How Excess Latent Dimensionality Delays Memorization in Diffusion Models

Anonymous authors
Paper under double-blind review

Abstract

Latent diffusion models operate on representations whose dimension can exceed the data's intrinsic dimension. Because diffusion models can reproduce individual training samples, it is important to understand whether this excess width affects memorization. We find that widening the latent space at fixed hidden width delays memorization in the tested vector-latent multilayer perceptron (MLP) score models. In controlled synthetic experiments, the memorized fraction after 5M training steps falls from 30% to 0.1% as the latent dimension grows from 5 to 40. This decline persists when memorization is measured only in the known signal subspace, showing that declining detector sensitivity alone cannot explain the result. On CelebA, above the width at which the variational autoencoder (VAE) preserves sample identity, the mean time to reach 5% memorization grows by $3.6\times$; CIFAR-10 provides supporting evidence. A frozen-feature model gives exact learning dynamics for individual feature directions. Its spectra and a controlled intervention support an explanation based on the variance entering the nonlinearity, although connecting this mechanism to memorization requires additional alignment assumptions. Spectral timing estimates calibrated on trained runs show dataset-dependent accuracy. These results identify latent width as a factor in memorization timing for vector-latent MLPs and motivate treating latent dimensionality as a variable when evaluating memorization and privacy risk in diffusion models. The mechanism in trained networks and transfer to other architectures remain unresolved.

1 Introduction

Score-based diffusion models learn a score network to reverse a noising process (Ho et al., 2020; Song et al., 2021; Karras et al., 2022), and latent diffusion and related flow models generate in learned latent spaces (Rombach et al., 2022; Esser et al., 2024) whose dimension d_{latent} typically exceeds the intrinsic dimension $d_{\text{intrinsic}}$ of the data (Stanczuk et al., 2024; Kamkari et al., 2024). Fitting the empirical score can lead to training-set replication (Yi et al., 2023); collapse during reverse sampling (Biroli et al., 2024) is distinct from the late-training memorization observed in some settings (Bonnaire et al., 2025). Memorization raises privacy and copyright concerns (Carlini et al., 2019; Somepalli et al., 2023; Carlini et al., 2023) and a basic question about what the model learns. We investigate how latent dimension affects its onset.

We study how widening the latent space at fixed hidden width changes memorization timing. Controlled synthetic experiments keep the signal distribution fixed; CelebA and CIFAR-10 experiments test learned image latents. Widening delays memorization in the synthetic sweeps and, once the VAE preserves sample identity, on CelebA. A frozen-feature model lets us examine a possible explanation through exactly solvable training dynamics. Our contributions:

- We measure how latent-space size affects memorization in controlled sweeps at fixed hidden width: on synthetic mixtures the memorized fraction after 5M steps falls from 30% at $d_{\text{latent}} = 5$ to near zero (0.1%) at $d_{\text{latent}} = 40$, and on CelebA the mean time to reach 5% memorization grows $3.6\times$ above the VAE identity threshold (Sections 2 and 4).
- The decline survives a dimension-controlled test restricted to the known five-dimensional signal subspace, falling from 29.9% to 1.2% against a 0.4% fresh-sample baseline (Table 1), so declining metric sensitivity is not the sole explanation.

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

- In the frozen-feature model, increasing d_{latent} lowers the pre-activation variance q of the $\tanh$ features and increases the measured ratio of signal to sample-block edges; holding q fixed largely removes that ratio's endpoint growth ($1.09\times$ rather than $3.17\times$ over $d_{\text{latent}} = 40 \to 240$), supporting an operating-point rather than a mode-count explanation (Section 3).
- We evaluate clocks computed from frozen VAE latent spectra and calibrated on observed hitting times. Held-out-width comparisons show dataset-dependent performance against a dimension-only baseline; these are diagnostics, not uncalibrated pre-training predictions (Section 4).

Our spectral analysis builds on Bonnaire et al. (2025), who relate random-feature diffusion training to generalization and memorization timescales and derive a limiting spectral description for Gaussian data with general clean-data covariance; George et al. (2026) study Gaussian data supported on a subspace. These results motivate our anisotropic-mixture approximation; finite-rank signal modes and noise averaging require additional arguments (Appendix G.1). Our theoretical contribution is the operating-point interpretation and its empirical control. Generative memorization has also been studied through membership inference, reconstruction attacks and data-copying tests (Hayes et al., 2017; Hilprecht et al., 2019; Meehan et al., 2020), and diffusion replication and extraction (Somepalli et al., 2023; Carlini et al., 2023); our nearest-neighbor ratio follows Bonnaire et al. (2025).

2 Synthetic experiments

The mixture's signal structure occupies a subspace of dimension $d_{\text{intrinsic}} = 5$ embedded in a latent space of dimension d_{latent}. In the signal coordinates we sample a Gaussian mixture model (GMM) with $k = 10$ centers of norm 3 and unit variance around each center ($\sigma_{\text{sig}} = 1$); the remaining $d_{\text{latent}} - d_{\text{intrinsic}}$ coordinates receive independent null noise with $\sigma_{\perp} = 0.5$; the full vector is then rotated by a random orthogonal matrix Q so that no coordinate is privileged. The training set has $n = 500$ points. We sweep $d_{\text{latent}} \in \{5, 8, 10, 12, 15, 20, 25, 30, 35, 40\}$ with everything else fixed; because the signal dimension is known, the excess is exact, $d_{\text{latent}} = 5$ has no excess directions, and the population score is known exactly at every d_{latent}.

We train a standard score-based diffusion model with a variance-preserving noising process and an MLP whose hidden width is fixed at 256. The denoising objective, architecture, optimizer and sampling details are in Appendix A.

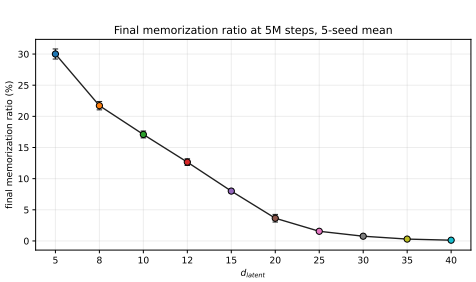
Measuring memorization. A generated sample $\hat{x}$ is counted as memorized when

$$\frac{\|\hat{x} - \mathrm{NN}_1\|}{\|\hat{x} - \mathrm{NN}_2\|} < \frac{1}{3},$$

where NN_1 and NN_2 are its nearest and second-nearest *training* points, following Bonnaire et al. (2025). We report the memorized fraction of generated samples, in latent space for synthetic data and in pixel space after decoding for images, as a mean over seeds. We also compute the ratio after projecting generated and training points onto the true $d_{\text{intrinsic}} = 5$ signal subspace; this signal-projected variant keeps its sensitivity to signal copies approximately constant across the sweep (Table 1).

Figure 1(a) shows the final memorized fraction after 5M steps as the latent dimension is swept from $d_{\text{latent}} = 5$, equal to the intrinsic dimension, to $d_{\text{latent}} = 40$ at a fixed hidden width of 256, five seeds per width. Once d_{latent} exceeds $d_{\text{intrinsic}}$ the fraction falls monotonically: 30.0, 21.7, 17.1, 12.7, 8.0, 3.7, 1.6, 0.8, 0.3 and 0.1% at $d_{\text{latent}} = 5, 8, 10, 12, 15, 20, 25, 30, 35, 40$. The added coordinates carry only isotropic noise, so the signal the network has to learn is the same at every width; what changes is how quickly it fits its training set. Holding the width fixed matters: scaling the hidden width with d_{latent} turns the sweep into a capacity sweep and no longer isolates the latent dimension. A separate width-scaled control is reported in Figure 10b. The full memorization trajectories, which rise later and lower as d_{latent} grows, are in Figure 10.

We also evaluate how the learned score departs from the population score. Figure 1(b) reports the per-coordinate error at $t = 0.1$, recovered from saved held-out denoising losses using the corrected analytic GMM score; the plotted values are estimates rather than direct checkpoint evaluations (Appendix C). Every width reaches its minimum by the first evaluation at 50k steps and then rises



(a) Final memorized fraction.



(b) Population score error per coordinate.

Figure 1: Synthetic memorization and score error for a fixed-width MLP, $d_{\text{intrinsic}} = 5$, five seeds, 5M steps. (a) Final memorized fraction falls from 30% at $d_{\text{latent}} = 5$ to 0.1% at $d_{\text{latent}} = 40$. (b) Population score error at $t = 0.1$ per coordinate, recovered from held-out denoising losses using the analytic GMM score. Every width reaches its minimum by 50k steps and then rises as the network fits its training set; the rise starts later at larger d_{latent}, and by 5M steps all widths have reached a similar level. The eventual score error is comparable across these widths.

as the network fits its training set. The rise starts later at larger d_{latent}: the error crosses 2.0 per coordinate at 0.35M steps for $d_{\text{latent}} = 5$ and at 1.05M for $d_{\text{latent}} = 40$, the same ordering as the memorization onset in panel (a). At 5M steps the error is 2.59 per coordinate at $d_{\text{latent}} = 5$ and 2.26–2.35 for $d_{\text{latent}} = 8$–40: final errors are comparable while the timing differs. Extending the same fixed-width sweep to $d_{\text{latent}} = 240$ makes the delay monotone over the full range, and it exceeds 25× across the sweep (Figure 11).

