How Excess Latent Dimensionality Delays Memorization in Diffusion Models

Trevor Chen TBCHEN@CALTECH.EDU
California Institute of Technology

Ryan Shahbaba SHAHBABA@CALTECH.EDU
California Institute of Technology

Avni Garg AVNI@CALTECH.EDU
California Institute of Technology

Kevin Peng KPENG2@CALTECH.EDU
California Institute of Technology

Abstract

We study how the gap between latent and intrinsic dimensionality shapes memorization in diffusion models. In practice, the latent dimensionality d_{latent} of diffusion models does not match the dataset's true intrinsic dimensionality $d_{\text{intrinsic}}$. Extending recent spectral analysis to the regime where $d_{\text{latent}} > d_{\text{intrinsic}}$, we find that the excess null dimensions populate a new "noise-dimension" bulk in the score network's feature-correlation spectrum, sitting between population-driven and sample-specific eigenmodes. Because the network learns modes in order of decreasing eigenvalue, this bulk acts as a buffer: each excess latent dimension adds a mode that must be absorbed before memorization can begin, increasing the gap between generalization and memorization. We further find that at very large d_{latent}, trainable score networks partially recover from the increase in score error that the frozen-feature theory predicts, consistent with the first layer learning sparse, signal-aligned features. These findings are supported by experiments on real and synthetic datasets and a random-feature analysis.

1. Introduction

Diffusion models are the dominant generative model class for high-dimensional continuous data such as images, video, audio, and molecules [8, 12, 13, 18, 20]. Sampling corresponds to numerically integrating the time-reversed stochastic differential equation associated with a forward noising process; the only learnable component is a score network $s_\theta(x_t, t) \approx \nabla \log p_t(x_t)$ trained to match the noisy marginals of the data distribution. Sampling can equivalently be expressed as a deterministic ordinary differential equation with the score as a velocity field, an approach taken by flow-matching methods [14].

Latent diffusion models [5, 16, 18] run forward and backward diffusion not in pixel space but in a learned latent space of much smaller dimensionality d_{latent}, typically obtained from a VAE encoder. By placing diffusion in a representation closer to the data's true intrinsic dimensionality $d_{\text{intrinsic}}$ [3, 10, 17, 21], though still well above it (with $d_{\text{latent}} \gg d_{\text{intrinsic}}$ in practice), latent diffusion achieves comparable or better sample quality at a fraction of the compute of pixel-space diffusion.

A model achieving zero training loss without regularization learns the empirical score and reproduces training samples at the end of the backward process [7]; the regime kicks in when the

© T. Chen, R. Shahbaba, A. Garg & K. Peng.

training set is small relative to model capacity and disappears past a dataset-size threshold [9]. What is missing is a sharp account of how this transition depends on the latent-space dimensionality, the design knob that distinguishes latent-diffusion from pixel-space diffusion. We ask: *how does the gap between d_{latent} and $d_{\text{intrinsic}}$ shape the transition from generalization to memorization, and is there a mechanism a practitioner can tune by adjusting the latent width?*

Existing theory provides a starting point. Bonnaire et al. [2] characterize diffusion training as a competition between a generalization time τ_{gen}, at which the score network learns the population distribution, and a memorization time τ_{mem}, at which it begins fitting individual training points. Their analysis assumes isotropic data covariance, which collapses the distinction between the $d_{\text{intrinsic}}$-dimensional signal subspace and the $d_{\text{latent}} - d_{\text{intrinsic}}$ null directions of the latent space. We extend their spectral analysis to the latent-diffusion regime $d_{\text{latent}} > d_{\text{intrinsic}}$, identify a new noise-dimension bulk that opens between the population and sample bulks, and show that it acts as a buffer that delays memorization.

Contributions.

- We extend Bonnaire's two-bulk theorem to anisotropic data and identify a four-bulk eigenvalue structure for the score network's feature correlation matrix, with bulk-count boundaries that land exactly at indices $d_{\text{intrinsic}}$ and d_{latent}.
- We identify the noise-dimension bulk as a memorization buffer: increasing d_{latent} adds $d_{\text{latent}} - d_{\text{intrinsic}}$ extra modes that the readout must absorb before reaching the sample bulk, delaying τ_{mem} while leaving τ_{gen} approximately fixed; we confirm the same mechanism in a trainable MLP on real and synthetic data.

2. Design

Synthetic data. We generate Gaussian-mixture data whose signal lies in a $d_{\text{intrinsic}}$-dimensional subspace embedded in a d_{latent}-dimensional ambient space ($d_{\text{latent}} \geq d_{\text{intrinsic}}$), with k cluster centers, intra-cluster signal variance σ^2_{sig} on the signal subspace, and isotropic null-direction variance $\sigma^2_{\perp}$. Gaussian mixtures give a closed-form population score (no noise floor), an exact intrinsic/ambient split that lets us vary d_{latent} and $d_{\text{intrinsic}}$ independently, and the block-diagonal $\mathbf{\Sigma}_{\text{data}}$ that the random-matrix-theory analysis of Section 4 requires. Specific hyperparameters and data-generation details are in Appendix B.

Real data. We additionally validate the picture empirically on MNIST and CelebA.

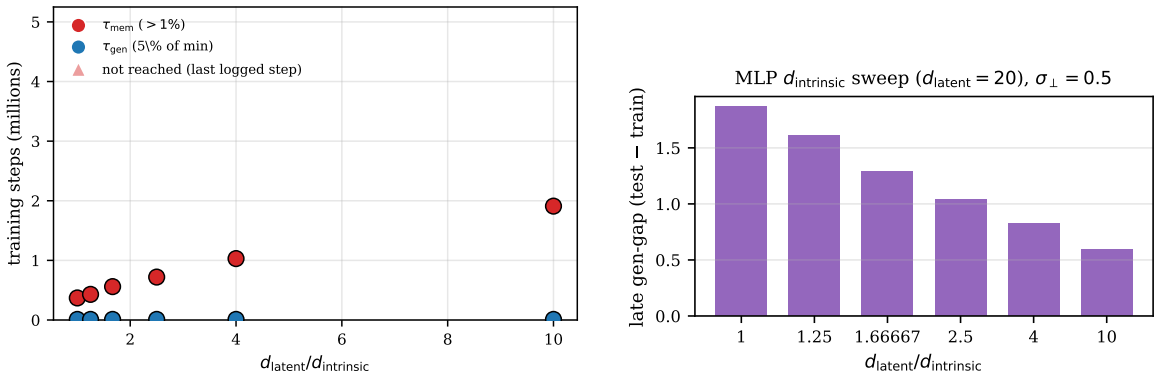
MLP score network. A trainable 3-layer MLP with sinusoidal time embedding maps $(x_t, t) \mapsto s_\theta(x_t, t) \in \mathbb{R}^{d_{\text{latent}}}$, trained by score-matching MSE against the analytic Gaussian-mixture score. The hidden width is held fixed across the d_{latent} sweep so parameter count and per-step FLOPs stay roughly constant, isolating the effect of d_{latent} from changes in model capacity.

RFNN score network. We also study the Random Feature Neural Network of Bonnaire et al. [2],

$$s_A(x_t) \;=\; \tfrac{1}{\sqrt{p}}\, A\, \tanh\!\big(\tfrac{1}{\sqrt{d_{\text{latent}}}}\, W x_t\big), \tag{1}$$

with $W \in \mathbb{R}^{p \times d_{\text{latent}}}$ frozen i.i.d. Gaussian and $A \in \mathbb{R}^{d_{\text{latent}} \times p}$ trained at fixed diffusion time $t = 0.01$. Because the first-layer features are frozen, the score-matching loss is quadratic in A and admits the closed-form spectral analysis of Section 4.

Timescales. τ_{gen} is the first step at which test loss reaches within 5% of its per-run minimum. τ_{mem} is the first step at which the Somepalli et al. [19] memorization fraction crosses 1%, with runs



(a) Timescales. (b) Late generalization gap.

Figure 1: **MLP $d_{\text{intrinsic}}$ sweep at fixed $d_{\text{latent}} = 20$, $\sigma_\perp = 0.5$.** Left: τ_{mem} grows sharply with $d_{\text{latent}}/d_{\text{intrinsic}}$ while τ_{gen} stays flat. Right: the late train–test gap shrinks as the buffer widens. The d_{latent}-sweep and low-$\sigma_\perp$ counterparts are in Appendices C, D, and F.

censored at their training budget if the threshold is not reached. For the RFNN at fixed t generation is ill-defined; we use train-test divergence (gen-gap > 0.02) as the τ_{mem} proxy. Full definitions and alternatives are in Appendix B.

Two complementary sweeps. We probe the noise-dimension buffer along two axes, each isolating one term of the buffer width $d_{\text{latent}} - d_{\text{intrinsic}}$: vary d_{latent} at fixed $d_{\text{intrinsic}} = 5$, and vary $d_{\text{intrinsic}}$ at fixed $d_{\text{latent}} = 20$. Exact sweep ranges and per-experiment configurations are in Appendix B.5.

3. The role of latent dimensionality in MLP score-network training

3.1. Synthetic data

The four-bulk predictions transfer to the trainable MLP on the same Gaussian-mixture data; full data, metric, and hyperparameter details are in Appendix A. We sweep $d_{\text{intrinsic}}$ at fixed $d_{\text{latent}} = 20$ in the body so $d_{\text{latent}}/d_{\text{intrinsic}}$ ranges over a regime in which memorization actually reaches the 1% threshold within budget for every configuration; the dual sweep over d_{latent} at fixed $d_{\text{intrinsic}} = 5$, where many high-ratio runs fail to memorize within the budget, is in Appendix C.

Delay in memorization. τ_{mem} grows sharply with $d_{\text{latent}}/d_{\text{intrinsic}}$ while τ_{gen} stays flat (Fig. 1).

Generalization gap. The late train–test loss gap is largest at $d_{\text{latent}} \approx d_{\text{intrinsic}}$ (no buffer) and shrinks monotonically as $d_{\text{latent}}/d_{\text{intrinsic}}$ grows (Fig. 1): the wider buffer keeps more of the training budget spent on real signal directions and less on sample-specific fitting.

3.2. Real data

The same buffer mechanism shows up on real-data latent diffusion over MNIST and CelebA, with the latent space learned by a VAE rather than imposed analytically. Estimated intrinsic dimensions are $d_{\text{intrinsic}} \approx 13$ for MNIST and $d_{\text{intrinsic}} \approx 26$ for CelebA [17], and our d_{latent} sweeps bracket these values. Three things worth noting on top of what the figures in Appendix G show: (i) the time before memorization starts grows roughly $8\times$ on MNIST (10k steps at the smallest d_{latent} to 80k at the largest) and roughly $15\times$ on CelebA (30k to 440k), while the time the model needs

to reach good sample quality (FID plateau) is essentially fixed at $\sim 10\,k$ steps on both datasets. (ii) Minimum FID is roughly flat across each sweep, so the delay-in-memorization benefit of a wider buffer does not come at the cost of sample quality. (iii) Even configurations with d_{latent} smaller than the estimated $d_{\text{intrinsic}}$ (MNIST $d_{\text{latent}} = 10$, CelebA $d_{\text{latent}} \in \{20, 25\}$) still show the memorization time growing with d_{latent}. This suggests either that the buffer effect is robust to the precise $d_{\text{latent}} = d_{\text{intrinsic}}$ boundary, or that the published intrinsic-dimension estimates [17] underestimate the effective intrinsic dimensionality of the VAE latent code itself.

4. Random matrix theory of RFNN training dynamics

4.1. Setup and assumptions

We adopt the RFNN framework of Bonnaire et al. [2] unchanged: a two-layer RFNN with frozen first-layer weights $W \in \mathbb{R}^{p \times d_{\text{latent}}}$ drawn i.i.d. from $\mathcal{N}(0,1)$, $\tanh$ activation, and a trained readout $A \in \mathbb{R}^{d_{\text{latent}} \times p}$ initialized to zero. Training is score matching at a single fixed diffusion time $t = 0.01$, and the analysis is taken in the proportional asymptotic limit $d_{\text{latent}}, n, p \to \infty$ with $\psi_p \equiv p/d_{\text{latent}}$ and $\psi_n \equiv n/d_{\text{latent}}$ held fixed. We use $\psi_p = 64$ throughout; ψ_n varies with the choice of d_{latent} across our sweeps. Experimental details for the synthetic data, diffusion process, and sweep settings are in Appendix A.

Training dynamics. Freezing W makes the score-matching loss quadratic in A, so full-batch gradient flow on A is the linear ODE

$$\tfrac{\mathrm{d}}{\mathrm{d}T} A(T) \;=\; -\big(A(T) - A^\star\big)\,\boldsymbol{U}, \tag{2}$$

where $A^\star$ is the train-loss minimizer and the empirical feature correlation matrix is

$$\boldsymbol{U} \;=\; \tfrac{1}{n} \sum_{\mu=1}^{n} \mathbb{E}_\xi\big[\phi(x_t^\mu)\phi(x_t^\mu)^\top\big], \tag{3}$$

$$\phi(x_t) \;:=\; \tfrac{1}{\sqrt{p}}\,\tanh\!\big(\tfrac{1}{\sqrt{d_{\text{latent}}}}\,W x_t\big), \tag{4}$$

with x_t^μ the noisy diffused state of training point μ and ξ the diffusion noise. Diagonalizing Eq. (2) in the eigenbasis of $\boldsymbol{U}$ gives an exponential per-mode evolution

$$a_i(T) - a_i^\star \;=\; \big(a_i(0) - a_i^\star\big)\,e^{-\lambda_i T}, \tag{5}$$

so eigenmode i of $\boldsymbol{U}$ is absorbed by the readout with characteristic timescale $\tau_i = 1/\lambda_i$.

