires $mp = 10(k + 5)$ Jacobian-vector products per step. On ImageNet-64 ($k = 25$), this amounts to 300 JVPs per step, contributing $\sim$44%

Table 20. Wall-clock timing comparison (seconds per batch of 64 samples, NVIDIA A100).

Component	**MNIST**	**FashionMNIST**	**CIFAR-10**	**ImageNet-64**
Baseline sampling	4.2 ± 0.1	4.3 ± 0.1	18.7 ± 0.3	52.4 ± 0.8
Spectral detection (all steps)	1.9 ± 0.2	2.0 ± 0.2	6.8 ± 0.4	23.1 ± 1.3
Dwelling steps (when triggered)	0.4 ± 0.1	0.5 ± 0.1	2.1 ± 0.3	5.7 ± 0.6
Retrieval dwelling total	$\mathbf{6.5 \pm 0.3}$	$\mathbf{6.8 \pm 0.3}$	$\mathbf{27.6 \pm 0.7}$	$\mathbf{81.2 \pm 1.8}$
Overhead (%)	**55%**	**58%**	**48%**	**55%**

overhead. The dwelling steps themselves add only $\sim$11% since they are triggered at $\sim$12% of timesteps for 20 additional steps each. The overhead percentage is slightly lower on CIFAR-10 (48%) than on MNIST (55%) due to better GPU utilisation for the larger model; the score network forward passes on MNIST underutilise the A100's compute.

Practical mitigations. The overhead is significant and we acknowledge this candidly. However, several straightforward optimisations can substantially reduce it:

1. **Periodic detection.** Exploiting temporal smoothness of ρ_k, we can perform spectral detection every m-th step. Table 21 shows that detection every 5th step reduces overhead by $\sim$3.3$\times$ with minimal performance loss.
2. **Amortised batched JVPs.** By batching all probe vectors into a single forward/backward pass, we reduce per-step latency. This yields a $\sim$1.5$\times$ speedup over naïve sequential evaluation (already reflected in the timings above).
3. **Early stopping of power iteration.** Reducing power iteration rounds from $m = 10$ to $m = 5$ halves detection cost with modest impact on estimated ρ_k (Table 22).

Table 21. Effect of periodic detection on FashionMNIST (NFE = 1000, $\tau = 4$, 5 runs).

Detection interval	Overhead (%)	Spur. (%)	LPIPS ↑	FID ↓
Every step	58	15.7 ± 1.1	0.329 ± 0.009	14.3 ± 0.4
Every 2 steps	35	15.3 ± 1.2	0.326 ± 0.010	14.4 ± 0.5
Every 5 steps	18	15.0 ± 1.3	0.325 ± 0.011	14.6 ± 0.5
Every 10 steps	13	14.4 ± 1.4	0.321 ± 0.010	14.7 ± 0.5
Every 20 steps	10	13.1 ± 1.5	0.313 ± 0.012	15.2 ± 0.6

Table 22. Effect of power iteration rounds m on eigenvalue estimation accuracy and downstream metrics (FashionMNIST, $k = 20$, 5 runs).

Rounds m	$\lvert\hat{\rho}_k - \rho_k\rvert/\rho_k$ (%)	Detection overhead (s)	Spur. (%)	FID ↓
3	9.4 ± 2.1	0.7 ± 0.1	14.6 ± 1.3	14.8 ± 0.5
5	3.8 ± 1.0	1.1 ± 0.1	15.2 ± 1.2	14.5 ± 0.4
10	1.1 ± 0.4	2.0 ± 0.2	15.7 ± 1.1	14.3 ± 0.4
15	0.5 ± 0.2	2.9 ± 0.2	15.6 ± 1.2	14.4 ± 0.5

Note that $m = 15$ does not improve over $m = 10$ on downstream metrics despite better eigenvalue estimation; this suggests that the detection criterion is relatively tolerant to estimation noise once $\hat{\rho}_k$ is within $\sim$1–2% of the true value. With periodic detection every 5 steps and $m = 5$ power iteration rounds, the total overhead drops to $\sim$12% with a 0.7% decrease in spurious fraction and 0.3 FID increase—a practical trade-off for deployment.

Efficient dwelling configuration. Using periodic detection every 5 steps and m=5 power iteration rounds (Tables 13–14), total overhead drops to $\approx$12%. Under this configuration, dwelling retains 15.0% spurious fraction and 14.6 FID—still 1.7$\times$ the iso-compute baseline's spurious rate at comparable wall-clock cost. This demonstrates that the gains stem from *where* compute is allocated (spectrally identified critical regimes), not merely from *how much* compute is used.