A nearest-neighbour ratio test can lose sensitivity as the dimension grows, so the decline in panel (a) could in principle reflect the test rather than the model. Table 1 rules out declining test sensitivity as the sole explanation. A sampler that reproduced a training point's signal coordinates with terminal sampler noise while redrawing its null coordinates would rarely be detected by the full-space test for $d_{\text{latent}} \geq 10$: such signal copies are detected at 57% at $d_{\text{latent}} = 5$ but at 0.8% and 0.0% beyond. Repeating the test after projecting generated and training points onto the true $d_{\text{intrinsic}}$-dimensional signal subspace, where signal copies are detected at every width, gives 29.9%, 16.8%, 5.8% and 1.2% for the model samples at $d_{\text{latent}} = 5, 10, 20, 40$, against a 0.4% chance level from fresh population draws. The decline remains 25× in the signal subspace. The full-space test reports a larger reduction, particularly at $d_{\text{latent}} = 40$, where it detects 0.1% of model samples compared with 1.2% after projection. The full table, including the exact-copy control, is Table 2.

Table 1: Dimension-controlled memorization test in a separate replication (5M steps, five seeds, 5000 samples each). The ratio test of Appendix A is applied in the full latent space and after projecting generated and training points onto the $d_{\text{intrinsic}} = 5$ signal coordinates. Signal copies add terminal sampler noise to a training point's signal coordinates and redraw its null coordinates. α is the alignment of the learned score with the empirical score of the training set (signal coordinates, $t = 0.1$). Full version: Table 2.

	$d_{\text{latent}} = 5$	10	20	40
model samples, full-space test	29.9%	16.9%	3.6%	0.1%
model samples, signal-projected test	29.9%	16.8%	5.8%	1.2%
fresh draws, signal-projected test	0.4%	0.4%	0.4%	0.4%
signal copies, full-space test	57%	0.8%	0.0%	0.0%
signal copies, signal-projected test	57%	57%	58%	58%
α at 5M steps	0.75	0.54	0.39	0.31

Empirical-score alignment provides a threshold-free second readout. Let $s^\star$ be the population score, s_{emp} the exact score of the empirical distribution of the n training points, and s_θ the learned score,

all in the signal coordinates at $t = 0.1$. The alignment coefficient

$$\alpha = \frac{\langle s_\theta - s^\star,\ s_{\text{emp}} - s^\star \rangle}{\|s_{\text{emp}} - s^\star\|^2} \tag{1}$$

is 0 for the population score and 1 for an exact fit of the training set. It reaches 0.75, 0.54, 0.39 and 0.31 at 5M steps for $d_{\text{latent}} = 5, 10, 20, 40$, and it passes 0.3 more than $30\times$ later at $d_{\text{latent}} = 40$ than at $d_{\text{latent}} = 5$. This is consistent with comparable final score errors in Figure 1(b) while memorization differs $25\times$: the per-coordinate gap $\|s_{\text{emp}} - s^\star\|^2$ grows $3.6\times$ from $d_{\text{latent}} = 5$ to 40, so a smaller α can still incur substantial population error. The projected sample test and empirical-score alignment both support a delay in fitting the training set as latent dimension increases.

Feature eigenvalues and learning times. To investigate a possible mechanism, we use the random-feature network (RFNN) of Bonnaire et al. (2025), whose readout dynamics are exactly solvable: freezing the first layer makes gradient flow diagonal in the eigenbasis of the feature-correlation matrix $\boldsymbol{U}$, so each eigenvalue sets the learning timescale of one feature direction. The RFNN is

$$s_A(x_t) = \tfrac{1}{\sqrt{p}}\, A\, \tanh\!\big(\tfrac{1}{\sqrt{d_{\text{latent}}}}\, W x_t\big), \tag{2}$$

with $W \in \mathbb{R}^{p \times d_{\text{latent}}}$ drawn i.i.d. from $\mathcal{N}(0,1)$ and frozen, and a readout $A \in \mathbb{R}^{d_{\text{latent}} \times p}$ trained by score matching at a single time $t = 0.01$. With $\phi(x_t) = p^{-1/2} \tanh(d_{\text{latent}}^{-1/2} W x_t)$, the empirical feature correlation matrix is

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

where x_t^μ is the diffused training point μ and ξ the diffusion noise. Full-batch gradient flow on A is linear, and in the eigenbasis of $\boldsymbol{U}$ each mode decays independently,

$$a_i(T) - a_i^\star = \big(a_i(0) - a_i^\star\big)\, e^{-\lambda_i T}, \tag{4}$$

We estimate $\boldsymbol{U}$ by averaging K noise draws, denoted $\boldsymbol{U}^{(K)}$; only $K = 1$ guarantees rank at most n (Appendix G.1). Thus eigenvalue λ_i is the learning rate of feature direction i and all modes are absorbed in parallel: no mode waits for another. For $d_{\text{intrinsic}} < d_{\text{latent}} < n < p$, partitioning the sorted spectrum of $\boldsymbol{U}^{(K)}$ at indices $d_{\text{intrinsic}}, d_{\text{latent}}, n$ gives four index groups, interpreted conditionally as: $d_{\text{intrinsic}}$ signal modes, $d_{\text{latent}} - d_{\text{intrinsic}}$ noise-dimension modes, $n - d_{\text{latent}}$ sample modes, and $p - n$ tail modes. Figure 2 shows representative measured spectra: the signal group is separated, and the top edge of the sample group changes with d_{latent}. Width and null-energy controls for these spectra are in Appendix B. Section 3 makes this quantitative.

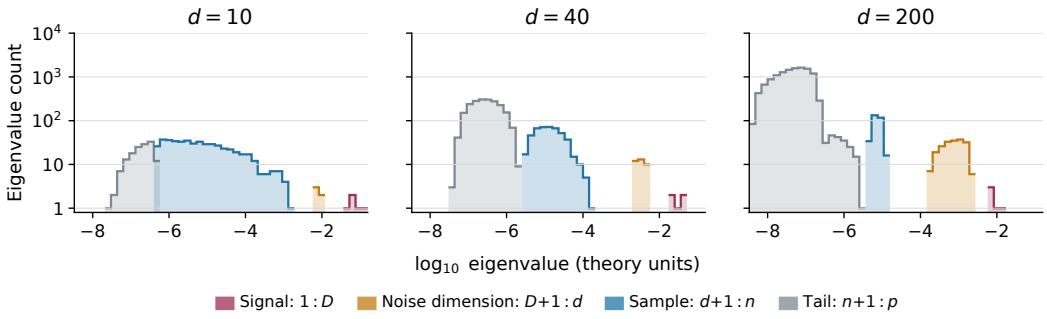


Figure 2: Measured RFNN spectra at three latent dimensions ($D = d_{\text{intrinsic}} = 5$, $d = d_{\text{latent}}$, $n = 500$, $p = 64 d_{\text{latent}}$, $t = 0.01$, $\sigma_\perp = 0.5$, seed 42). Saved eigenvalues are divided by p to match (3). Colors mark sorted index groups, not established eigenvector alignment. All eigenvalues from the $K = 50$ noise-draw averages are shown; the tail is positive and need not be separated from the sample group by a clear gap.

3 SPECTRAL EXPLANATION OF THE SYNTHETIC RESULTS

The frozen-feature model lets us ask why widening might delay memorization. The idea has two steps: eigenvalues determine how fast feature directions are learned, and widening changes those eigenvalues by changing the input variance seen by the nonlinearity. The first step follows exactly from gradient flow; the second uses measured spectra and a conditional approximation. Connecting these directions to generated-sample memorization requires additional alignment assumptions. Full derivations and the scope of these assumptions are in Appendix G.

3.1 FROM EIGENVALUES TO LEARNING TIMES

By (4), each active readout mode approaches its target independently as $1 - e^{-\lambda_i T}$: a smaller eigenvalue means slower learning (Lemma G.1). For the signal group $\mathcal{S}$ and index-defined sample group $\mathcal{M}$ introduced in Section 2, define edge-absorption times at a common relative parameter tolerance ε:

$$\tau_{\text{gen}} := \log(1/\varepsilon)/\lambda^{\mathcal{S}}_{\min}, \tag{5}$$

$$\tau^{\text{on}}_{\text{mem}} := \log(1/\varepsilon)/\lambda^{\mathcal{M}}_{\max}. \tag{6}$$

These track the slowest signal mode and the fastest sample-group mode. They are spectral proxies, not observed sample-generation times. In particular, the sample group can contain population-aligned directions; interpreting its edge as memorization onset requires the alignment and spike/bulk conditions in Proposition G.2 and Definition G.3. All active modes learn in parallel, so adding noise-dimension modes cannot itself create a queue that delays sample modes.

The measured spectra in Figure 2 motivate the four index groups, but their interpretation is conditional on separation and alignment. Noise averaging also matters: the released spectra average $K = 50$ draws and can be full rank, unlike a single-draw feature matrix. The tail and sample group need not be separated. Appendices G.1 and G.3 give the rank bounds, approximate group scales and distinctions from prior asymptotic results.