Estimating $\boldsymbol{U}$. $\boldsymbol{U}$ is computed numerically per run: we fix the training set and the random first layer W, draw a Monte Carlo estimate of the expectation over diffusion noise ξ in Eq. (3) (50 noise samples per training point; Appendix B), assemble the $p \times p$ matrix, and then take its eigendecomposition. The reported spectra are the eigenvalues of this empirical estimate; no analytic surrogate is used.

What the eigenvalues mean. Eq. (5) says each λ_i is a learning rate for one feature direction: large λ_i are absorbed quickly, small ones slowly, and the order in which the readout fits its mass is exactly the sorted eigenvalue list. The RFNN is therefore performing weighted PCA on a frozen nonlinear basis: the random map ϕ defines a fixed set of candidate directions in $\mathbb{R}^p$, $\boldsymbol{U}$ is the empirical covariance of the training data after that lift, and λ_i measures how strongly the training samples vote for direction i. Modes are learned in order of how well the data supports them. Our extension operates entirely inside this framework; only the data covariance $\boldsymbol{\Sigma}_{\text{data}}$ changes.

4.2. From two bulks to four

Bonnaire et al. prove (replica limit, isotropic $\mathbf{\Sigma}_{\text{data}} = \mathbb{I}$) that the spectrum of $\boldsymbol{U}$ splits into a population bulk ρ_2 and a sample bulk ρ_1. With anisotropic data covariance – variance σ_{sig}^2 on the $d_{\text{intrinsic}}$ signal coordinates and variance $\sigma_\perp^2$ on the $d_{\text{latent}} - d_{\text{intrinsic}}$ null coordinates, with $\sigma_{\text{sig}} \gg \sigma_\perp$ – the sorted spectrum separates into four populations: signal data modes, null-direction data modes, sample-specific modes, and the rank-null tail. The new population is the null-direction data bulk, whose width is $d_{\text{latent}} - d_{\text{intrinsic}}$; it sits between population learning and sample-specific fitting, creating the memorization buffer. We observe this structure in individual spectra and across the d_{latent} sweep (Figs. 2 and 3); additional intuition and sweep-level diagnostics are in Appendix H.1 and Appendix K.

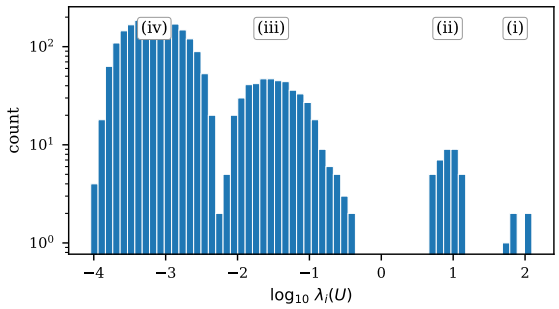


Figure 2: **Clean four-bulk example** ($d_{\text{latent}} = 40$). The four bulks appear as distinct peaks in the eigenvalue density.

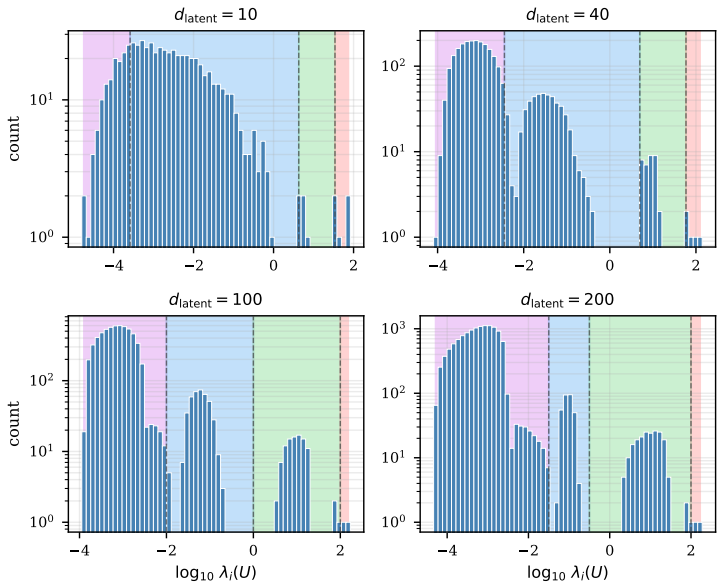


Figure 3: **Four-bulk eigenvalue spectrum, density view** across the d_{latent} sweep at $\sigma_\perp = 0.5$. Signal and noise-dim peaks shift right with d_{latent}; sample mode stays fixed. Additional density and sorted-spectrum diagnostics are in Appendix J and Appendix K.

Mechanism: a noise-dimension buffer. The RFNN absorbs modes in order of decreasing λ_i, so the trajectory traverses the four bulks sequentially: signal bulk first (defining τ_{gen}), noise-dim bulk next (additional training time with no sample-quality benefit), and sample bulk last (defining τ_{mem}). Because the buffer grows as $d_{\text{latent}} - d_{\text{intrinsic}}$, increasing excess latent width delays memorization while leaving τ_{gen} approximately fixed. Appendix H.2 formalizes this picture with bulk counts $d_{\text{intrinsic}}$, $d_{\text{latent}} - d_{\text{intrinsic}}$, n, and $p - d_{\text{latent}} - n$.

5. Discussion and Related Work

Summary. We extended Bonnaire's two-timescale picture of diffusion-model training to anisotropic data and identified a four-bulk eigenvalue structure whose bulk-count boundaries land at indices $d_{\text{intrinsic}}$ and d_{latent} across every configuration we tested. The new noise-dimension bulk acts as a buffer that mechanistically delays memorization, while generalization stays cheap at high ambient noise and grows only mildly at low ambient noise. The same delay holds in trainable MLPs, with one notable feature-learning-driven caveat: a non-monotonic score-error peak at low $\sigma_\perp$ that recovers (though does not fall below the $d_{\text{latent}} = d_{\text{intrinsic}}$ baseline) at large d_{latent}.

Broader impact. Memorization in diffusion models has tangible deployment-time consequences: verbatim training-data extraction [4], near-duplicate generations [19], and downstream legal exposure on copyright and patient-privacy grounds. Production latent-diffusion systems [5, 16, 18] already use $d_{\text{latent}} \gg d_{\text{intrinsic}}$ for reasons of computational efficiency; our analysis identifies a second function of this design choice: the latent-space width acts as a memorization buffer whose width $d_{\text{latent}} - d_{\text{intrinsic}}$ a practitioner can tune without modifying the loss or sacrificing sample quality. As a concrete recipe, our experiments suggest picking $d_{\text{latent}}/d_{\text{intrinsic}} \gtrsim 3$ to put the model in the buffer-dominated regime $\tau_{\text{mem}} \gg \tau_{\text{gen}}$, while ratios $d_{\text{latent}}/d_{\text{intrinsic}} \lesssim 1.5$ leave the buffer too narrow to delay memorization meaningfully. The trade-off is compute: per-step FLOPs and total training time grow roughly linearly with the buffer width, so in compute-bound regimes a smaller d_{latent} paired with explicit regularization may be preferable. Importantly, the buffer lengthens τ_{mem} but does not eliminate it; sufficient training time still drives memorization, and practitioners relying on the buffer for safety must still measure τ_{mem} and stop training in time. More broadly, the four-bulk picture should extend to other block-structured anisotropic data: curved manifolds, hierarchical class structure, and multi-modal sub-populations.

Limitations. (i) The data manifold is a linear Gaussian mixture; curved manifolds and real images are open. (ii) Real-data validation on MNIST and CelebA is in progress. (iii) Quantitative predictions on real data are conditional on the intrinsic-dimension estimator, which disagrees at the $2\times$ level on common image datasets [6, 10, 17, 21]. (iv) All runs are single-seed; multi-seed bands will be added for the camera-ready. (v) The bulk-structure interpretation rests on qualitative random-matrix arguments and the sketch in Appendix H; closed-form bulk-edge formulas remain future work.

References

[1] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26(none):1 – 37, 2021. doi: 10.1214/21-EJP699.

[2] Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 141266–141286. Curran Associates, Inc., 2025.

[3] Bradley CA Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations*, 2023.

[4] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3.

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR, 21–27 Jul 2024.

[6] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017. doi: 10.1038/s41598-017-11873-y.

[7] Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[9] Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024.

[10] Hamidreza Kamkari, Brendan Leigh Ross, Rasa Hosseinzadeh, Jesse Cresswell, and Gabriel Loaiza-Ganem. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 38307–38354. Curran Associates, Inc., 2024. doi: 10.52202/079017-1211.

[11] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38 (1):1 – 50, 2010. doi: 10.1214/08-AOS648.

[12] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26565–26577. Curran Associates, Inc., 2022.

[13] Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saee Gopal Paliwal, Weili Nie, and Arash Vahdat. Genmol: A drug discovery generalist with discrete diffusion. In *Forty-second International Conference on Machine Learning*, 2025.

[14] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[16] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[19] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6048–6058, 2023.

[20] 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.

[21] Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46412–46440. PMLR, 21–27 Jul 2024.

Appendix A. Main-text design details

Synthetic data. We generate Gaussian-mixture data whose signal lies in a $d_{\text{intrinsic}}$-dimensional subspace embedded in a d_{latent}-dimensional ambient space ($d_{\text{latent}} \geq d_{\text{intrinsic}}$), with k cluster centers, intra-cluster signal variance σ_{sig}^2 on the signal subspace, and isotropic null-direction variance $\sigma_\perp^2$. Gaussian mixtures give a closed-form population score (no noise floor), an exact intrinsic/ambient split that lets us vary d_{latent} and $d_{\text{intrinsic}}$ independently, and the block-diagonal $\mathbf{\Sigma}_{\text{data}}$ used by the random-matrix-theory analysis.

Model families. The trainable score model is a 3-layer MLP with sinusoidal time embedding, trained by score-matching MSE against the analytic Gaussian-mixture score. Hidden width is held fixed across the d_{latent} sweep so parameter count and per-step FLOPs stay roughly constant. The RFNN model follows Bonnaire et al. [2]: first-layer features are frozen, only the readout is trained, and the resulting quadratic score-matching loss admits a spectral analysis. We also validate the picture empirically on MNIST and CelebA.

Timescales and sweeps. τ_{gen} is the first step at which test loss reaches within 5% of its per-run minimum. τ_{mem} is the first step at which the Somepalli et al. [19] memorization fraction crosses 1%, with runs censored at their training budget if the threshold is not reached. For the RFNN at fixed t, generation is ill-defined, so we use train-test divergence (gen-gap > 0.02) as the τ_{mem} proxy. We probe the buffer along two axes: varying d_{latent} at fixed $d_{\text{intrinsic}} = 5$, and varying $d_{\text{intrinsic}}$ at fixed $d_{\text{latent}} = 20$.

Appendix B. Experimental methods and design choices

B.1. Data generation

We generate anisotropic Gaussian-mixture data:

- Cluster centers $\{m_c\}_{c=1}^k$ are drawn iid uniformly on a sphere of radius s inside the first $d_{\text{intrinsic}}$ coordinates of $\mathbb{R}^{d_{\text{latent}}}$; the remaining $d_{\text{latent}} - d_{\text{intrinsic}}$ "null-space" coordinates of every center are exactly zero.
- For each training point μ, a cluster index $c(\mu)$ is drawn uniformly, then $x^\mu = m_{c(\mu)} + Q\,\xi^\mu$ where ξ has independent components: variance $\sigma_{\text{sig}}^2 = 1$ along the $d_{\text{intrinsic}}$ signal dimensions and variance $\sigma_\perp^2$ along the $d_{\text{latent}} - d_{\text{intrinsic}}$ null-space dimensions. Q is a fixed orthogonal basis whose first $d_{\text{intrinsic}}$ columns span the signal subspace.
- Default values: $k = 10$, $d_{\text{intrinsic}} = 5$, $s = 3$, $n = 500$ training points, 2048 held-out test points, $\sigma_{\text{sig}} = 1$, $\sigma_\perp \in \{0.01, 0.5\}$.

B.2. Diffusion process

We use the variance-preserving Ornstein-Uhlenbeck process with $t \in [t_{\min}, t_{\max}] = [0.01, 3.0]$. Score targets are computed analytically from the closed-form Gaussian-mixture density at noise level t; this avoids the standard score-matching MSE loss noise floor and gives a clean "score error" metric independent of training-set fluctuations.

B.3. Models

MLP. 3-layer ReLU MLP with sinusoidal time embedding (32 frequency bands), input width $d_{\text{latent}} + 64$, output width d_{latent}. Hidden width $h = 256$ held fixed across the d_{latent} sweep. Adam with learning rate 10^{-4}, batch size 256, $T = 3 \times 10^5$ training steps (extended to 5×10^6 for the body timescale-scatter figure), evaluation every 5,000 steps with 5,000 generated samples.