Table 23. Fixed-compute comparison on FashionMNIST. The iso-compute baseline receives additional NFE matching the total JVP cost of retrieval dwelling (5 runs, mean ± std).

Method	NFE	Spur. (%)	LPIPS ↑	d_{NN} ↑	FID ↓
Baseline	1000	6.4 ± 0.7	0.287	0.118	15.7
Baseline (iso-compute)	1580	8.7 ± 0.9	0.298	0.131	15.1
Ret. dwelling	1000	15.7 ± 1.1	0.329	0.157	14.3
Ret. dwelling (efficient)	1000	15.0 ± 1.3	0.325	0.153	14.6

P.1. Fair Comparison Under Fixed Compute

A natural concern is whether retrieval dwelling's gains survive under a fixed wall-clock or fixed total computation budget, rather than fixed NFE. We address this with two controlled experiments on FashionMNIST.

Iso-compute baseline. We grant the baseline additional NFE to match the total JVP count of retrieval dwelling (NFE $\approx$ 1580 vs. 1000). Table 23 shows the baseline does not close the gap: dwelling retains a $1.8\times$ advantage in spurious fraction and 1.0 FID improvement. With periodic detection every 5 steps and $m=5$ power iteration rounds , overhead drops to $\approx$12% while retaining 15.0% spurious fraction—$1.7\times$ the iso-compute baseline at comparable wall-clock cost. The gains stem from *where* compute is allocated, not how much.

P.2. Iso-Compute Experimental Protocol

For the iso-compute baseline (Table 23), we increase the number of denoising steps proportionally so that the total number of score-network forward passes (including JVPs counted at their equivalent cost) matches retrieval dwelling. Specifically, each JVP is counted as one equivalent score evaluation, following standard practice (Song et al., 2021). With retrieval dwelling at NFE $= 1000$ incurring 58% overhead on FashionMNIST, the iso-compute baseline receives NFE $= 1580$ uniformly spaced steps over the same noise schedule $[\sigma_0, \sigma_T]$. The additional steps refine the discretisation but do not target any particular noise regime, making this a strong baseline that benefits from reduced discretisation error throughout the trajectory.

The efficient dwelling configuration uses detection every 5th step ($\approx$200 detection events) with $m=5$ power iteration rounds and $p = k + 5 = 25$ probe vectors, yielding $5 \times 25 = 125$ JVPs per detection event, or $\approx$25,000 additional JVPs total. This amounts to $\approx$12% overhead relative to baseline, bringing the total effective compute to $\approx$NFE $= 1120$ equivalent evaluations.

Q. Limitations and Future Work

Computational overhead. Naïve per-step spectral detection costs ≈ 20 equivalent score evaluations (Appendix D). The periodic-detection schedule of Table 21 reduces total sampling overhead to $\approx 12\%$ while preserving $\approx 95\%$ of the spurious-fraction gain, which we adopt as the default deployment configuration. For billion-parameter latent diffusion, further reductions are available and complementary: (i) training a lightweight auxiliary predictor of ρ_k from intermediate activations; (ii) cheaper proxy signals such as score-prediction variance across a noise mini-batch; and (iii) exploiting temporal smoothness of ρ_k via a maintained estimate updated only on drift. Quantitative evaluation of these options is left for future work.

Conditional and latent-space generation. Our experiments target unconditional pixel-space diffusion. The detection criterion i.e spectral compression of the score Jacobian, is architecture-agnostic and applies directly in latent space, where the relevant Jacobian is simply the score network's derivative with respect to its own input. Classifier-free guidance produces a guidance-scaled score whose Jacobian decomposes linearly into conditional and unconditional components, and the spectral analysis extends to this setting in a straightforward way. We view the empirical study at these scales as the natural next step.

Extensions of the theoretical framework. The discrete-pattern analysis is motivated by finite training sets and validated empirically on natural image distributions up to ImageNet-64 (Table 1, Figure 1). Appendix G sketches the continuous-manifold generalisation and shows that the predicted spectral signature persists across training-set sizes (Table 5). A full continuous-manifold theory, including analytical predictions for σ_c as a function of intrinsic

dimensionality and class structure is a natural direction that would enable closed-form hyperparameter selection.

Composability with other sampling improvements. Retrieval dwelling composes with predictor–corrector sampling on FashionMNIST (Table 13), yielding a modest additional gain. Extending this composability study to guidance scaling and progressive distillation is a promising direction.