3.2 HOW WIDENING CHANGES THE INPUT SCALE

Widening spreads a fixed amount of signal energy over more input coordinates. With center norm $s = 3$ and $\Delta_t = 1 - e^{-2t}$, the expected input second moment per coordinate is

$$q(d_{\text{latent}}) = \Delta_t + e^{-2t}\left[\sigma^2_\perp + \frac{s^2 + d_{\text{intrinsic}}(\sigma^2_{\text{sig}} - \sigma^2_\perp)}{d_{\text{latent}}}\right]. \tag{7}$$

The null-noise floor stays fixed, while the excess signal contribution decreases as $1/d_{\text{latent}}$. Under the RFNN normalization, widening therefore lowers q while preserving the signal distribution, and moves $\tanh$ toward its linear regime. This is an expectation over the data construction; realized training sets fluctuate (Lemma G.4). Write $a_1(q)$ for its linear response and $a_*(q)^2$ for the residual nonlinear variance; their definitions and exact variance identity are in Lemma G.4. For $z \sim \mathcal{N}(0,1)$, these coefficients separate the linear and nonlinear parts of the activation:

$$a_1(q) = \mathbb{E}[\operatorname{sech}^2(\sqrt{q}z)], \qquad a_*(q)^2 = \mathbb{E}[\tanh^2(\sqrt{q}z)] - q\,a_1(q)^2. \tag{8}$$

When q is small, $a_1(q) \to 1$ while $a_*(q)^2 = \frac{2}{3}q^3 + O(q^4)$. The nonlinear residual therefore shrinks faster than the linear response as the input scale falls. This explains the direction of the proposed effect: sample-group edges associated with the residual can slow relative to signal edges. The released numerical comparisons use quadrature at the measured q, rather than treating the small-q expansion as accurate throughout the sweep. The conditional edge approximation gives

$$R(d_{\text{latent}}) := \frac{\tau^{\text{on}}_{\text{mem}}}{\tau_{\text{gen}}} = \frac{\lambda^{\mathcal{S}}_{\min}}{\lambda^{\mathcal{M}}_{\max}} \simeq \frac{n}{d_{\text{latent}}} \cdot \frac{a_1(q)^2}{a_*(q)^2} \cdot \frac{e^{-2t}\ell_{\min} + \Delta_t}{\kappa_{\text{top}}}, \tag{9}$$

where $\Delta_t = 1 - e^{-2t}$, $\ell_{\min}$ is the smallest clean signal-subspace eigenvalue, and κ_{top} is a measured sample-edge coefficient (Proposition G.7). The first equality is exact; the approximation is a decomposition using measured information, not an independent prediction of the sample edge. It shows why

lower nonlinear variance can increase the separation of learning times, but the normalization, signal geometry and residual edge also change. Growth of a_1^2/a_*^2 alone does not guarantee increasing R.

On the released spectra, q falls from 2.76 to 0.33 and a_*^2 falls about $17\times$. Signal e-folds grow $11\times$ over $d_{\text{latent}} = 5 \to 200$, sample-edge e-folds grow $221\times$, and R rises from 21 to 426. Tables 4 and 7 report the calibration and the $d_{\text{latent}} = 5$ case where $p < n$.

The calibration also separates what is predicted from what is measured. Using the realized signal geometry gives conditional signal-edge estimates within approximately 5% of the measured values at the tabulated widths. The sample-edge coefficient is extracted from those same spectra: it falls about $13\times$, in addition to the $17\times$ decline in nonlinear residual variance. Thus the agreement tests the signal scale, while the full ratio remains a measured decomposition (Table 4).

To test the contribution of the changing operating point, we scale the null variance to hold the clean input variance per coordinate at $q_{\text{in}} = 0.57$, approximately its baseline value at $d_{\text{latent}} = 40$ (Remark G.8). Holding this variance fixed largely removes the endpoint growth of R: $1.09\times$ rather than $3.17\times$ over $d_{\text{latent}} = 40 \to 240$, even though the noise-dimension count grows $6.7\times$ (Figure 3). The fixed-q curve is non-monotone, and the intervention also changes the null covariance. It therefore supports an operating-point contribution rather than isolating a complete causal mechanism. An isotropic-noise control likewise yields a non-monotone ratio (Remark G.9). The edge changes locate the difference: the signal edge falls by 4.3–$4.4\times$ in both sweeps, whereas the sample edge falls by $13.6\times$ in the baseline and $4.8\times$ at fixed q. This supports a relative slowing of the residual-associated directions, beyond the shared normalization effect.

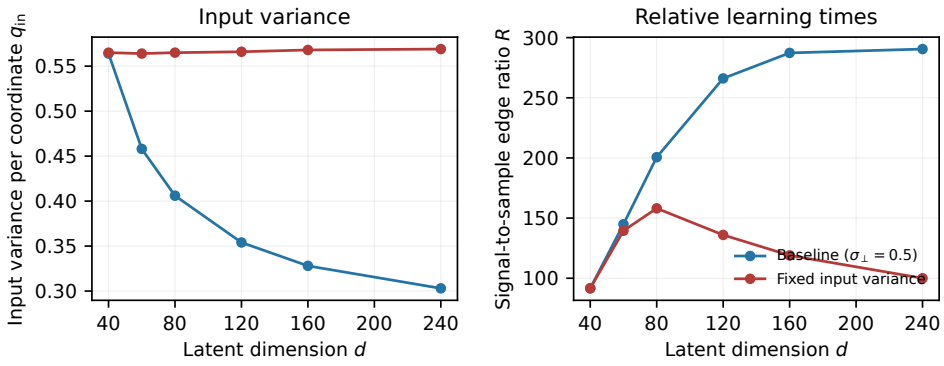


Figure 3: Fixed-q control ($n = 500$, $p = 4096$, three seeds). Left: input variance per coordinate in the baseline and fixed-q sweeps. Right: measured signal-to-sample edge ratio. Holding the operating point fixed largely removes its endpoint growth ($1.09\times$ against $3.17\times$ over $d_{\text{latent}} = 40 \to 240$) while the noise-dimension count grows $6.7\times$. Protocol and numerical decomposition: Remark G.8.

A linear comparison isolates the role of the nonlinear response. For a population score model $s_B(x) = Bx$, second moment $M = \mathbb{E}[x_t x_t^\top]$ and initialization $B(0) = 0$, gradient flow on the half-scaled squared score loss has the exact solution

$$\dot{B} = -(BM + I), \qquad B(T) = -M^{-1}(I - e^{-MT}). \tag{10}$$

Each independent Gaussian null direction relaxes at rate $\beta_t^2 = \Delta_t + e^{-2t}\sigma_\perp^2$, regardless of how many such directions are added. Widening increases their total contribution to the error but does not change their individual rate. This comparison explains why a dimension count alone cannot supply the proposed delay: changes in the nonlinear feature spectrum are needed (Appendix G.5).

The analysis thus connects input scale to learning times inside the frozen-feature model. It does not derive near-duplicate generation from spectral edges, and trainable first layers and adaptive optimizers remain outside the model. Section 4 evaluates its usefulness for timing estimates in trained networks.

4 REAL-DATA EXPERIMENTS

We study CelebA and CIFAR-10 images using a separately trained convolutional VAE at each latent width. Its encoder ends in linear heads, producing a vector of dimension d_{latent} without a spatial grid. We freeze the VAE and train an MLP score model at fixed hidden width on the encoder means of a diverse subset of 1k images, for 5M steps and five seeds. Architecture and training details are in Appendix A. Memorization uses the same nearest-neighbor ratio as the synthetic test, measured in pixel space after decoding; FID uses the same 1k reference set.

For each width, we decode 10k generated samples at evaluations every 100k training steps. Hitting times are the first saved evaluations crossing a specified memorized fraction. Runs that do not cross by 5M steps are censored; reported crossing-time means use crossing seeds, with censoring counts provided alongside them in Table 3.

A narrow VAE can erase sample identity before the diffusion model can copy it. We therefore compare widths above the smallest value at which 95% of VAE reconstructions pass the memorization test: $d_{\text{latent}} \approx 70$ on CelebA (Appendix D). This is an operational reconstruction threshold, not an estimate of manifold dimension. Even above it, the decoder changes with width, so this experiment isolates added null coordinates less directly than the synthetic construction.

Above this threshold, memorization starts later and reaches lower final levels as d_{latent} grows (Figure 4). The mean time to reach 5% memorization rises from 0.32M steps at $d_{\text{latent}} = 80$ to 1.16M at $d_{\text{latent}} = 200$, a 3.6× slowdown. The final fraction peaks near $d_{\text{latent}} = 70$–80 and falls by roughly half by $d_{\text{latent}} = 200$. Hitting times, uncertainty and censoring are reported in Table 3. CIFAR-10 gives supporting, less monotone evidence with a later reconstruction threshold ($d_{\text{latent}} \approx 140$) and a 1.4× slowdown: the mean 5% hitting time rises from 1.04M at $d_{\text{latent}} = 160$ to 1.50M at $d_{\text{latent}} = 240$; the full sweeps are in Appendix D and Figure 13.

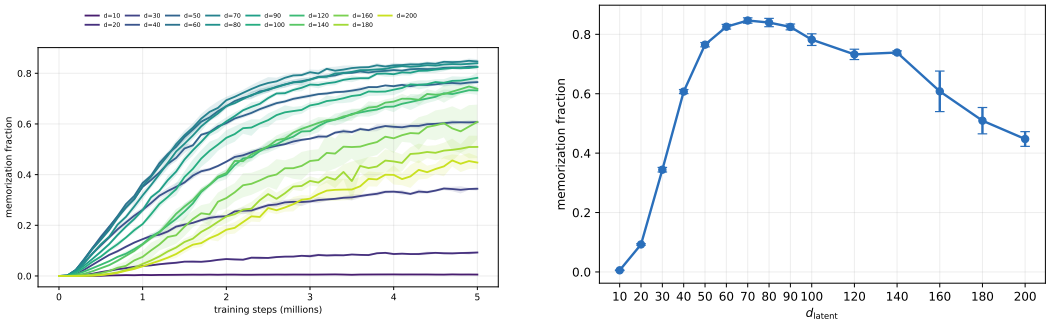


Figure 4: CelebA memorization, five seeds with 95% confidence bands. Left: memorized fraction over training by latent width. Right: final fraction after 5M steps against width. Above the identity threshold ($d_{\text{latent}} \approx 70$) memorization starts later and ends lower as width grows; below it the VAE erases sample identity. Full timing is in Table 3.

FID improves as the bottleneck is relieved, is best at intermediate widths, and worsens again at the widest latents: on CelebA from 38.9 at $d_{\text{latent}} = 140$ to 47–50 at $d_{\text{latent}} = 160$–200 (Figure 5). Quality and copying are separate outcomes: lower memorization at large width comes with a loss in this measure of image quality.