RFNN. Two-layer RFNN with $\tanh$ activation; first layer $W \in \mathbb{R}^{p \times d_{\text{latent}}}$ is fixed at random Gaussian initialization, the second-layer readout A is trained. $p = 64\, d_{\text{latent}}$ to keep the rank-null bulk visible at every d_{latent}. Same optimizer, batch, and evaluation cadence as the MLP.

B.4. Evaluation metrics

- **Score error**: integrated mean-squared error of the learned score against the analytic Gaussian-mixture score, averaged over 5,000 test points.
- **Memorization fraction**: fraction of generated samples whose nearest training-set neighbor is closer than one third of that neighbor's own nearest training-set neighbor [19].
- τ_{gen}: first step at which test loss is within 5% of its per-run minimum.
- τ_{mem}: first step at which the memorization fraction exceeds 1%. Censored to ">300k" if not crossed within budget.

Why these definitions. We use the test-loss-plateau criterion for τ_{gen} rather than an FID- or MMD-based criterion because (i) it is well-defined on both the MLP and the RFNN, the latter trained at fixed t with no usable generation process, (ii) the analytic Gaussian-mixture score gives noise-floor-free target values, removing the need for finite-sample FID estimates that fluctuate with the number of generated samples, and (iii) the test-loss plateau aligns directly with the moment when the population score has been internalised. We use the Somepalli et al. [19]-style memorization fraction for τ_{mem} on the MLP because it is the same threshold Bonnaire et al. [2] use, allowing direct comparison with the isotropic baseline. For the fixed-t RFNN, no generation is available; we substitute the gen-gap proxy (test loss $-$ train loss > 0.02), which monotonically tracks the canonical τ_{mem} on the MLP and calibrates the two-track comparison.

B.5. Per-experiment configurations

Both sweeps run on the synthetic Gaussian-mixture data described above. Shared constants across every run: $n = 500$ training points, 2,048 held-out test points, $k = 10$ cluster centers, $s = 3$, $\sigma_{\text{sig}} = 1$, OU diffusion at $t \in [0.01, 3.0]$, seed 42 for both data and the random first layer W. Tables 1 and 2 list the sweeps and training hyperparameters.

B.6. Empirical bulk sizes

Table 3 reports the per-panel sizes of the four detected bulks in Figure 23, ordered left-to-right by increasing eigenvalue (B1: rank-null tail; B4: signal block). $\sigma_\perp = 0.01$, $d_{\text{intrinsic}} = 5$, $n = 500$, $p = 64\, d_{\text{latent}}$.

Table 1: **Sweep configurations.** Both $\sigma_\perp$ values used unless noted.

Model	Vary	Range	Fixed	Width
MLP	d_{latent}	$\{5, 8, 10, 15, 20, 30, 40\}$	$d_{\text{intrinsic}} = 5$	$h = 256$
RFNN	d_{latent}	$\{5, 8, 10, 15, 20, 30, 40\}$	$d_{\text{intrinsic}} = 5$	$p = 64\, d_{\text{latent}}$
RFNN (ext., $\sigma_\perp$=0.5)	d_{latent}	$\{60, 80, 100, 150, 200\}$	$d_{\text{intrinsic}} = 5$	$p = 64\, d_{\text{latent}}$
MLP	$d_{\text{intrinsic}}$	$\{2, 5, 8, 12, 16, 20\}$	$d_{\text{latent}} = 20$	$h = 256$
RFNN	$d_{\text{intrinsic}}$	$\{2, 5, 8, 12, 16, 20\}$	$d_{\text{latent}} = 20$	$p = 1280$

Table 2: **Training hyperparameters.**

	MLP	RFNN
Optimizer	Adam	full-batch GD
Learning rate	10^{-4}	closed-form
Batch size	256	full ($n = 500$)
Total steps T	3×10^5	3×10^5
Eval interval	5,000	5,000
Diffusion time	$t \sim \text{Unif}[0.01, 3.0]$	fixed $t = 0.01$
Time embedding	sinusoidal, 32 frequencies	n/a
Activation	ReLU	$\tanh$
$\boldsymbol{U}$ Monte Carlo	n/a	50 noise samples per training point

Table 3: **Empirical bulk sizes per d_{latent}.** B4 collapses to $d_{\text{intrinsic}} = 5$ as predicted; B3 grows linearly with $d_{\text{latent}} - d_{\text{intrinsic}}$ (the noise-dim buffer); B1 absorbs the rank-null tail.

d_{latent}	B1	B2	B3	B4
5	150	138	27	5
8	344	133	30	5
10	295	305	35	5
15	699	216	40	5
20	916	311	48	5
30	1397	458	60	5
40	1907	578	70	5

B.7. Known pitfalls

The $\tanh$ activation in the RFNN saturates whenever the data scale divided by $\sqrt{d_{\text{latent}}}$ exceeds approximately 2, collapsing all eigenvalues into a degenerate single bulk. Configurations were chosen to keep the operating point well inside the linear regime; see Appendix I.

Appendix C. d_{latent} sweep (synthetic data)

The body sweeps $d_{\text{intrinsic}}$ at fixed $d_{\text{latent}} = 20$, the regime where every configuration crosses the 1% memorization threshold within budget. The dual sweep varies $d_{\text{latent}} \in \{5, 8, 10, 15, 20, 30, 40\}$ at fixed $d_{\text{intrinsic}} = 5$ on the same Gaussian-mixture data, hidden $= 256$, $n = 500$, with $\sigma_\perp = 0.5$

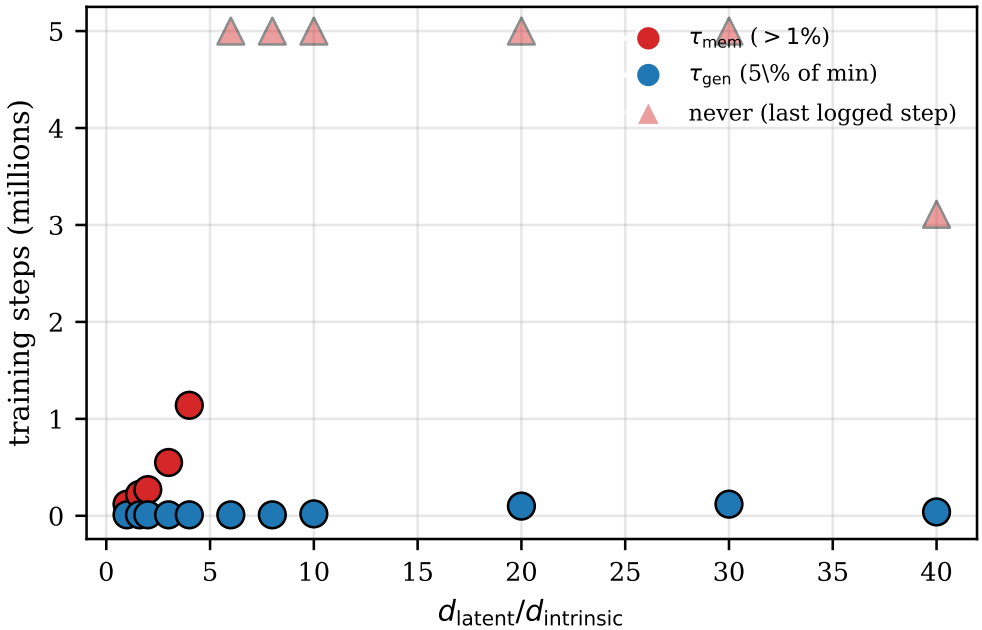


Figure 4: d_{latent} **sweep: timescales versus** $d_{\text{latent}}/d_{\text{intrinsic}}$ **at** $\sigma_\perp = 0.5$**.** 5M-step MLP. Triangles mark runs that do not reach 1% memorization within budget. Dual of body Fig. 1.

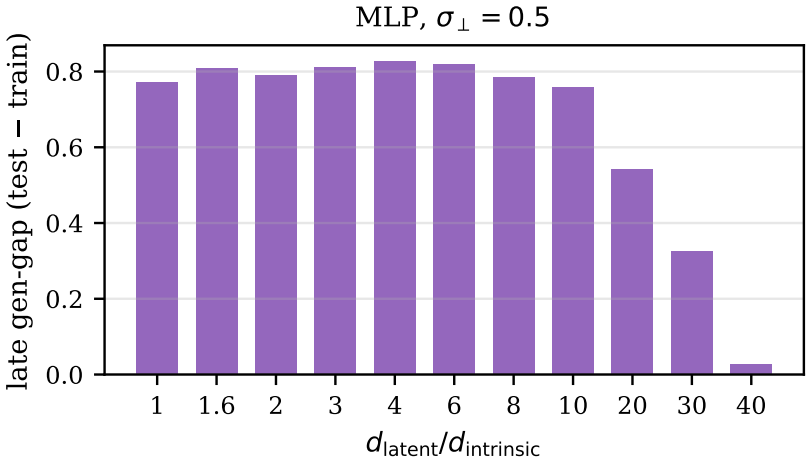


Figure 5: d_{latent} **sweep: late generalization gap versus** $d_{\text{latent}}/d_{\text{intrinsic}}$ **at** $\sigma_\perp = 0.5$**.** Dual of body Fig. 1.

runs at 5M steps and the $\sigma_\perp = 0.01$ counterparts at 300k. Many high-ratio runs in this dual sweep do *not* reach memorization within budget, which is exactly the buffer mechanism prediction.

Appendix D. d_{latent} sweep at $\sigma_\perp = 0.01$

The $\sigma_\perp = 0.01$ counterparts of the d_{latent} sweep figures in Appendix C. The non-monotonic score-error n-shape appears here: when the data/sample cliff is soft, increasing $d_{\text{latent}}/d_{\text{intrinsic}}$ first

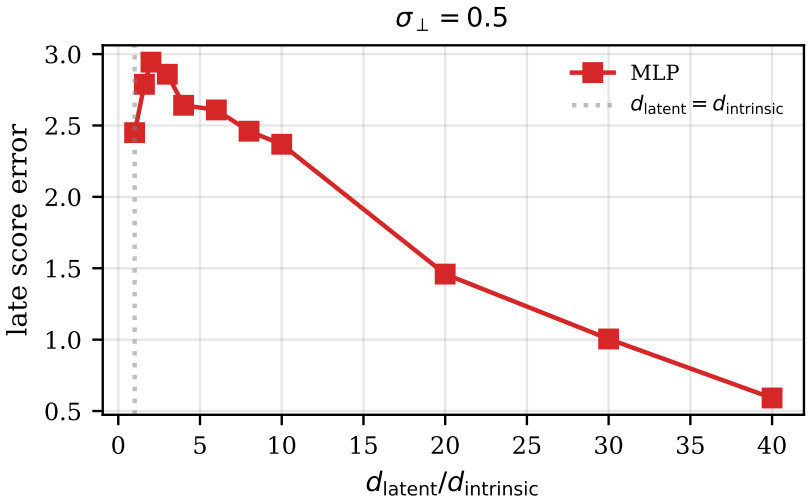


Figure 6: d_{latent} **sweep: late score error versus** $d_{\text{latent}}/d_{\text{intrinsic}}$ **at** $\sigma_\perp = 0.5$**.** This is the wide-range score-error view moved out of the body.

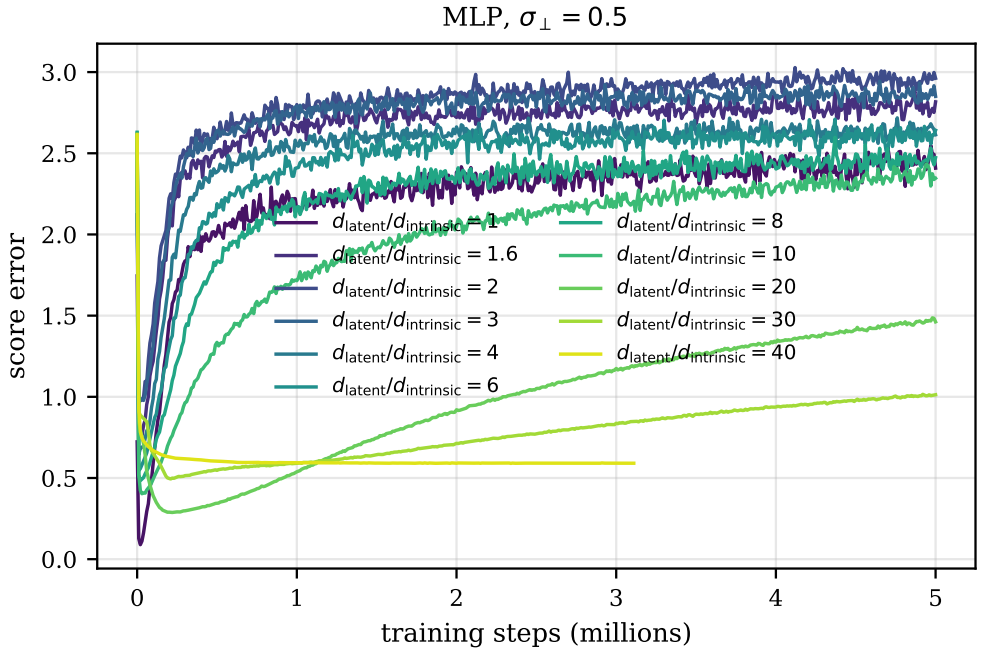


Figure 7: d_{latent} **sweep: score error over training at** $\sigma_\perp = 0.5$**.** Per-step trajectories underlying the aggregate in Fig. 6, one curve per d_{latent} (viridis colormap, low to high). Compare to Fig. 13 (the $d_{\text{intrinsic}}$-sweep counterpart).

worsens late score error and then partially recovers at larger buffer widths. The peak is visible in Fig. 12.