Spectral timing diagnostics. The frozen VAE latent spectrum supplies the eigenvalues of a diffused second moment M_t, from which the released predictor constructs an absorption clock. One rate constant and up to six monotone threshold levels are calibrated on observed diffusion-model hitting times to obtain per-width curves and timing estimates (Appendix F). Figure 6 overlays these estimates on the measured trajectories.

Writing $d = d_{\text{latent}}$, let $Z_d \in \mathbb{R}^{n \times d}$ contain the frozen encoder means for the training subset. At diffusion time $t = 0.1$, define

$$M_t(d) = e^{-2t} Z_d^\top Z_d / n + \Delta_t I_d, \qquad P_d(s) = \frac{1}{d} \sum_{i=1}^{d} \left(1 - e^{-\kappa \lambda_i(d) s}\right)^2, \tag{11}$$

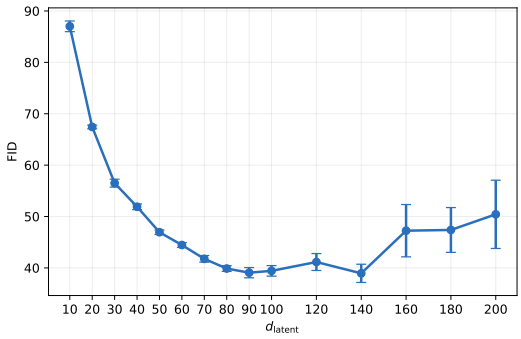


Figure 5: Final CelebA FID after 5M steps against latent width, five-seed mean with 95% confidence intervals: best at intermediate widths, worse at the widest. This FID is a subset-relative quality statistic (Appendix A).

where s denotes training steps and $\lambda_i(d)$ are the eigenvalues of $M_t(d)$. The pressure P_d summarizes absorption in the frozen data spectrum. These are data-side eigenvalues, rather than the RFNN feature eigenvalues in Section 2; using them for trained-network timing is an empirical modeling step.

A shared rate κ and ordered pressure levels θ_f give the predicted crossing time $\widehat{\tau}_f(d) = \inf\{s : P_d(s) \geq \theta_f\}$, for fractions $f \in \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75\}$. Calibration shares these quantities across widths within each dataset, with no per-width fitting. Available observed crossings determine the levels; fully censored levels are filled from neighbors and supply no validation target. This construction tests whether a common spectral description summarizes the observed timing variation, while requiring calibration from diffusion training runs.

The representative panels distinguish a reconstruction-limited regime from the above-threshold comparisons. At narrow widths, low detected copying can reflect lost image identity. Once identity is preserved, the later and lower trajectories at larger widths provide evidence for delayed memorization. The curves do not imply that every added VAE coordinate is an independent null direction, as in the controlled synthetic experiment.

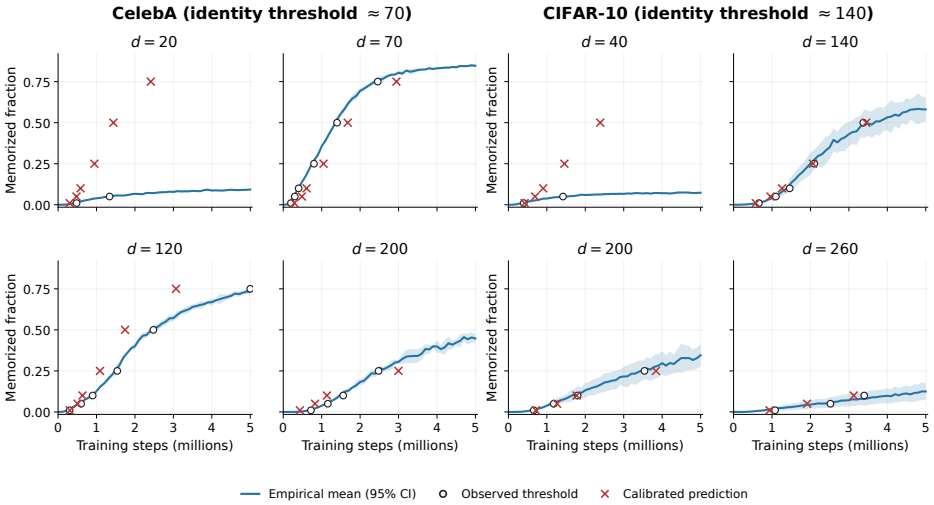


Figure 6: Spectral memorization timing on CelebA and CIFAR-10 at four representative widths per dataset. Blue curves show five-seed means with 95% confidence bands; circles and crosses mark observed and calibrated predicted hitting times. Below the VAE identity thresholds ($d_{\text{latent}} \approx 70$ and 140), reconstruction limits detectable copying. At larger widths, memorization is delayed and reduced. Predictions are fitted in-sample across widths; the full sweeps and calibration details are in Appendix F and Figures 49 and 50.

A separate one-constant spectral clock is evaluated by fitting on all widths but one. It uses $\tau_f(d_{\text{latent}}) = Cn/g(q_d, 1 - \Delta_t/q_d)$, with $q_d = \operatorname{tr} M_t/d_{\text{latent}}$, and reaches a log error of 0.41 on CelebA ($d_{\text{latent}} \geq 50$), compared with 0.13 for a dimension-only clock and 0.34 for the released predictor's in-sample fit (Table 5). Errors are mean absolute differences in log hitting time for the 5% threshold. The clock constant is fitted on all but the evaluated width, so each error tests transfer across widths rather than across independent datasets. The dimension-only baseline therefore performs better here. On CIFAR-10 ($d_{\text{latent}} \geq 60$), the corresponding errors are 0.15 for the spectral clock and 0.20 for the dimension-only baseline; the released in-sample fit is also 0.15. The comparison therefore reverses, but the fitted constant differs by $2\times$ between datasets; both evaluations include widths below the identity threshold. The overlays describe an in-sample calibration, and the held-out comparison tests its predictive value: neither establishes a mechanism for trained-network memorization. A separate variance-based time-rescaling diagnostic accounts for 0.55–0.82 of the CelebA log-slowdown over $160 \to 200$, but only 0.28 of the final-fraction log-decline (Table 6).

5 Conclusion and limitations

Widening the latent space at fixed hidden width delays memorization in the tested MLPs. In the synthetic sweep, the memorized fraction falls from 30% to 0.1% over $d_{\text{latent}} = 5 \to 40$, and the decline survives the signal-subspace test (Table 2). Above the VAE identity threshold, the time to reach 5% memorization grows $3.6\times$ on CelebA and $1.4\times$ on CIFAR-10. Spectral clocks calibrated on observed training runs provide timing estimates, but their advantage over a dimension-only baseline is dataset dependent (Section 4).

The frozen-feature analysis establishes exact learning dynamics for individual feature directions. Its main implication is that counting added dimensions alone misses how their variance changes the nonlinearity: holding the variance per coordinate fixed largely removes the measured spectral-ratio growth ($1.09\times$ against $3.17\times$ over $d_{\text{latent}} = 40 \to 240$, Figure 3). This intervention also changes the null covariance. Relating the measured eigenvalue groups to memorization still requires spectral alignment and a link from feature learning to generated copies; the analysis does not prove delayed near-duplicate generation in trained networks.

Limitations. In trained synthetic MLPs, the first-layer pre-activation variance is similar across widths at the first saved checkpoint ($q_{\text{M}}(40)/q_{\text{M}}(5) = 0.96$ at 50k steps) and separates as memorization proceeds (Appendix C). This does not establish that an earlier variance difference causes the delay. In a CelebA pilot with a spatial $4 \times 4 \times d_{\text{latent}}$ VAE and a DiT-style score network, memorization rises from about 22% at $d_{\text{latent}} = 5$ to 48% by $d_{\text{latent}} = 8$–10, with no clear decline at larger widths (Appendix E and Figures 14 and 15). The demonstrated effect is therefore limited to the tested vector-latent MLP setting. Absolute percentages depend on the nearest-neighbor criterion, and all diffusion-model training sets have $n \leq 1000$, leaving larger-scale transfer untested. Hidden width is fixed, although input/output parameter counts change with latent dimension; a separate capacity sweep is in Figure 10b.

Latent width deserves consideration when auditing memorization alongside reconstruction and generation quality. The observed delay motivates interventions on trained features and validation on new settings before spectral timing estimates are used to choose widths or training horizons. Increasing latent dimension alone is not an established privacy guarantee.

AI USE STATEMENT

We used AI tools to help edit the manuscript and identify technical terminology or phrasing that could be unclear to readers, to find potentially related work, and to assist with writing code used to implement and run our models. The authors reviewed the AI-assisted material and take responsibility for the paper's code, analyses, claims, and citations.

REFERENCES

Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1 – 37, 2021. doi: 10.1214/21-EJP699.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024. doi: 10.1038/s41467-024-54281-3.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. Why diffusion models dont 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 (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 141266–141286. Curran Associates, Inc., 2025.

Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

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)*, pp. 5253–5270, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3.

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 (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 21–27 Jul 2024.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7710–7721, 2020. arXiv:2005.11879.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1405–1413. PMLR, 2026. arXiv:2502.00336.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference (MSML)*, volume 145 of *Proceedings of Machine Learning Research*, pp. 426–471. PMLR, 2022. arXiv:2006.14709.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Reconstruction and membership inference attacks against generative models. *arXiv preprint arXiv:1906.03006*, 2019.

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

Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023. arXiv:2009.07669 (2020).

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 (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 38307–38354. Curran Associates, Inc., 2024. doi: 10.52202/079017-1211.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022.

Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3063–3071. PMLR, 2018. arXiv:1805.11916.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018. arXiv:1702.05419.

Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. *arXiv preprint arXiv:2004.05675*, 2020.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022. arXiv:1908.05355.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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)*, pp. 10684–10695, June 2022.

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*, pp. 6048–6058, 2023.

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.