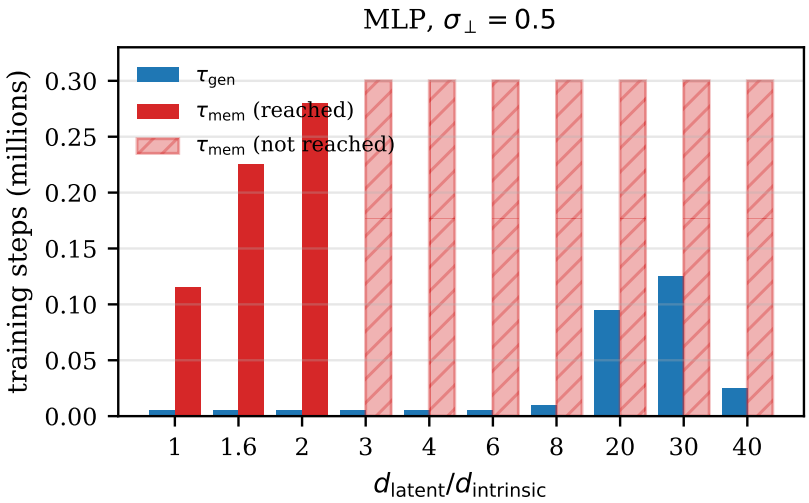


Figure 8: d_{latent} **sweep: 300k-step bar-chart timescales at** $\sigma_\perp = 0.5$**.** The 300k-step bar-chart view of paired τ_{gen} (blue) and τ_{mem} (red), with hatched bars indicating τ_{mem} not reached within budget. The 5M-step extended-budget version is Fig. 4 above.

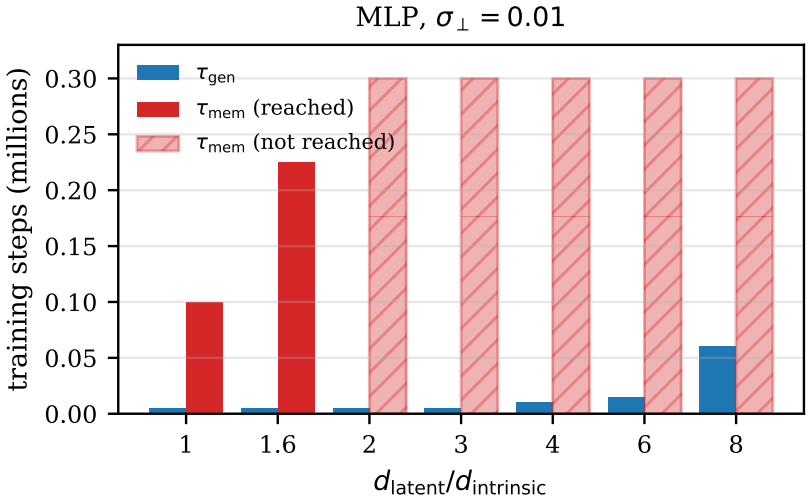


Figure 9: $\sigma_\perp = 0.01$ **counterpart of Fig. 8.**

Appendix E. $d_{\text{intrinsic}}$ sweep additional figures

The body $d_{\text{intrinsic}}$-sweep figures (Figs. 1, 1) report aggregate timescales and gen-gap; this section collects the per-step score-error trajectories and the $d_{\text{intrinsic}}$-sweep late score error, complementing the wider-range d_{latent}-sweep view in Fig. 6.

Appendix F. $d_{\text{intrinsic}}$ sweep at $\sigma_\perp = 0.01$

The $\sigma_\perp = 0.01$ counterparts of the body's $d_{\text{intrinsic}}$ sweep at fixed $d_{\text{latent}} = 20$. These runs are at 300k training steps (the 5M extension was only run at $\sigma_\perp = 0.5$).

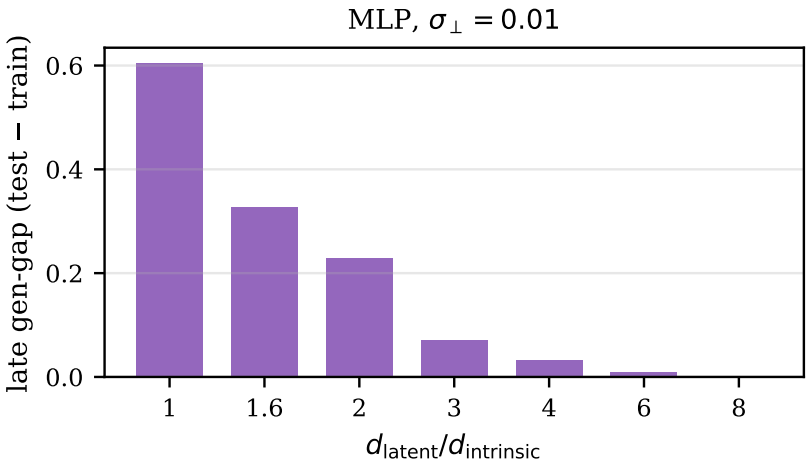


Figure 10: $\sigma_\perp = 0.01$ **counterpart of Fig. 5.**

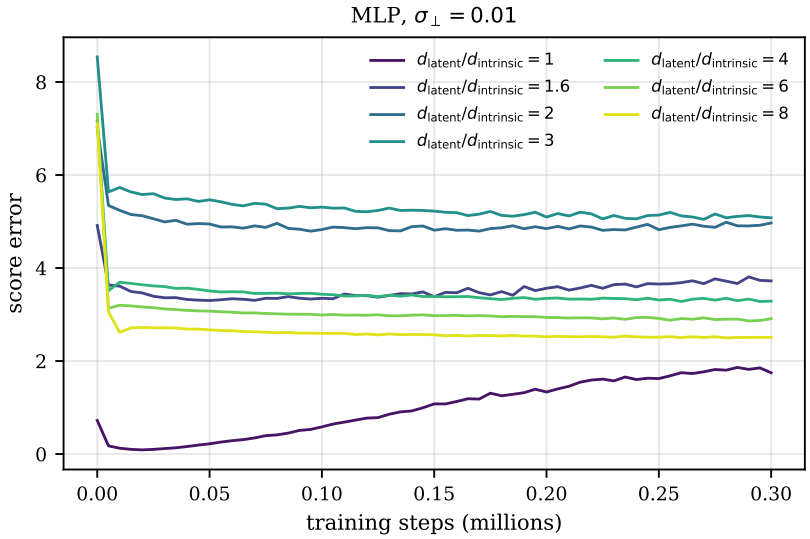


Figure 11: $\sigma_\perp = 0.01$ **counterpart of Fig. 7.**

Appendix G. Real-data experiments

G.1. Setup

We test the buffer mechanism on two real-image datasets, MNIST and CelebA, with the latent space learned end-to-end by a VAE.

Intrinsic dimension. MLE-based estimates on raw pixels [17] place MNIST at $d_{\text{intrinsic}} \approx 13$ and CelebA at $d_{\text{intrinsic}} \approx 26$. Our d_{latent} sweeps—$\{10, 15, 20, 25, 30, 40\}$ for MNIST and $\{20, 25, 30, 35, 40, 50\}$ for CelebA—bracket these values to probe the $d_{\text{latent}} > d_{\text{intrinsic}}$ buffer regime predicted by the synthetic theory. Alternative estimators [6, 10, 21] disagree at the $2\times$ level, so the specific numerical predictions of buffer width $d_{\text{latent}} - d_{\text{intrinsic}}$ are estimator-conditional (Sec. 5).

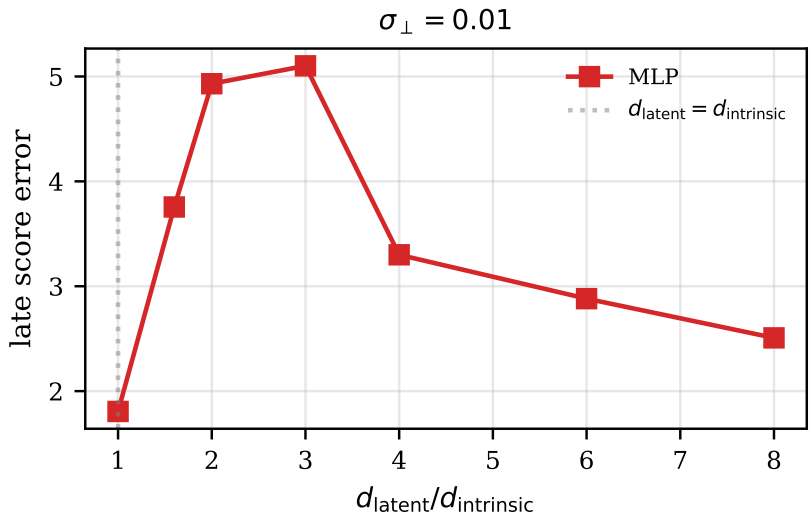


Figure 12: $\sigma_\perp = 0.01$ **counterpart of Fig. 6.** The non-monotonic peak discussed in Sec. 5 is visible here.

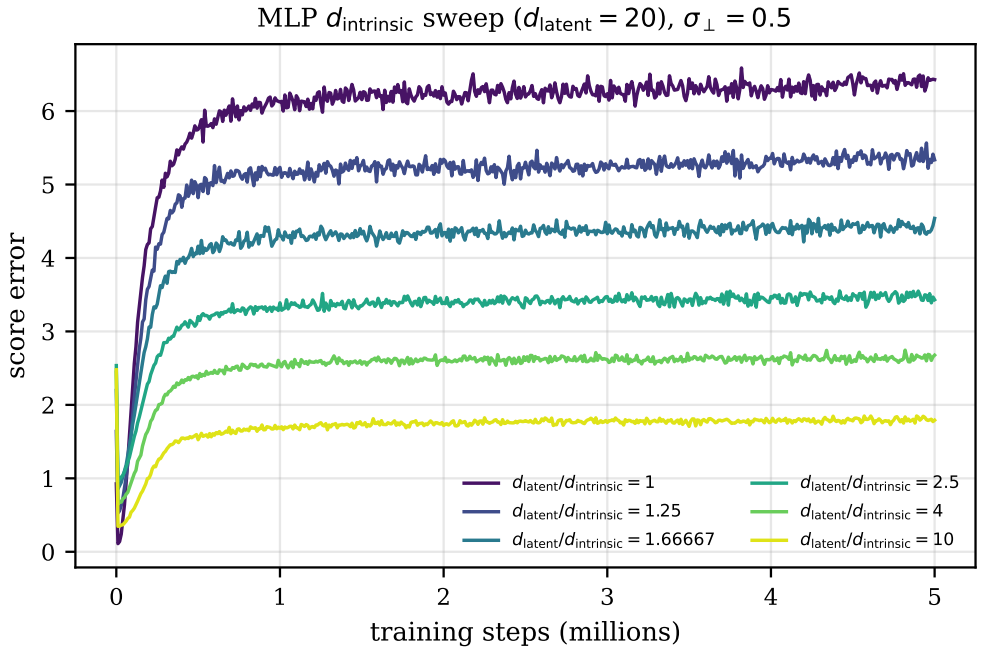


Figure 13: $d_{\text{intrinsic}}$ **sweep: score error over training at** $\sigma_\perp = 0.5$**.** One curve per $d_{\text{latent}}/d_{\text{intrinsic}}$ ratio (viridis colormap). Per-step trajectories underlying the body's $d_{\text{intrinsic}}$-sweep aggregates.

VAE. Trained on half the dataset (the "encoder half"), varying the bottleneck d_{latent} across the same sweep used in synthetic experiments. The remaining half (the "diffusion half") is reserved for diffusion training and evaluation.

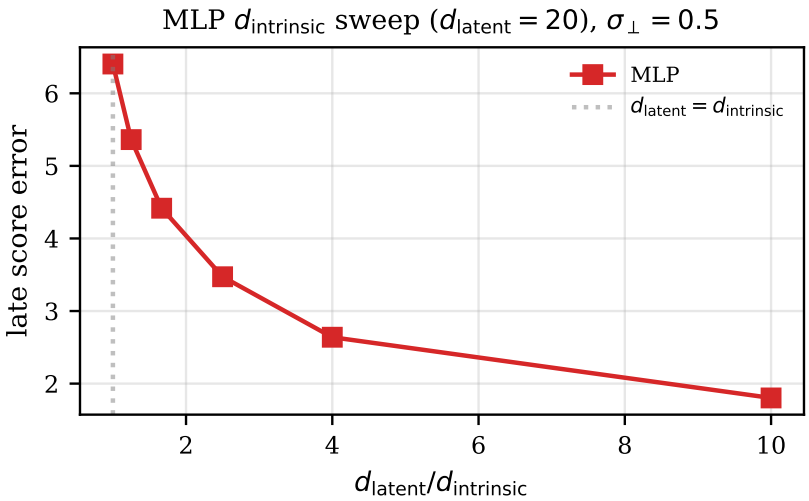


Figure 14: $d_{\text{intrinsic}}$ **sweep: late score error versus** $d_{\text{latent}}/d_{\text{intrinsic}}$ **at** $\sigma_\perp = 0.5$**.** $d_{\text{intrinsic}}$-sweep counterpart of Fig. 6.