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 (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46412–46440. PMLR, 21–27 Jul 2024.

Mingyang Yi, Jiacheng Sun, and Zhenguo Li. On the generalization of diffusion model. *arXiv preprint arXiv:2305.14712*, 2023.

A EXPERIMENTAL PROTOCOL

This appendix gives the training and evaluation protocol, the controls needed to interpret the main results, and the architecture and prediction failures that limit their scope. Further implementation details are in Appendix I; Appendix G distinguishes exact gradient-flow identities from conditional spectral approximations. A guide to the extended diagnostics is in Appendix H.

A.1 DATA, SCORE MODELS, AND SAMPLING

The main synthetic sweep has $n = 500$, $d_{\text{intrinsic}} = 5$, ten equiprobable Gaussian mixture components, center norm 3, within-component signal standard deviation $\sigma_{\text{sig}} = 1$, and independent null-coordinate standard deviation $\sigma_{\perp} = 0.5$. A random orthogonal matrix rotates the full vector before training. The main comparison uses $d_{\text{latent}} \in \{5, 8, 10, 12, 15, 20, 25, 30, 35, 40\}$; the extension adds $\{60, 80, 100, 120, 140, 160, 180, 200, 220, 240\}$. Changing d_{latent} preserves the signal distribution while adding independent null coordinates. It also changes the input dimension and total null energy; the controls below separate these effects from changes in model width and the feature operating point.

All vector-latent experiments use the Ornstein–Uhlenbeck perturbation

$$x_t = e^{-t}x_0 + \sqrt{\Delta_t}\,\xi, \qquad \Delta_t = 1 - e^{-2t}, \qquad \xi \sim \mathcal{N}(0, I).$$

The trainable MLPs minimize denoising score matching, $\mathbb{E}\|\sqrt{\Delta_t}s_\theta(x_t, t) + \xi\|^2$, with the squared residual summed over coordinates and averaged over the minibatch. Times are drawn uniformly from $[0.01, 3]$. The analytic population score is an evaluation target; it is not used to train the synthetic MLP. The frozen-feature RFNN instead trains at $t = 0.01$ and its saved-sweep implementation averages the loss over coordinates as well as samples. This normalization matters when converting its eigenvalue clocks to optimization steps (Lemma G.1).

The score model has three GELU hidden layers of width 256 and a linear output of dimension d_{latent}. The input concatenates x_t with 32 sine and 32 cosine time features, using frequencies logarithmically spaced from 1 to 1000. Adam uses learning rate 10^{-4}, batch size 256, and a 5M-step horizon for seeds 42–46. Hidden width is fixed across the sweep. The alternative width rule $8d_{\text{latent}}$ is a capacity control, not the main dimension intervention.

The vector experiments initialize $x \sim \mathcal{N}(0, I)$ at $t = 3$ and use 500 Euler–Maruyama steps to $t = 0.01$. With $h = (3 - 0.01)/500$, each reverse-time update is

$$x \leftarrow x + \{x + 2s_\theta(x, t)\}h + \sqrt{2h}\,\zeta, \qquad \zeta \sim \mathcal{N}(0, I).$$

The sampler stops at positive diffusion time and therefore retains terminal noise; a memorized sample need not be an exact numerical copy. The dimension-controlled tests in Table 2 use 5000 generated samples per seed. The image evaluations use 10k decoded samples per checkpoint against the 1k diffusion training subset.

Each latent width has its own convolutional VAE trained on the full available training split, before the diffusion subset is selected. CelebA uses the local CelebA-HQ split with 21k training and 6k validation images, resized to 32×32 grayscale; CIFAR-10 uses the standard 50k/10k split at 32×32 RGB. The final VAE checkpoints use a five-stage ResNet-style encoder/decoder with channel widths $(32, 64, 128, 256, 512)$, stride-two downsampling, GroupNorm/SiLU residual blocks, transpose-convolution upsampling, and a Tanh output. Their capacity objective is $\mathcal{L}_{\text{recon}} + \gamma|\text{KL} - C(t)|$, with an early warm-up of $C(t)$; the checkpoint is selected by validation objective. The same farthest-first diverse subset of 1000 training images is used across latent widths for each dataset. Diffusion trains on encoder means, not samples from the VAE posterior.

The vector-latent score model is a depth-five MLP with hidden width 1024 and sinusoidal time features. It uses SGD with learning rate 10^{-3}, momentum 0.8, batch size 512, and 5M steps over five seeds. Hidden width is fixed, although input/output parameter counts grow with d_{latent}. Training loss is logged every 10k steps; pixel-space memorization and FID are evaluated every 100k steps, both against the same 1k subset. This FID is a subset-relative quality statistic, not a full-dataset benchmark FID. Nonfinite runs are optimization failures. The final CIFAR-10 aggregate uses stable, unclipped replacement runs with the same SGD settings. Clipped diagnostic runs, which suppressed memorization, are not mixed into it.

A.2 METRICS AND THEIR LIMITATIONS

A generated point is counted as memorized when its distances to the nearest and second-nearest training points satisfy $\|\hat{x} - \mathrm{NN}_1\|/\|\hat{x} - \mathrm{NN}_2\| < 1/3$ (Bonnaire et al., 2025). Distances are in latent space for synthetic data and decoded pixel space for images. The reference is the diffusion training subset, not the full VAE training split. Curves report five-seed means with 95% confidence bands unless stated otherwise. Hitting times use the first saved evaluation reaching a specified fraction; they are resolution-limited and right-censored at 5M steps, not continuously observed crossing times.

For synthetic data we also project both training and generated points onto the known five-dimensional signal subspace. This removes the growing null distance from the test without making detection perfect: signal-copy controls are detected at about 57–58% at every width under the terminal-noise convention. For images we first test immediate VAE reconstructions of the same 1k subset. The smallest width at which 95% of these reconstructions pass the ratio test is approximately 70 on CelebA and 140 on CIFAR-10. We call this an *identity threshold*; it is an operational property of the VAE and metric, not an estimate with a theorem identifying the manifold's intrinsic dimension. Below it, low pixel-space memorization may be caused by the decoder erasing identity. Above it, decoder changes and reconstruction distortion can still affect the test.

B FROZEN-FEATURE CONTROLS

These experiments measure feature spectra; they do not generate samples or measure a memorized fraction. The default setting is the synthetic GMM above, $t = 0.01$, and the RFNN of Equation (2). The controlled-width and fixed-width panels in Figures 7 and 8 cover two of the three released rules: $p = 64d_{\mathrm{latent}}$, $p = d_{\mathrm{latent}} + n + 300$, and fixed $p = 1800$ with fixed total null energy. The latter uses $\sigma_\perp^2 = E_{\mathrm{null}}/(d_{\mathrm{latent}} - d_{\mathrm{intrinsic}})$ where null coordinates are present. It removes increasing feature count and increasing total null variance; it does not hold the operating point q fixed. The fixed-q intervention in Figure 3 addresses that separate mechanism directly. The complete RFNN width and noise sweeps are in Appendix J.

The feature correlation is estimated with $K = 50$ noise draws in the $p = 64d_{\mathrm{latent}}$ raw-data spectra and $K = 500$ in the $p = d_{\mathrm{latent}} + n + 300$ set. A single-draw estimate has rank at most n; an average of K estimates has rank at most $\min(p, nK)$ and can be full rank. Accordingly, purple tails in the noise-averaged plots are not exact rank-null spaces. Colored groups are assigned by sorted eigenvalue index, not fitted clusters or established population/sample alignment. At $d \geq n$, the $n - d$ group count used in the main theoretical regime no longer applies. Those extended panels are stress tests outside $p > n > d$.

The complementary source-dimension sweep fixes $d_{\mathrm{latent}} = 20$ and varies $d_{\mathrm{intrinsic}} \in \{2, 5, 8, 12, 16, 20, 25, 30, 35, 40\}$ under the same width rules. Below $d_{\mathrm{intrinsic}} = 20$ the number of observed null coordinates decreases; beyond that boundary, source variation is compressed into the fixed observed space and there is no literal negative-sized null block. Gaussian robustness runs remove mixture centers and vary $\sigma_\perp \in \{0.1, 0.3, 0.5\}$ at $\sigma_{\mathrm{sig}} = 1$, or $\sigma_{\mathrm{sig}} \in \{0.5, 1, 1.5\}$ at $\sigma_\perp = 0.3$. Figure 9 retains the null-noise sweep. The location and visibility of the groups depend on these scales; the noise-dimension/sample gap nearly closes in the separate $\sigma_\perp = 0.01$ runs. Thus four separated components are not a universal consequence of adding dimensions.