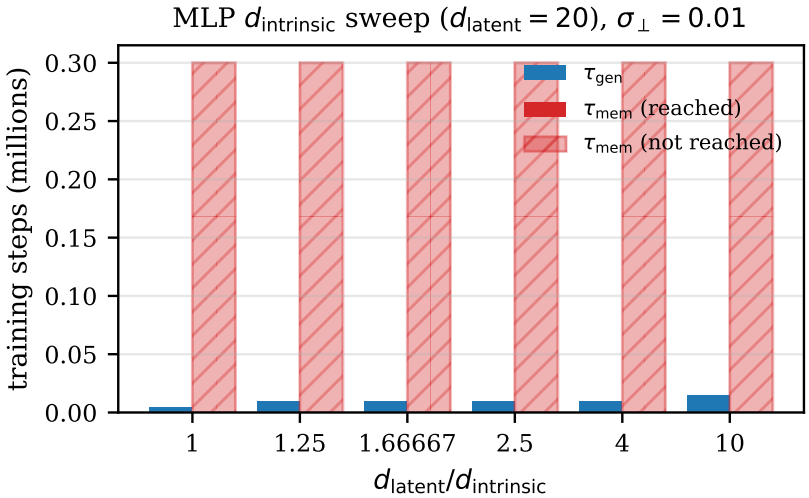


Figure 15: $\sigma_\perp = 0.01$ **counterpart of body Fig. 1.**

Diffusion training set. Each diffusion model is trained on only $n = 1000$ latents drawn from the diffusion half, kept small to put the model in the memorization-prone overparameterized regime.

Score model. Identical architecture to the synthetic MLP (Sec. B): 3-layer ReLU MLP with sinusoidal time embedding, hidden width $h = 256$ held fixed across the d_{latent} sweep, same optimizer, batch size, and step budget.

Metrics.

- τ_{gen}: first step at which FID is within 5% of its per-run minimum, i.e. the FID plateau is reached. We use FID here rather than the synthetic test-loss-plateau criterion because for real data FID is the canonical sample-quality signal.

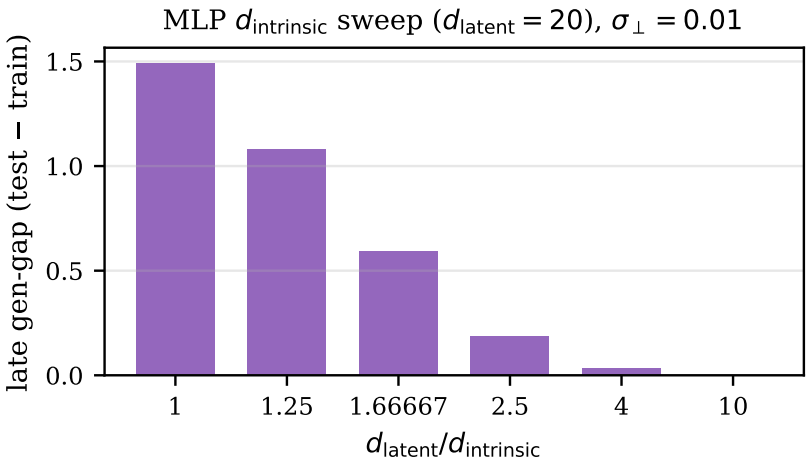


Figure 16: $\sigma_\perp = 0.01$ **counterpart of body Fig. 1.**

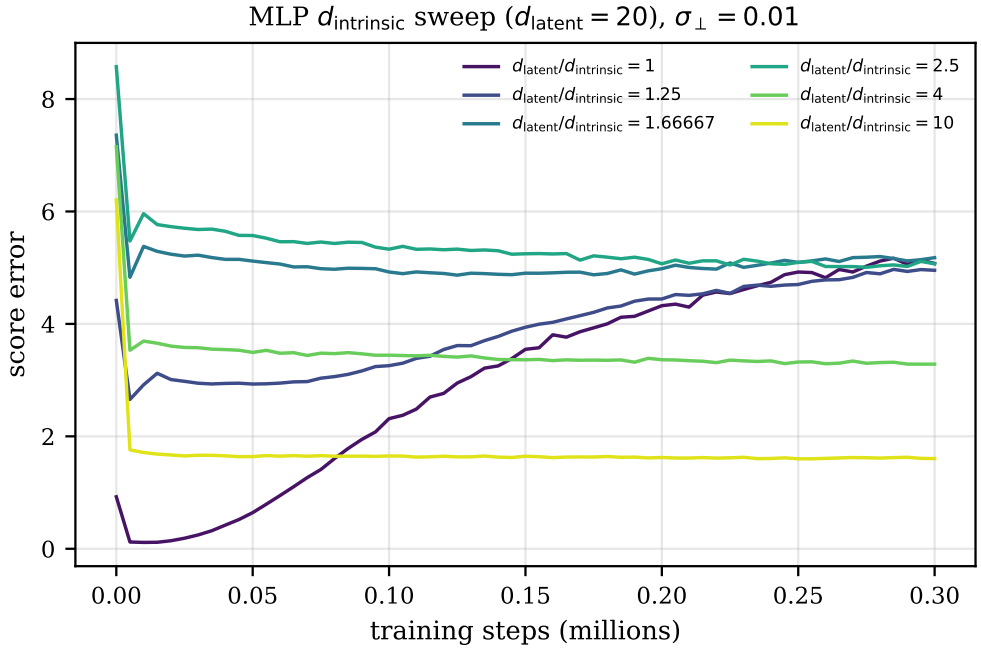


Figure 17: $\sigma_\perp = 0.01$ **counterpart of Fig. 13.**

- τ_{mem}: first step at which the gen-gap proxy (test loss − train loss > 0.02) is crossed; we use the gen-gap proxy here rather than the Somepalli et al. [19]-style memorization fraction because the pixel-space nearest-neighbor threshold rarely triggers at $n = 1000$ in the latent space and the gen-gap proxy monotonically tracks the canonical τ_{mem} on synthetic data.

G.2. Results

The three synthetic findings transfer to real data at $n = 1000$. Fig. 19 shows that τ_{mem} grows with d_{latent} on both datasets, going from $10\,k$ to $80\,k$ steps on MNIST and from $30\,k$ to $440\,k$ steps on

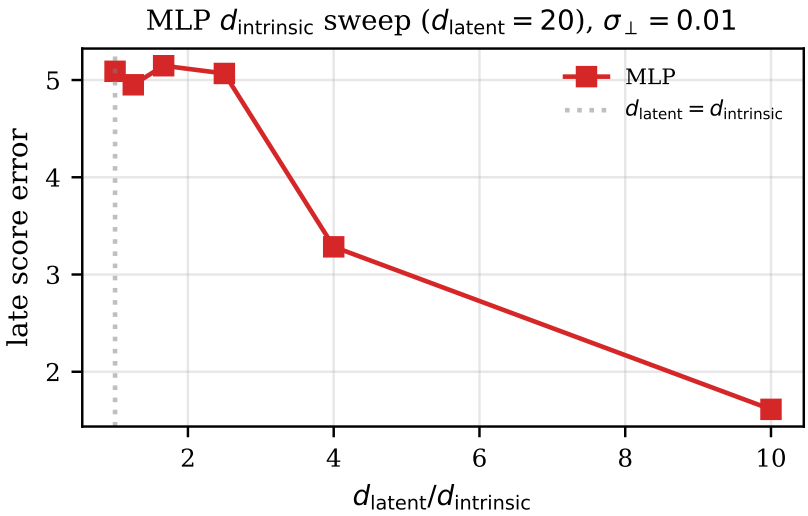


Figure 18: $\sigma_\perp = 0.01$ **counterpart of Fig. 14.**

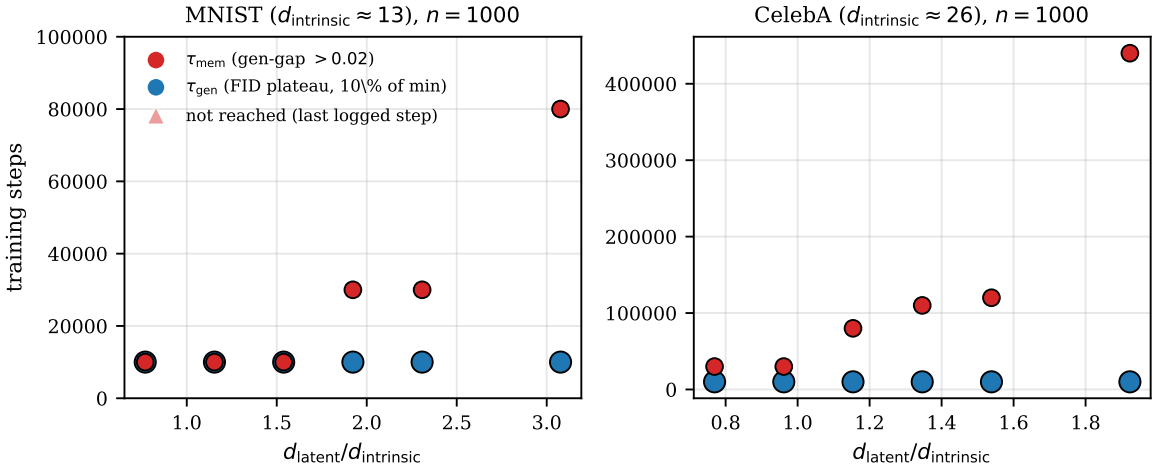


Figure 19: **Real data: timescales versus** $d_{\text{latent}}/d_{\text{intrinsic}}$**.** 5M-step MLP, $n = 1000$. τ_{gen} (blue) is the first step at which test loss is within 5% of its per-run minimum; τ_{mem} (red) is the first step at which the gen-gap proxy crosses 0.02. τ_{mem} grows by roughly $10\times$ across each sweep ($10\,k \to 80\,k$ on MNIST, $30\,k \to 440\,k$ on CelebA), tracking the buffer-mechanism prediction.

CelebA across the sweep—a $\sim 10\times$ delay in memorization onset attributable to the noise-dim buffer. Fig. 20 reports the late gen-gap. Fig. 21 shows minimum FID is roughly flat across the sweep, so the delay benefit of larger d_{latent} does not come at the cost of generation quality; Fig. 22 reports the per-step FID trajectories.

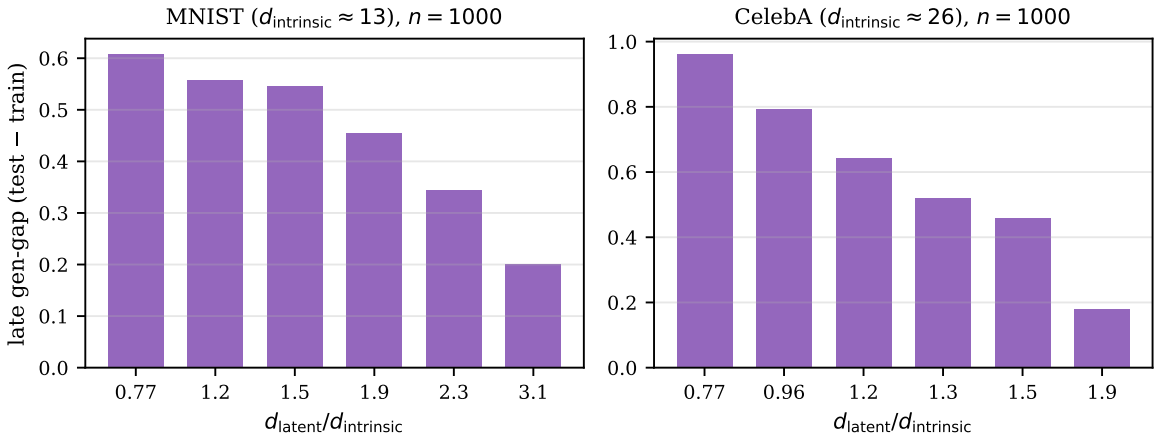


Figure 20: **Real data: late generalization gap versus** $d_{\text{latent}}/d_{\text{intrinsic}}$**.** Mean of test loss − train loss over the last 5% of training.

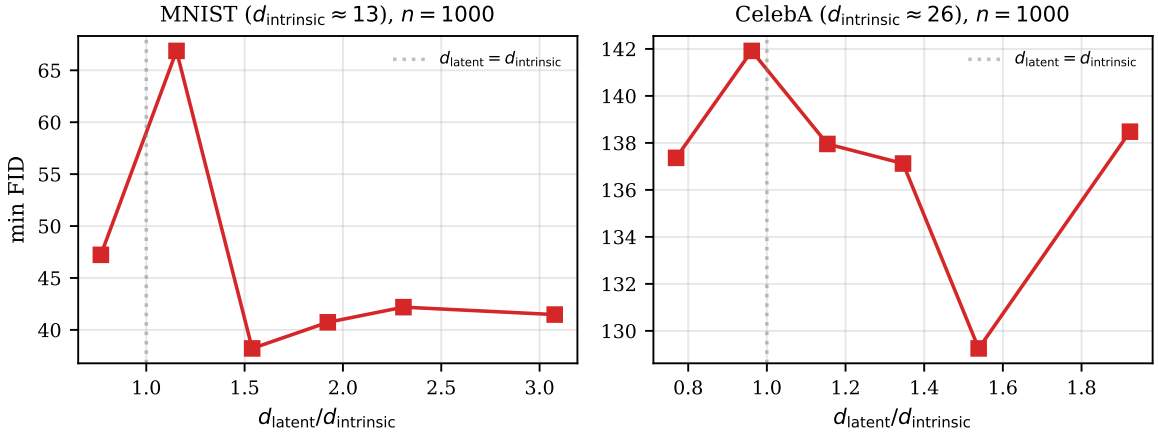


Figure 21: **Real data: minimum FID versus** $d_{\text{latent}}/d_{\text{intrinsic}}$**.** Sample quality is roughly preserved across the sweep, so the delay-in-memorization benefit of larger d_{latent} does not come at the cost of generation quality.

Appendix H. RFNN: sketch of an analytical derivation

H.1. Four-bulk intuition

Bonnaire et al. [2] note that the population bulk ρ_2 develops internal structure when $\mathbf{\Sigma}_{\text{data}}$ has multiple eigenvalues, but at signal-to-null ratio $3\times$ they treat it as a single deformed bulk. The latent-diffusion regime sits at the opposite extreme ($\sigma_{\text{sig}}^2/\sigma_\perp^2 \approx 10^4$ at $\sigma_\perp = 0.01$, ≈ 4 at $\sigma_\perp = 0.5$): the humps inside ρ_2 separate into well-isolated bulks decades apart, which is why we count them as distinct.

What the gen-gap measures. The training set carries two kinds of structure. The data bulks correspond to *population* structure: directions of variance that exist in the population and happen to also be visible in the training sample. The sample bulk corresponds to *null structure*: directions that look like signal in the n training points but do not exist in the population. The RFNN absorbs

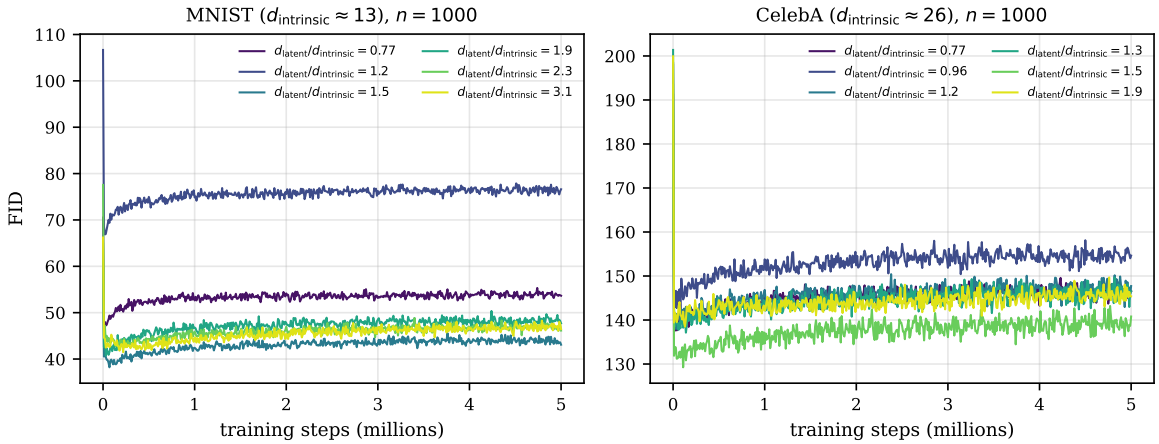


Figure 22: **Real data: FID over training.** Per-step trajectories underlying Fig. 21, one curve per $d_{\text{latent}}/d_{\text{intrinsic}}$ ratio.

eigenmodes in order of decreasing λ_i, so it fits the population structure first and the null structure later. While it is fitting population structure, train and test loss drop together. Once it begins fitting null structure, train loss keeps dropping but test loss does not, and the gap opens. The gen-gap is therefore a direct measure of how much null structure the readout has absorbed, which is why it opens at τ_{mem} and why the noise-dim buffer also delays its onset: widening the buffer pushes back the moment when null structure begins to be learned.

Decomposing the four-bulk structure. The split is easiest to see in sorted-index rather than density language. In the isotropic setting, all population directions have the same scale, so the data-supported part of the spectrum appears as one bulk. In our block-anisotropic setting, the signal coordinates and the null coordinates have different variances before the random feature lift. The Hermite-linear part of the $\tanh$ features therefore produces two data-side eigenvalue scales: a high-eigenvalue signal block of width $d_{\text{intrinsic}}$ and a lower-eigenvalue null-direction block of width $d_{\text{latent}} - d_{\text{intrinsic}}$.

Equivalently, the four populations can be viewed along two axes: shared across the training set versus sample-specific, and signal subspace versus null subspace. The four resulting groups are: (i) BULK_DATA_SIGNAL, the $d_{\text{intrinsic}}$ meaningful signal directions shared by the population; (ii) BULK_DATA_NOISE, the $d_{\text{latent}} - d_{\text{intrinsic}}$ null directions that carry small but population-level variance; (iii) BULK_SAMPLE_SIGNAL, sample-specific fluctuations inside the signal subspace; and (iv) BULK_SAMPLE_NOISE, sample-specific fluctuations inside the null subspace. The first two refine Bonnaire's population bulk ρ_2, while the latter two make up the sample-specific bulk ρ_1.

The signal and noise-dim bulks of $\boldsymbol{U}$ inherit the block structure of $\boldsymbol{\Sigma}_{\text{data}}$ directly: $X^\top X/n$ on the same training data has $d_{\text{intrinsic}}$ large eigenvalues in the signal subspace and $d_{\text{latent}} - d_{\text{intrinsic}}$ small eigenvalues in the null directions, separated by roughly $\sigma^2_{\text{sig}}/\sigma^2_{\perp}$, and the lift by W carries them into the data bulks of $\boldsymbol{U}$ up to a constant from W's spectrum. The sample and rank-null bulks are feature-space artifacts of n training points lifted into $p > d_{\text{latent}}$ random features. The buffer mechanism is thus gradient flow following the data's own variance hierarchy. As d_{latent} grows at fixed $d_{\text{intrinsic}}$, the signal block keeps the same width, but the null-direction block widens. In the sorted spectrum, the cliff associated with sample-specific fitting is pushed to later indices, while the signal-learning cliff stays tied to the intrinsic dimension. Thus the extra latent coordinates form

a trainable population buffer: they absorb optimization time after the model has learned signal directions but before it reaches sample-specific directions.

H.2. Main-text theoretical statement

Using a Hermite expansion of the RFNN feature map, the four-bulk picture predicts that in the proportional limit $d_{\text{latent}}, n, p \to \infty$, the spectrum of $\boldsymbol{U}$ separates into four contiguous bulks with counts

$$d_{\text{intrinsic}}, \qquad d_{\text{latent}} - d_{\text{intrinsic}}, \qquad n, \qquad p - d_{\text{latent}} - n,$$

corresponding respectively to data-signal, data-noise, sample, and rank-null modes. Combining this spectral decomposition with the per-mode decay in Eq. (5) gives the buffer prediction: increasing $d_{\text{latent}} - d_{\text{intrinsic}}$ adds modes between population learning and sample-specific memorization. Thus τ_{mem} is delayed while τ_{gen} remains approximately fixed.

Equivalently, the memorization delay scales with the width of the noise-dimension bulk,

$$\begin{aligned} \tau_{\text{mem}} - \tau_{\text{gen}} &\gtrsim (d_{\text{latent}} - d_{\text{intrinsic}}) \frac{1}{\lambda_{\min}^{\text{noise-dim}}} \\ &\approx (d_{\text{latent}} - d_{\text{intrinsic}}) \frac{\psi_p}{\mu_1^2 \beta_t^2}. \end{aligned} \tag{6}$$

This is the formal version of the noise-dimension buffer mechanism: larger excess latent dimensionality lengthens the path from generalization to memorization without changing the leading signal timescale.

H.3. Derivation sketch

We extend Bonnaire et al.'s replica computation to block-diagonal data covariance $\boldsymbol{\Sigma}_{\text{data}} = \operatorname{diag}\big(\sigma_{\text{sig}}^2 \mathbb{I}_{d_{\text{intrinsic}}}, \sigma_{\perp}^2 \mathbb{I}_{d_{\text{latent}} - d_{\text{intrinsic}}}\big)$. Throughout, $x^\mu = m_{c(\mu)} + \xi^\mu$ with $m_{c(\mu)}$ a cluster center in the $d_{\text{intrinsic}}$-dimensional signal subspace and $\xi^\mu \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\text{data}})$. The diffused state at time t is $x_t^\mu = e^{-t} x^\mu + \sqrt{\Delta_t}\, \eta^\mu$ with $\Delta_t = 1 - e^{-2t}$ and $\eta^\mu \sim \mathcal{N}(0, \mathbb{I}_{d_{\text{latent}}})$. We work in the proportional limit $d_{\text{latent}}, n, p \to \infty$ with $\psi_p \equiv p/d_{\text{latent}}$ and $\psi_n \equiv n/d_{\text{latent}}$ fixed and $\psi_p > 1 + \psi_n$ (so the rank-null tail has room to exist).

Hermite expansion of $\tanh$**.** For data scale s small compared with $\sqrt{d_{\text{latent}}}$ (Appendix I), the pre-activations $W x_t^\mu / \sqrt{d_{\text{latent}}}$ have order-one variance per coordinate and stay inside the analytic regime of $\tanh$. Let $\mu_k = \mathbb{E}_{z \sim \mathcal{N}(0,1)}[h_k(z) \tanh(z)]$ denote the Hermite coefficients of $\tanh$ against the probabilists' Hermite polynomials $\{h_k\}$; $\tanh$ is odd so $\mu_k = 0$ for even k, and numerically $\mu_1 \approx 0.606$, $\mu_3 \approx -0.099$. For each pair of inputs $x, y \in \mathbb{R}^{d_{\text{latent}}}$ the random vector $((Wx)_a, (Wy)_a)$ is jointly Gaussian with mean zero and covariance $\operatorname{diag}(\|x\|^2, \|y\|^2)/d_{\text{latent}}$ plus off-diagonal $x^\top y / d_{\text{latent}}$, *regardless of whether x or y has nonzero mean as a sample* (the cluster centers shift the distribution of x, not the conditional Gaussian over W at fixed x). Consequently the Hermite expansion of $\tanh$ against this Gaussian gives $\mu_0 = 0$ (odd function against a zero-mean Gaussian), and the pairwise kernel admits the proportional-limit decomposition [1, 15]

$$\mathbb{E}_W\big[\tanh\big((Wx)_a\big) \tanh\big((Wy)_a\big)\big] \;=\; \mu_1^2 \frac{x^\top y}{d_{\text{latent}}} \;+\; \eta_\star(x,y) \mathbb{1}[x = y] \;+\; o_d(1), \tag{7}$$

where $\eta_\star(x,x) = \sum_{k\geq 3} \mu_k^2 \, (\|x\|^2/d_{\text{latent}})^k$ is the higher-Hermite mass that survives only on the diagonal. Karoui [11] establishes the same off-diagonal/diagonal split at the level of kernel random matrices.

Two-piece decomposition of $\boldsymbol{U}$. Substituting (7) into $\boldsymbol{U} = \frac{1}{n}\sum_\mu \mathbb{E}_\eta[\phi(x_t^\mu)\phi(x_t^\mu)^\top]$, $\phi(x) = \frac{1}{\sqrt{p}}\tanh(Wx/\sqrt{d_{\text{latent}}})$, and averaging over the diffusion noise yields, to leading order in $1/d_{\text{latent}}$,

$$\boldsymbol{U} \;=\; \boldsymbol{U}^{\text{lin}} + \boldsymbol{U}^{\text{diag}} + R_t, \qquad \|R_t\|_{\text{op}} = o_d(1), \tag{8}$$

where

$$\boldsymbol{U}^{\text{lin}} \;=\; \frac{\mu_1^2}{p\, d_{\text{latent}}}\, W\, M_t\, W^\top, \tag{9}$$

$$\boldsymbol{U}^{\text{diag}} \;=\; \frac{1}{p\, n}\sum_{\mu=1}^{n} \eta_\star\big(x_t^\mu,\, x_t^\mu\big)\, \phi^\perp(x_t^\mu)\, \phi^\perp(x_t^\mu)^\top, \tag{10}$$

and $\phi^\perp(x)$ is the residual feature vector after projecting out the Hermite-1 component. By construction $\mathbb{E}_W[\phi^\perp(x)\phi^\perp(y)^\top] = 0$ for $x \neq y$ and $\mathbb{E}_W[\|\phi^\perp(x)\|^2] = \eta_\star(x,x) + o_d(1)$, so the rank-one outer products in (10) are mutually W-orthogonal in the proportional limit.