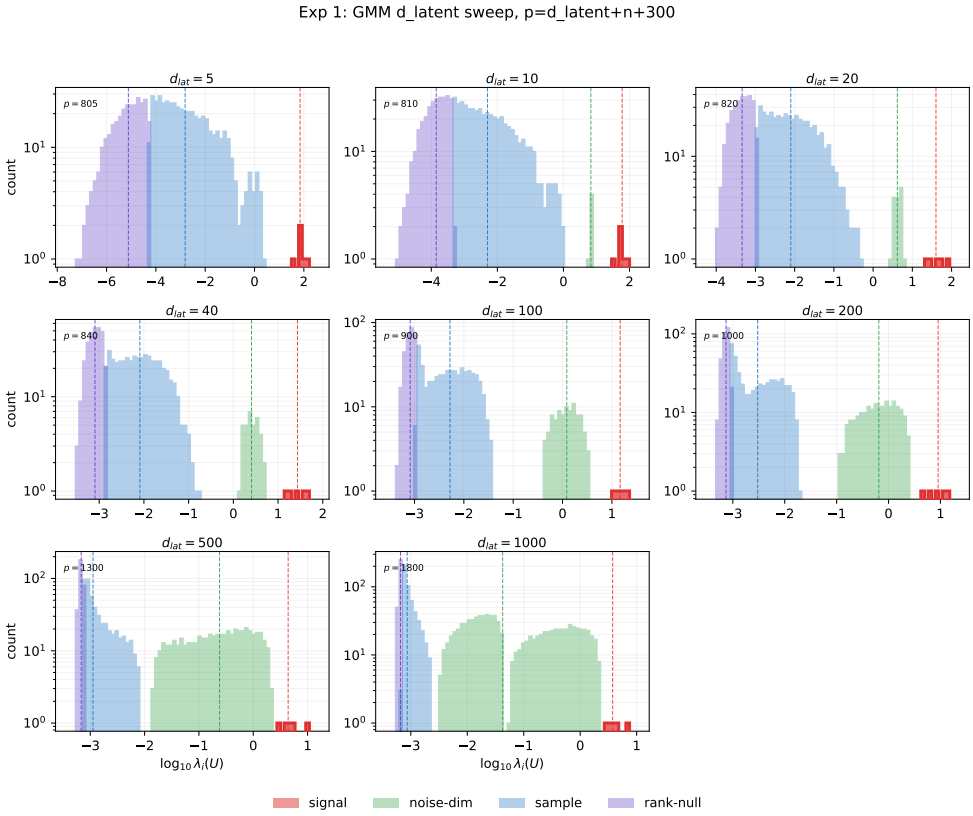


Figure 7: Representative latent-dimension spectra with $p = d_{\text{latent}} + n + 300$, $d_{\text{intrinsic}} = 5$, $n = 500$, $\sigma_{\text{sig}} = 1$, $\sigma_{\perp} = 0.5$, and center norm 3. Red, green, blue, and purple mark signal, noise-dimension, sample, and tail index groups; dashed lines show group median log-eigenvalues. Visible gaps support a spectral description in this configuration, but coloring alone does not prove alignment or separation. The extended high-d panels leave the $d < n$ regime.

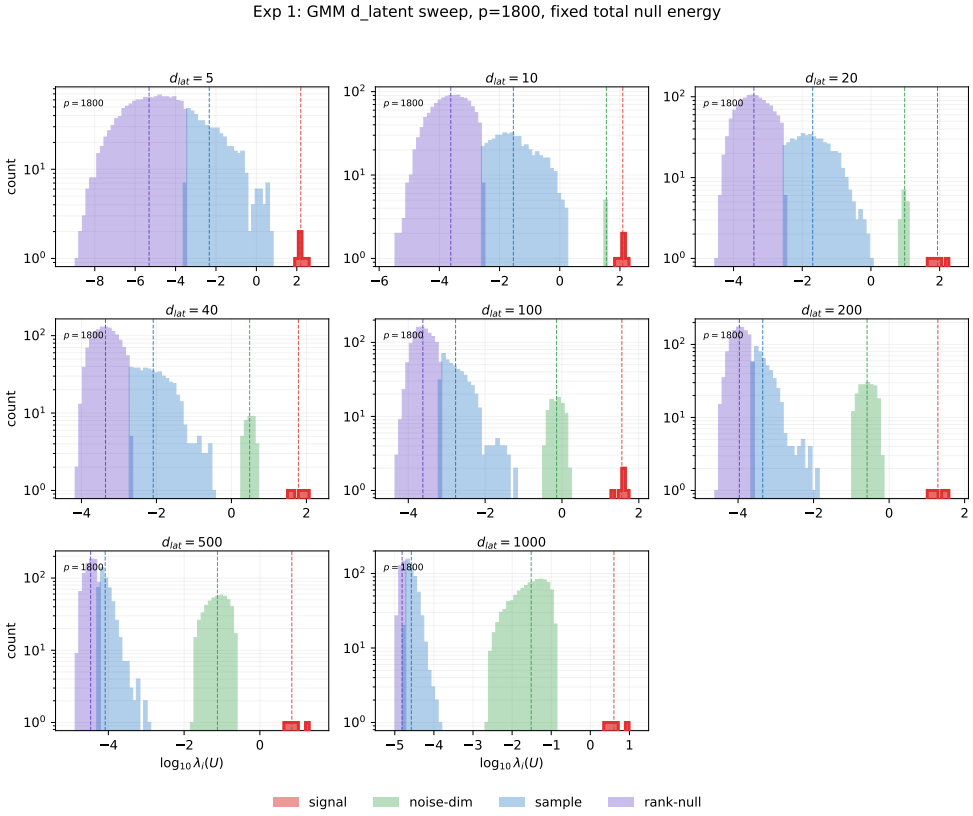


Figure 8: Fixed-width, fixed-total-null-energy control: $p = 1800$ and $\sigma_\perp^2 = E_{\text{null}}/(d_{\text{latent}} - d_{\text{intrinsic}})$. The signal distribution and feature count are fixed while null variance is distributed across more coordinates. Separation persists in parts of this sweep, but the control does not isolate a count-of-modes explanation: the per-coordinate variance entering $\tanh$ still changes. Colors use the index convention of Figure 7.

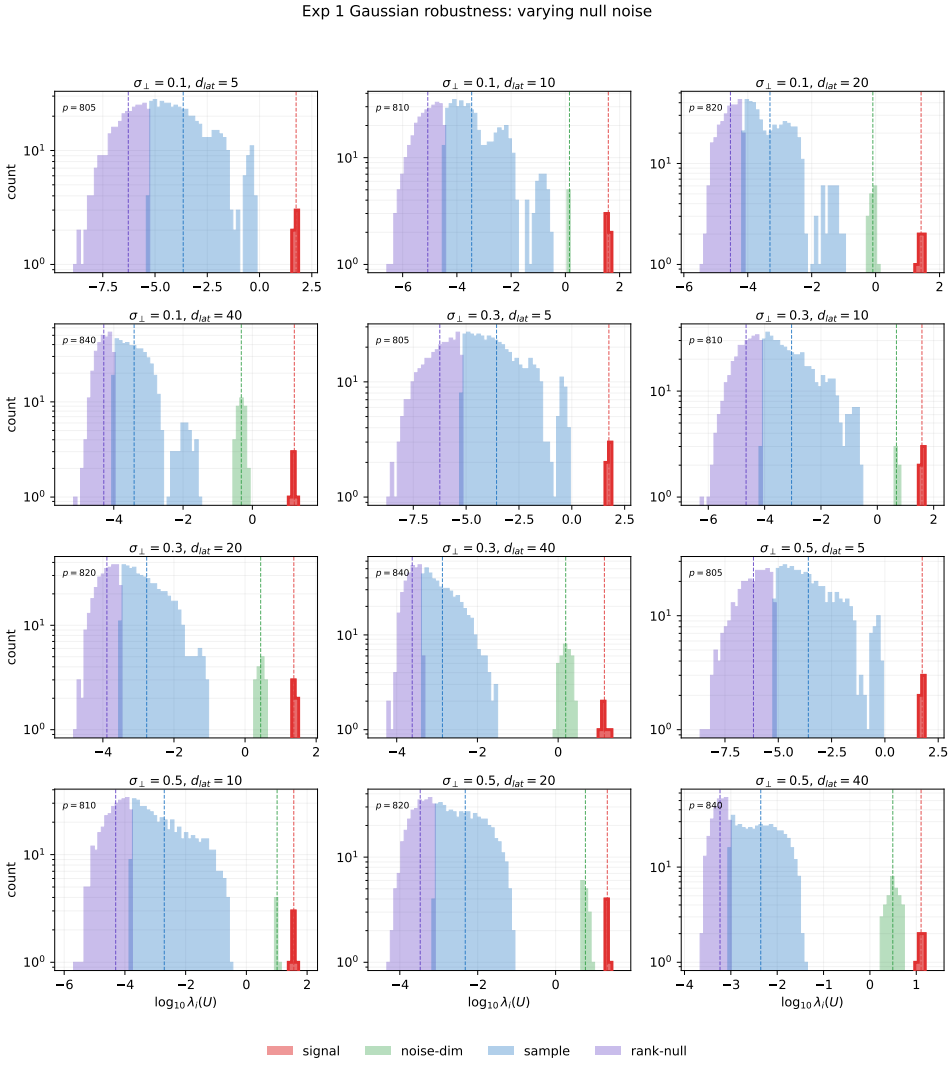
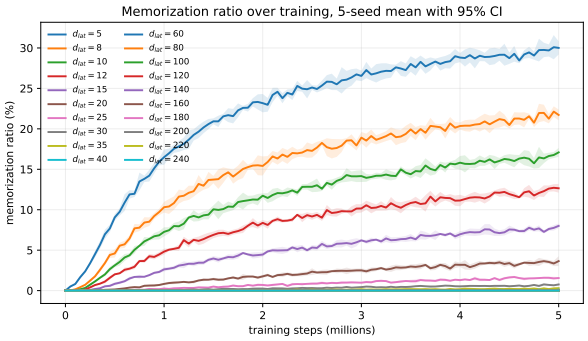


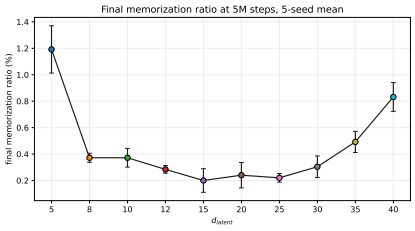
Figure 9: Robustness to removing the mixture centers. Pure anisotropic Gaussian data with $d_{\text{intrinsic}} = 5$, $\sigma_{\text{sig}} = 1$, and rows $\sigma_\perp \in \{0.1, 0.3, 0.5\}$, using $p = d_{\text{latent}} + n + 300$. Changing null variance moves the noise-dimension group and its separation from the sample group. The effect therefore depends on covariance and nonlinear feature scale, not only on group cardinalities.

C SYNTHETIC MLP CONTROLS AND SCORE DIAGNOSTICS

The full fixed-width sweep retains the delayed memorization seen in the main-text range. Figure 10 shows trajectories rather than only final fractions, and the width-scaled control shows why fixing the hidden width matters. A later rise in memorization does not establish that the trained MLP follows the frozen-feature dynamics: all RFNN modes learn in parallel, and a trainable first layer can change its features. The extended trajectories and score diagnostics are in Appendix K.



(a) Fixed hidden width 256, full dimension sweep.



(b) Hidden width $8d_{\text{latent}}$, final fraction.

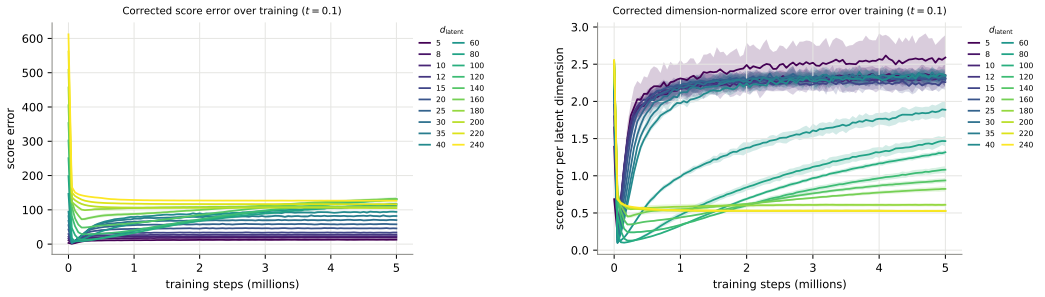
Figure 10: Synthetic MLP controls, five seeds and 5M steps. Left: the fixed-width trajectories extend through $d_{\text{latent}} = 240$. Right: scaling hidden width from 40 to 320 changes capacity; memorization stays near the 1% level and rises again for $d_{\text{latent}} \geq 30$. This second sweep cannot isolate latent dimension.

Table 2: Dimension-controlled test at 5M steps (five seeds, 5000 samples each). Exact-copy controls add the sampler's terminal noise to training points; signal-copy controls add the same terminal noise to copied signal coordinates and redraw null coordinates. The added noise has per-coordinate variance 0.0269 in these controls. α is learned-score alignment with the empirical score in signal coordinates at $t = 0.1$, with 0 corresponding to the population score and 1 to an exact empirical fit.

	$d_{\text{latent}} = 5$	10	20	40
model samples, full-space test	29.9%	16.9%	3.6%	0.1%
model samples, signal-projected test	29.9%	16.8%	5.8%	1.2%
fresh draws, signal-projected test	0.4%	0.4%	0.4%	0.4%
exact copies, full-space test	57%	69%	84%	97%
signal copies, full-space test	57%	0.8%	0.0%	0.0%
signal copies, signal-projected test	57%	57%	58%	58%
α, 0.25M steps	0.54	0.37	0.22	0.09
α, 5M steps	0.75	0.54	0.39	0.31

The plotted error is recovered from saved held-out denoising losses by $\widehat{e} = L_{\text{test}}/\Delta_t - \widehat{C}/d_{\text{latent}}$, per coordinate, where $\widehat{C}$ estimates $\mathbb{E}\|s^\star - s_{\text{cond}}\|^2$ on the same 2048 held-out noisy pairs using the corrected analytic GMM score. This corrects a rotation-frame error in the originally logged analytic diagnostic; training did not use that diagnostic. The DSM identity is exact in expectation, but a finite-sample cross term remains, so these are recovered estimates rather than direct checkpoint evaluations. At 5M steps the mean per-coordinate error is 2.59 for $d_{\text{latent}} = 5$, 2.26–2.35 for 8–40, and 0.53 for 240. Larger dimensions delay departure from the population score; within 5–40 they do not preserve a lower final error at this horizon.

We measure $q_{\text{M}} = (NH)^{-1} \sum_{i,j} h_j(x_{t,i}, t)^2$, where h is the first linear layer's output including bias and Fourier time embedding, $H = 256$, $N = 4096$ held-out population draws, and $t = 0.1$. This uncentered second moment includes learned offsets and differs from the isotropic RFNN operating point. Test draws use seed 9999 and diffusion noise seed 123, fixed across checkpoints. Averaged over seeds 42–46, at 0.05M, 0.25M, 1M, and 5M steps the values are $(0.710, 1.118, 1.621, 2.813)$



(a) Summed over coordinates.

(b) Per coordinate.

Figure 11: Recovered population score error at $t = 0.1$ over the complete fixed-width sweep, five-seed mean with 95% CI. Per-coordinate normalization removes the mechanical increase in the number of outputs. The widest latents have not returned to the small-dimension error plateau by 5M steps. The recovery and finite-sample limitation are stated above.

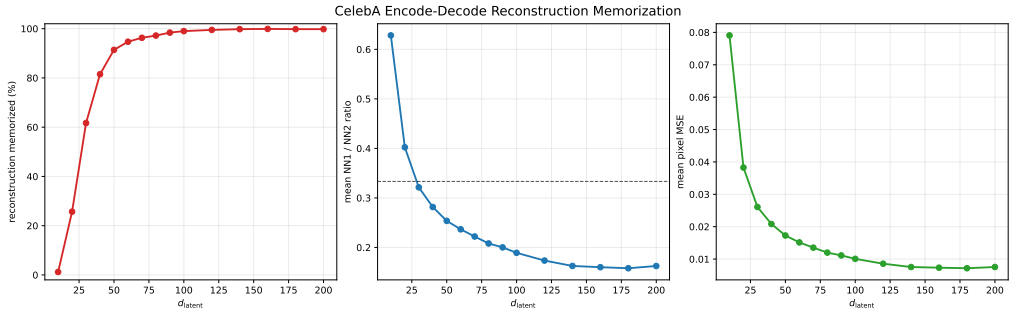
for $d_{\text{latent}} = 5$ and $(0.685, 0.780, 0.762, 0.836)$ for $d_{\text{latent}} = 40$. No initialization checkpoint is available. The first measurement is at 50k steps, when these values are similar; their later separation does not establish a width-set operating-point difference preceding memorization.

D IMAGE IDENTITY, QUALITY, AND MEMORIZATION TIMING

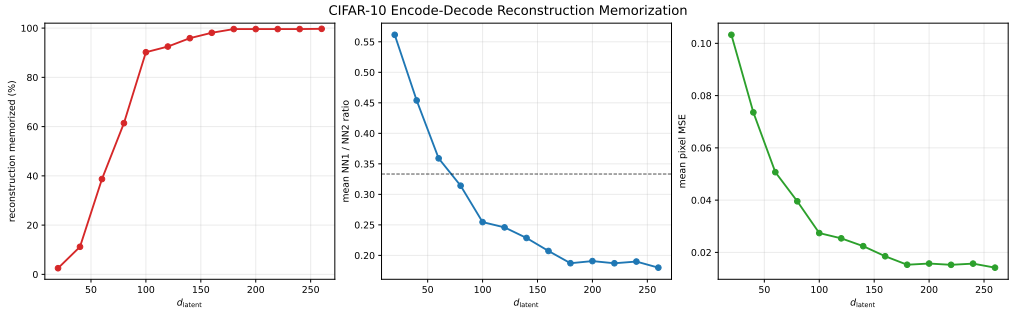
Figure 12 provides the reconstruction check behind the identity thresholds. Figure 13 shows the aggregate memorization and FID trajectories. Quality and copying are separate outcomes: on CelebA the best reported FID is 38.9 at $d_{\text{latent}} = 140$, worsening to 47–50 at 160–200 while memorization falls. The post-threshold CIFAR-10 hitting-time trend is not strictly monotone. VAE training curves and reconstruction grids are in Appendix L; the per-width diffusion trajectories are in Appendix M.

Table 3: Empirical hitting times on the real-data sweeps: first evaluation step (in millions, mean ± sd over seeds) at which the pixel-space memorized fraction exceeds the threshold; c = seeds censored at the 5M-step horizon. Source: {celeba,cifar10}_empirical_tau_summary.csv in the analysis repository. Means and standard deviations use only seeds that cross; censored seeds are shown separately. These times support Appendix G.6.

	CelebA ($n = 1000$)				CIFAR-10 ($n = 1000$)		
d_{latent}	thr. 0.01	0.05	0.10	d_{latent}	thr. 0.01	0.05	0.10
10	– (c=5)	– (c=5)	– (c=5)	20	1.90±0.19	– (c=5)	– (c=5)
20	0.48±0.04	1.34±0.05	– (c=5)	40	0.40±0.00	1.42±0.13	– (c=5)
30	0.30±0.00	0.50±0.00	0.76±0.05	60	0.30±0.00	0.54±0.05	0.90±0.00
40	0.20±0.00	0.34±0.05	0.50±0.00	80	0.30±0.00	0.66±0.05	0.98±0.04
50	0.20±0.00	0.30±0.00	0.40±0.00	100	0.32±0.04	0.72±0.04	1.12±0.04
60	0.20±0.00	0.30±0.00	0.40±0.00	120	0.48±0.04	1.02±0.08	1.44±0.09
70	0.20±0.00	0.30±0.00	0.40±0.00	140	0.66±0.05	1.10±0.10	1.46±0.11
80	0.20±0.00	0.32±0.04	0.50±0.00	160	0.68±0.04	1.04±0.05	1.26±0.05
90	0.20±0.00	0.40±0.00	0.58±0.04	180	0.66±0.05	1.02±0.04	1.34±0.11
100	0.20±0.00	0.40±0.00	0.64±0.05	200	0.66±0.05	1.18±0.13	1.80±0.32
120	0.30±0.00	0.60±0.00	0.90±0.07	220	0.68±0.04	1.42±0.08	2.38±0.15
140	0.40±0.00	0.70±0.00	0.92±0.04	240	0.70±0.00	1.50±0.07	2.68±0.13
160	0.54±0.05	0.92±0.08	1.14±0.05	260	1.08±0.68	2.52±1.39	3.40±0.24 (c=1)
180	0.72±0.04	1.08±0.04	1.36±0.05		–	–	–
200	0.72±0.04	1.16±0.05	1.56±0.05		–	–	–