Block decomposition of M_t. Direct computation gives

$$M_t \;=\; \tfrac{1}{n}\sum_\mu \mathbb{E}_\eta\big[x_t^\mu (x_t^\mu)^\top\big] \;=\; e^{-2t}\big(\hat{C} + \hat{S}\big) + \Delta_t\, \mathbb{I}_{d_{\text{latent}}}, \tag{11}$$

with $\hat{C} = \frac{1}{n}\sum_\mu m_{c(\mu)} m_{c(\mu)}^\top$ the empirical centroid Gram and $\hat{S} = \frac{1}{n}\sum_\mu \xi^\mu(\xi^\mu)^\top$ the empirical within-cluster covariance. In population, $\mathbb{E}\hat{C}$ has rank $d_{\text{intrinsic}}$ with eigenvalues $\Theta(s^2/d_{\text{intrinsic}})$ in the signal subspace and $\mathbb{E}\hat{S} = \boldsymbol{\Sigma}_{\text{data}}$ contributes σ_{sig}^2 on the signal block and $\sigma_\perp^2$ on the null block. Writing $\alpha_t^2 \equiv e^{-2t}(s^2/d_{\text{intrinsic}} + \sigma_{\text{sig}}^2) + \Delta_t$ for the signal-block eigenvalue and $\beta_t^2 \equiv e^{-2t}\sigma_\perp^2 + \Delta_t$ for the null-block eigenvalue, the population M_t has the block-diagonal spectrum $\text{spec}(M_t) = \{\alpha_t^2\}^{d_{\text{intrinsic}}} \cup \{\beta_t^2\}^{d_{\text{latent}} - d_{\text{intrinsic}}}$ in the Q-basis (the cluster signal subspace and its orthogonal complement).

Linear part: signal and noise-dim bulks.

Lemma 1 (Rank and block structure of $\boldsymbol{U}^{\text{lin}}$) $\text{rank}(\boldsymbol{U}^{\text{lin}}) = d_{\text{latent}}$ *almost surely, and its nonzero spectrum is the spectrum of M_t rescaled by μ_1^2/ψ_p up to Marchenko–Pastur fluctuations of relative size $O(1/\sqrt{\psi_p})$. In particular it splits into a* signal bulk *of size $d_{\text{intrinsic}}$ with eigenvalues $\Theta(\mu_1^2\alpha_t^2/\psi_p)$ and a* noise-dim bulk *of size $d_{\text{latent}} - d_{\text{intrinsic}}$ with eigenvalues $\Theta(\mu_1^2\beta_t^2/\psi_p)$.*

Proof W has i.i.d. Gaussian entries, so $W^\top W/d_{\text{latent}}$ is full rank a.s. for $p \geq d_{\text{latent}}$, with $\mathbb{E}[W^\top W/d_{\text{latent}}] = \psi_p\, \mathbb{I}_{d_{\text{latent}}}$ and operator-norm fluctuations $O(1/\sqrt{\psi_p})$ [15]. The nonzero eigenvalues of $WM_tW^\top$ equal those of $M_t^{1/2}W^\top W M_t^{1/2}$, which are the eigenvalues of M_t rescaled by ψ_p up to those fluctuations. The block counts and eigenvalue magnitudes follow from (11). ■

Diagonal part: the sample bulk.

Lemma 2 (Rank and edge of U^{diag}) *In the proportional limit,* $\mathrm{rank}(\boldsymbol{U}^{\mathrm{diag}}) = n$ *almost surely, and its nonzero eigenvalues are concentrated around* $\bar{\eta}_\star/\psi_p$ *where* $\bar{\eta}_\star = \mathbb{E}_{x\sim P_t}[\eta_\star(x,x)] = \sum_{k\geq 3}\mu_k^2\,\mathbb{E}[(\|x\|^2/d_{\mathrm{latent}})^k]$ *with* $x \sim P_t$ *a sample of the diffused mixture distribution.*

Proof $\boldsymbol{U}^{\mathrm{diag}}$ in (10) is a sum of n rank-one outer products $v_\mu v_\mu^\top$ scaled by $\Lambda_\mu = \eta_\star(x_t^\mu, x_t^\mu)/(pn)$, where $v_\mu = \phi^\perp(x_t^\mu)/\|\phi^\perp(x_t^\mu)\|$. The kernel decomposition (7) gives $v_\mu^\top v_\nu = \delta_{\mu\nu} + o_d(1)$, so the $\{v_\mu\}$ are asymptotically orthonormal and the rank is n for $n < p$. The eigenvalues are $\Lambda_\mu \cdot \|\phi^\perp(x_t^\mu)\|^2 \to \eta_\star(x_t^\mu, x_t^\mu)^2/(pn)$ and concentrate around their mean as a self-averaging quadratic form in the Gaussian noise component of x_t^μ. ■

Remark 3 (Why a linear-only derivation cannot resolve the sample bulk) $\boldsymbol{U}^{\mathrm{lin}}$ *has rank exactly* d_{latent}*, so any spectral content beyond the two data-side blocks must come from a different contribution. The Hermite-*$k \geq 3$ *terms in* (7) *are the smallest piece of* tanh *on off-diagonal kernel entries but the* only *piece that survives on the diagonal, and that survival is what produces the rank-*n *sample bulk. An earlier version of this derivation tried to extract the sample bulk from the rank-*n *fluctuation* $\hat{S} - \boldsymbol{\Sigma}_{\mathrm{data}}$ *in* (11)*; those modes lie inside the rank-*d_{latent} *range of* $\boldsymbol{U}^{\mathrm{lin}}$ *and contribute Marchenko–Pastur deformations of the two data-side bulks, not a separately gapped class.*

Remark 4 (Fixed-W vs. W-averaged kernel) *The empirical* $\boldsymbol{U}$ *averages over the diffusion noise and training data at* fixed W*, so the kernel decomposition* (7)*, which holds in expectation over* W*, must be upgraded to a concentration statement. By Benigni and Péché [1] for sub-Gaussian entries of* W *(in particular Gaussian, our case) the off-diagonal kernel entries* $\phi(x_t^\mu)^\top\phi(x_t^\nu)$ *concentrate around their* W*-mean at rate* $O(1/\sqrt{p})$ *in the proportional limit, and the diagonal entries concentrate at the same rate. The rank-*n *structure of* $\boldsymbol{U}^{\mathrm{diag}}$ *therefore survives at fixed* W *up to operator-norm corrections of order* $n/\sqrt{p} = \sqrt{\psi_n/(\psi_p p)} \to 0$*, which is the self-averaging argument referenced implicitly in the Lemma 2 proof.*

Remark 5 (Marchenko–Pastur spreading of the bulk edges) *Each of the four bulks in Theorem 6 is a single* leading-order *edge; the actual bulk in finite* d_{latent} *is a Marchenko–Pastur-deformed band around that edge with relative width* $\Delta_{\mathrm{MP}} = (\sqrt{\psi_p}+1)^2/(\sqrt{\psi_p}-1)^2 - 1 \approx 4/\sqrt{\psi_p}$ *at large* ψ_p*. At our default* $\psi_p = 64$ *this gives* $\Delta_{\mathrm{MP}} \approx 0.50$*, so the two data-side bulks remain spectrally resolvable whenever* $\alpha_t^2/\beta_t^2 \gtrsim 1.5$*; the empirical* $\alpha_t^2/\beta_t^2 \geq 4$ *at* $\sigma_\perp = 0.5$ *and* $\geq 10^4$ *at* $\sigma_\perp = 0.01$ *satisfy this comfortably. The noise-dim-to-sample gap is governed instead by the relative size of the Hermite-1 and Hermite-*≥ 3 *coefficient masses; see the "Bulk-gap scaling" paragraph below.*

Four-bulk theorem.

Theorem 6 (Four-bulk structure) *Under the setup above and in the proportional asymptotic limit, the spectrum of* $\boldsymbol{U}$ *in* (8) *splits almost surely into four contiguous bulks with the following counts and leading-order edges (fixed* $t \in (0,\infty)$*):*

Bulk	*Count*	*Leading-order edge*
Signal	$d_{\mathrm{intrinsic}}$	$\mu_1^2\,\alpha_t^2/\psi_p \;=\; \mu_1^2\big[e^{-2t}(s^2/d_{\mathrm{intrinsic}} + \sigma_{\mathrm{sig}}^2) + \Delta_t\big]/\psi_p$
Noise-dim	$d_{\mathrm{latent}} - d_{\mathrm{intrinsic}}$	$\mu_1^2\,\beta_t^2/\psi_p \;=\; \mu_1^2\big[e^{-2t}\sigma_\perp^2 + \Delta_t\big]/\psi_p$
Sample	n	$\bar{\eta}_\star/\psi_p$
Rank-null	$p - d_{\mathrm{latent}} - n$	$o(1)$

Proof Combine Lemmas 1 and 2 with the decomposition (8). The ranges of $\boldsymbol{U}^{\text{lin}}$ and $\boldsymbol{U}^{\text{diag}}$ are asymptotically W-orthogonal: the former lies in the column space of W (rank d_{latent}), and the latter is spanned by the n residual directions $\phi^{\perp}(x_t^{\mu})$ which are orthogonal to that column space in expectation. Together they span a subspace of dimension $d_{\text{latent}} + n < p$, leaving a rank-null tail of dimension $p - d_{\text{latent}} - n$ with eigenvalues vanishing as $o_d(1)$ from R_t in (8). ■

Bulk-gap scaling. The signal-to-noise-dim gap is the eigenvalue ratio $\alpha_t^2/\beta_t^2 = \big(e^{-2t}(s^2/d_{\text{intrinsic}} + \sigma_{\text{sig}}^2) + \Delta_t\big)/\big(e^{-2t}\sigma_{\perp}^2 + \Delta_t\big)$. At small diffusion times this is approximately $(s^2/d_{\text{intrinsic}} + \sigma_{\text{sig}}^2)/\sigma_{\perp}^2$ (large at $\sigma_{\perp} = 0.01$, modest at $\sigma_{\perp} = 0.5$); as t grows both bulks are dragged toward the diffusion floor Δ_t and the ratio collapses to 1. The noise-dim-to-sample gap is the ratio $\mu_1^2\beta_t^2/\bar{\eta}_{\star}$, governed by the relative sizes of the linear-Hermite coefficient and the higher-Hermite residual mass: for $\tanh$, $\mu_1^2 \approx 0.367$ while $\bar{\eta}_{\star}$ is dominated by $\mu_3^2 \approx 0.0097$ at our operating scales, so for $\beta_t^2 \gtrsim \bar{\eta}_{\star}/\mu_1^2$ the noise-dim edge sits above the sample edge and the four bulks are spectrally resolvable. Below that threshold (small $\sigma_{\perp}$ at large t) the two merge, which is the regime where the four-bulk cliff at index d_{latent} becomes a smooth step in Fig. 24.

Toward a full Stieltjes derivation. Theorem 6 gives leading-order edge magnitudes that match (within Marchenko–Pastur fluctuations $O(1/\sqrt{\psi_p})$) the empirical bulk locations in Table 3 and Fig. 23. A full Stieltjes-transform / replica derivation that produces the bulk edges as exact roots of a polynomial fixed-point equation, extending the isotropic-$\boldsymbol{\Sigma}_{\text{data}}$ analysis of Bonnaire et al. [2] to the block-diagonal case using the tools of Karoui [11] and Benigni and Péché [1], is left to follow-up work. The route taken above is sufficient for the buffer-mechanism corollary in Section 4, which only requires the bulk counts and leading-order edges established here.

Appendix I. RFNN: eigenvalue saturation and the $\tanh$ pitfall

We observed that for sufficiently large data scale s, the spectrum of $\boldsymbol{U}$ collapses to a single bulk with eigenvalues bunched at $\lambda \approx 1$. The cause is $\tanh$ saturation: when $\|Wx\|/\sqrt{d_{\text{latent}}} \gtrsim 2$, $\tanh$ outputs ± 1 almost deterministically and the feature map becomes effectively constant in x. The constraint $s/\sqrt{d_{\text{latent}}} \lesssim 2$ defines the safe operating regime; all reported sweeps satisfy this constraint.

Appendix J. RFNN: empirical bulk detection

Fig. 23 shows the $\sigma_{\perp} = 0.01\, d_{\text{latent}}$ sweep with the empirical bulk-detection overlay (peaks found by `scipy.signal.find_peaks` on the histogram counts, capped at 4 by prominence). At the larger anisotropy ratio $\sigma_{\text{sig}}^2/\sigma_{\perp}^2 \approx 10^4$ the four bulks separate cleanly and the detected peaks land where the theoretical index-based cuts in Fig. 3 put them.

Appendix K. RFNN: additional 4-bulk spectrum figures

The body shows the $\sigma_{\perp} = 0.5$ density view (Figure 3); for completeness we include the remaining three corners of the $\{\text{cliffs}, \text{density}\} \times \{\sigma_{\perp} = 0.01, 0.5\}$ grid here. Figures 24 and 25 are the cliffs (sorted-spectrum) view at the two noise scales, and Figure 26 is the density view at $\sigma_{\perp} = 0.01$. Figure 27 sweeps $d_{\text{intrinsic}}$ instead of d_{latent}.

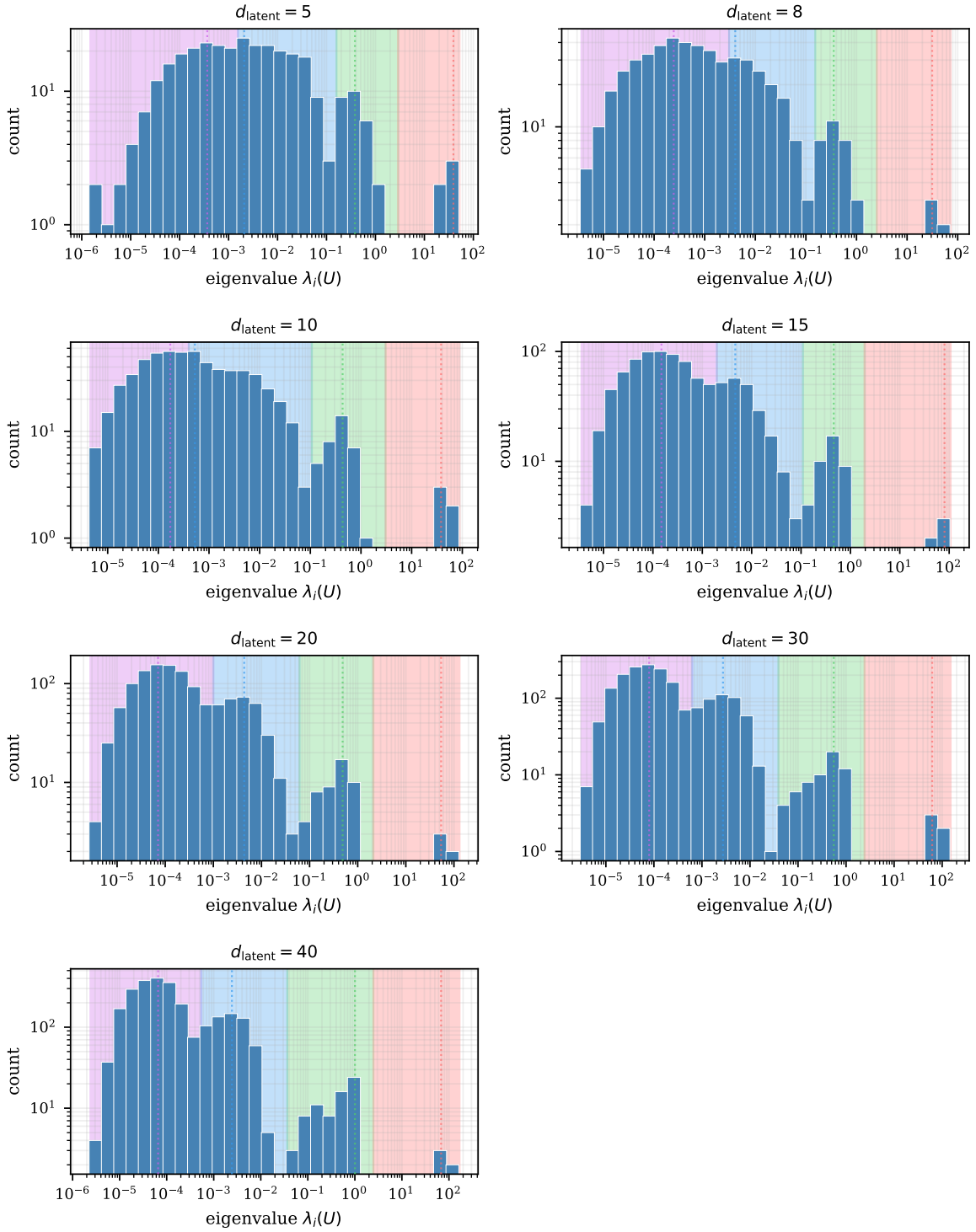


Figure 23: **Empirical bulk detection across the d_{latent} sweep at $\sigma_\perp = 0.01$.** Each colored region is one detected peak in the eigenvalue density; four bulks are detected in every panel. As $d_{\text{latent}}/d_{\text{intrinsic}}$ grows, the bulks pull apart: at $d_{\text{latent}} = d_{\text{intrinsic}}$ the four peaks sit close together with shallow valleys, and by $d_{\text{latent}} = 40$ ($d_{\text{latent}}/d_{\text{intrinsic}} = 8$) they have separated into well-isolated peaks decades apart in eigenvalue. Per-panel bulk sizes are tabulated in Table 3; detection algorithm in Appendix B.

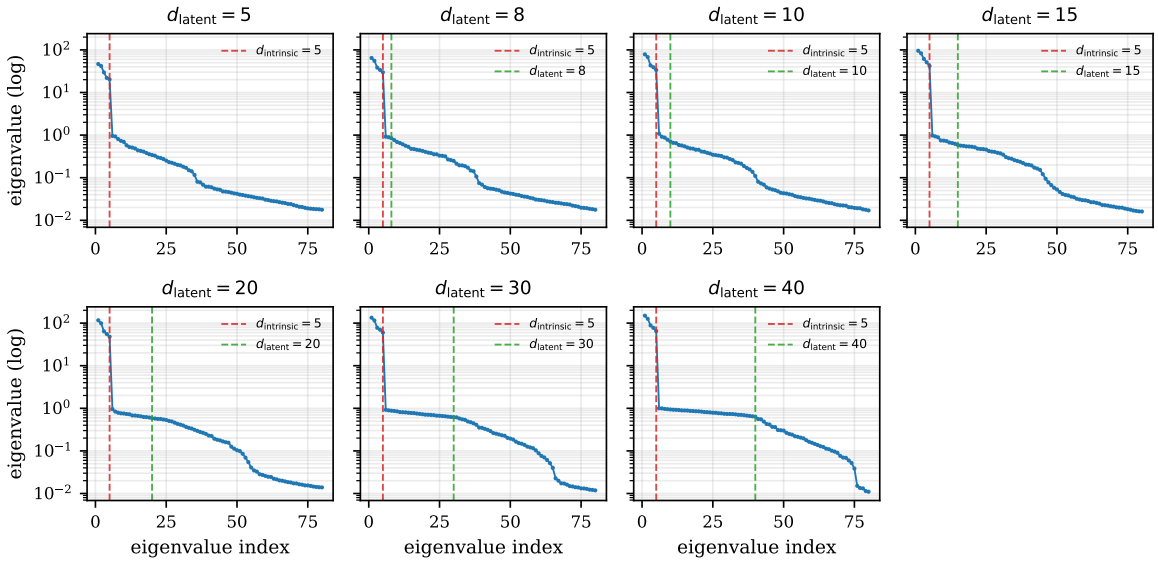


Figure 24: **Cliffs view,** $\sigma_\perp = 0.01$**.** Sorted log-spectrum across the d_{latent} sweep. Red and green dashed lines mark indices $d_{\text{intrinsic}}$ and d_{latent}. The BULK_DATA_NOISE region widens between them as d_{latent} grows.

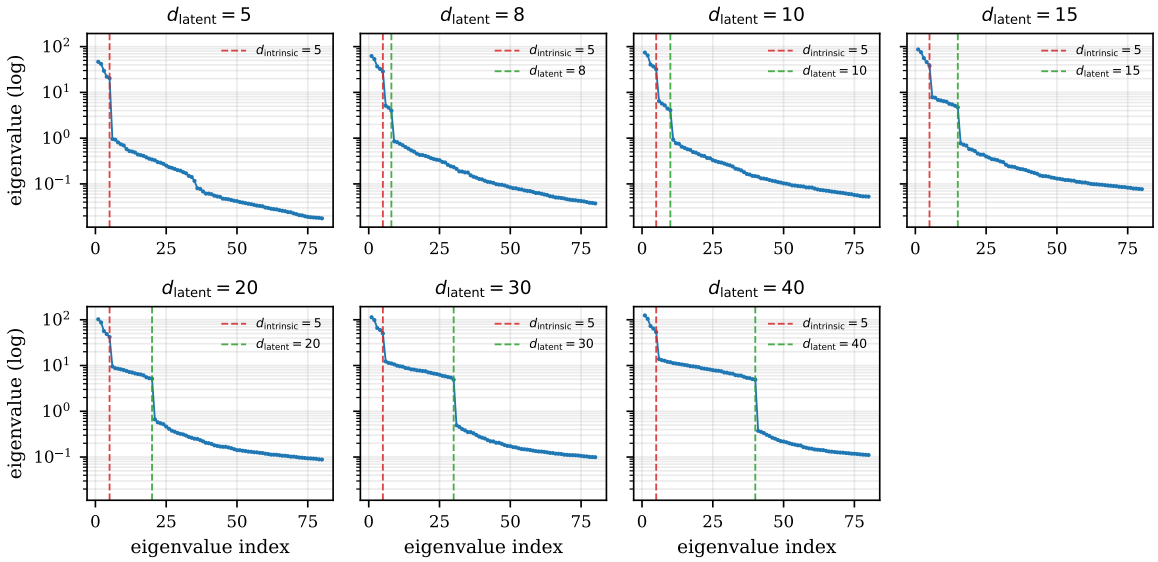


Figure 25: **Cliffs view,** $\sigma_\perp = 0.5$**.** Same sweep as Figure 24 at the higher noise scale: the signal/noise-dim cliff at index $d_{\text{intrinsic}}$ is sharp, while the data/sample cliff at index d_{latent} softens as the gap $\sigma_{\text{sig}}^2/\sigma_\perp^2$ shrinks.

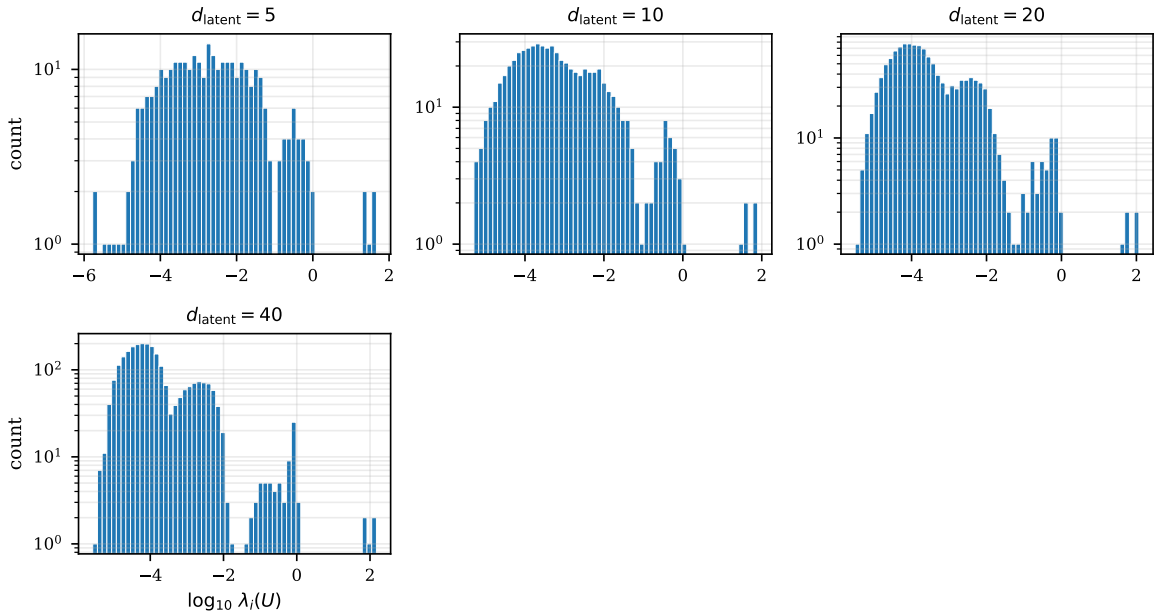


Figure 26: **Density view,** $\sigma_\perp = 0.01$**.** Counterpart to the body's Figure 3. The larger $\sigma_{\text{sig}}^2/\sigma_\perp^2 \approx 10^4$ pushes the signal bulk decades to the right of the noise-dim bulk.

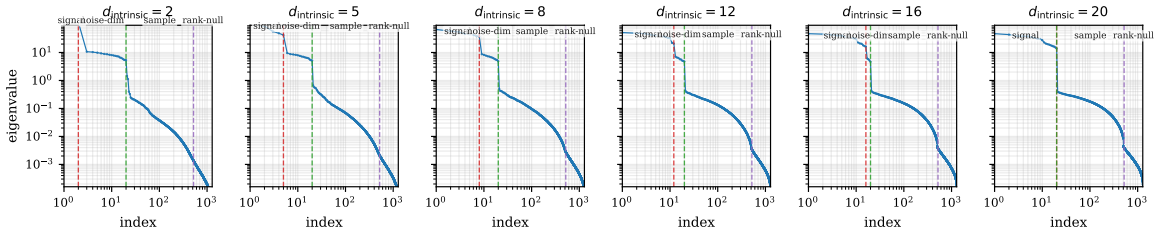


Figure 27: **Varying** $d_{\text{intrinsic}}$ at $d_{\text{latent}} = 20$, $\sigma_\perp = 0.5$. As $d_{\text{intrinsic}} \to d_{\text{latent}}$ the noise-dim bulk shrinks toward zero width and we recover Bonnaire's two-bulk picture.