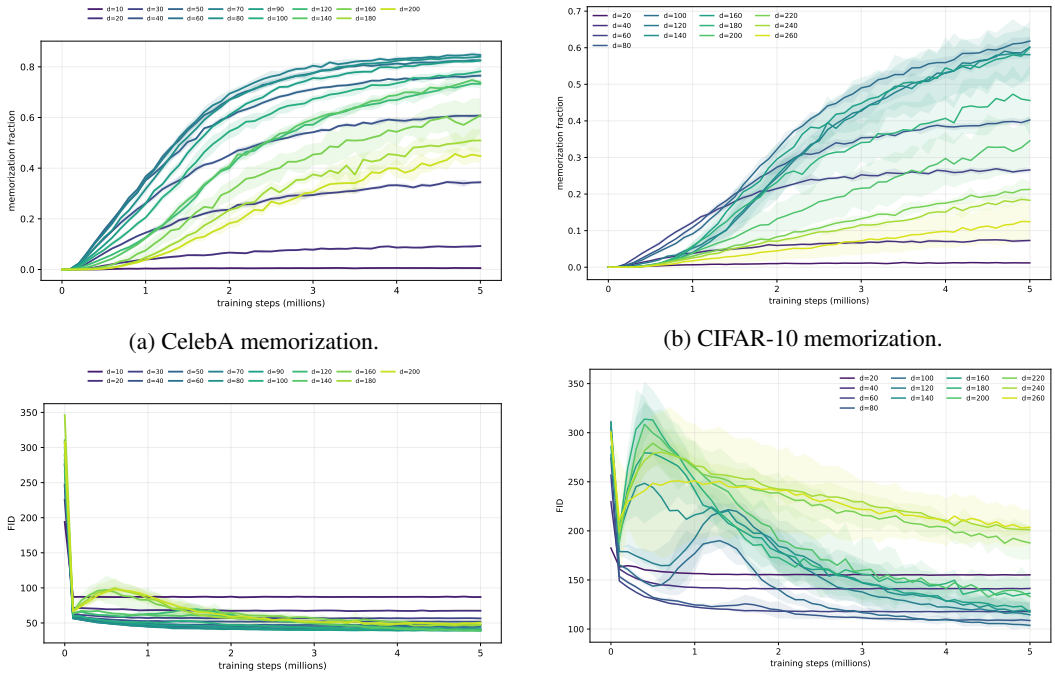


(a) CelebA.



(b) CIFAR-10.

Figure 12: Pixel-space ratio test on immediate VAE reconstructions of the 1k diffusion subset. Each row shows the detected fraction, mean distance ratio, and reconstruction MSE. These curves measure identity preservation by the encoder/decoder before score-model training. They locate the operational 95% identity thresholds near latent widths 70 and 140; low reconstruction copying is not evidence of a nonmemorizing diffusion process.



(a) CelebA memorization.

(b) CIFAR-10 memorization.

(c) CelebA FID.

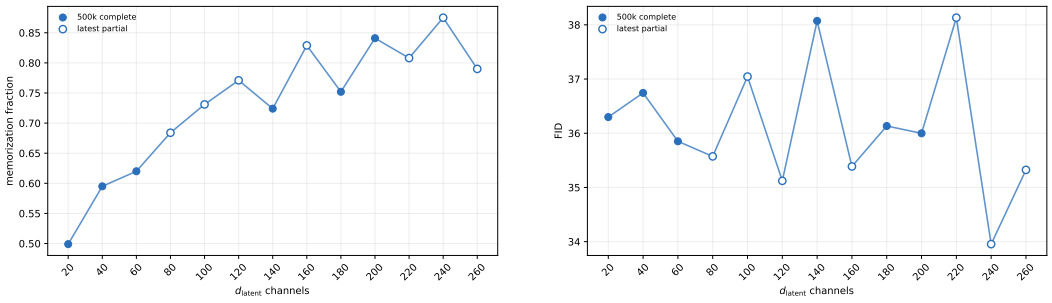
(d) CIFAR-10 FID.

Figure 13: Five-seed real-data trajectories with 95% confidence bands. Memorization applies the pixel-space ratio test to decoded samples; FID uses the same 1k training subset as its reference. The two statistics measure different behavior, and comparisons below the VAE identity threshold mix diffusion learning with representation loss.

E SPATIAL-DIT PILOT: A LIMIT OF ARCHITECTURAL TRANSFER

The CelebA pilot uses a spatial VAE with a $4 \times 4 \times d_{\text{latent}}$ latent grid. Here d_{latent} is channel count, so each increment adds sixteen scalar coordinates and changes the cross-position/channel covariance. The broad seed-42 sweep uses an unconditional DiT-S-style trunk with hidden size 384, depth 12, six attention heads, MLP ratio 4, batch size 1024, learning rate 10^{-4}, EMA decay 0.9999, and gradient clipping at 1. Evaluations at 20, 40, 60, 80, 100, 200, 300, 400, and 500k steps use 1000 generated samples and the same diverse 1k subset for pixel-space memorization and FID. Full per-channel trajectories and optimization diagnostics are in Appendix N.

The broad sweep does not show a clean high-width decline in memorization (Figure 14). The low-channel follow-up over seeds 39–43 instead increases from about 22% memorization at five channels to 48% at eight to ten channels, while FID improves from about 50 to 36 (Figure 15). This is negative transfer evidence for an architecture-independent width rule. Because the representation, score architecture, optimizer treatment, and scalar dimension all differ from the vector-MLP setting, it does not identify which change causes the difference.



(a) Memorization. (b) FID.

Figure 14: Latest available evaluation of the broad spatial-DiT pilot (seed 42). The horizontal axis counts channels; the scalar latent dimension is $16d_{\text{latent}}$. Memorization remains high across many widths and does not reproduce the fixed-vector-MLP decline.

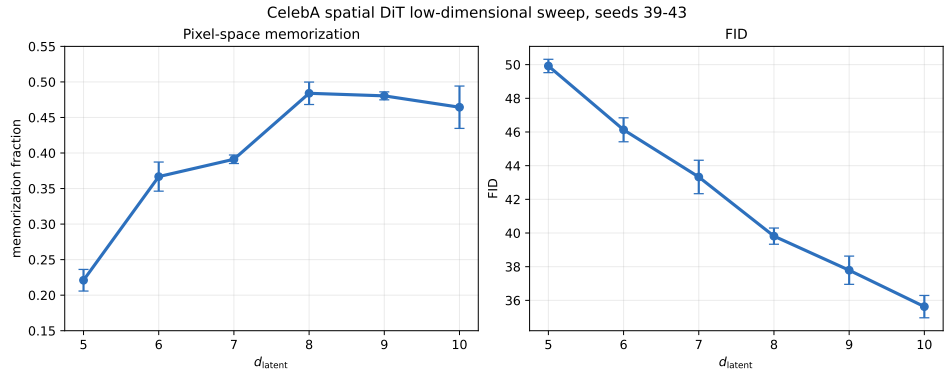


Figure 15: Low-channel spatial-DiT follow-up, seeds 39–43. Channels 5–10 correspond to 80–160 scalar coordinates. Memorization rises as FID improves; these results limit the main empirical claim to the tested vector-latent, fixed-width MLP setup.

F WHAT THE RELEASED SPECTRUM PREDICTOR FITS

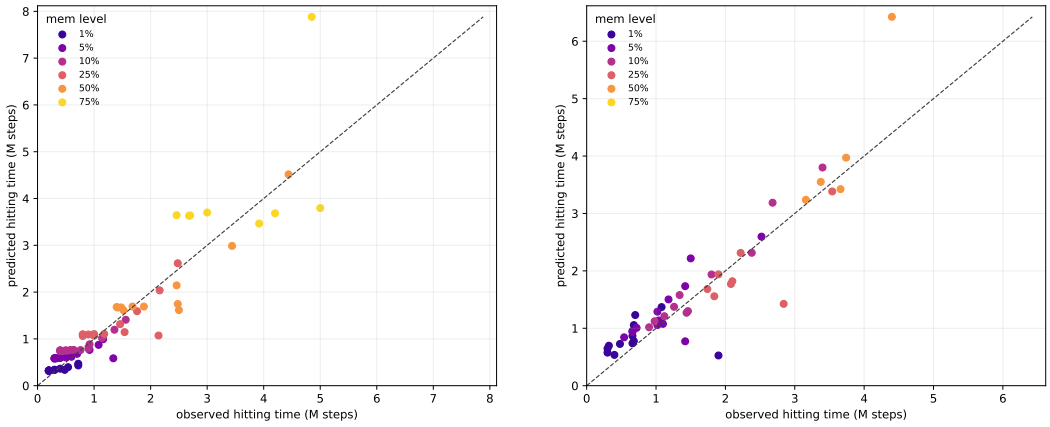
The released predictor uses frozen VAE encoder means, but it is calibrated against measured diffusion hitting times. It therefore does not inspect score-model weights yet is not an uncalibrated prediction from data alone. The full pressure curves, fitted threshold diagnostics and per-width overlays are in

Appendix O. For the same 1k subset let Z_d contain encoder means as rows. At $t = 0.1$,

$$M_t(d) = e^{-2t} Z_d^\top Z_d/n + \Delta_t I, \qquad P_d(s) = \frac{\sum_i w_i(d)(1 - e^{-\kappa\lambda_i(d)s})^2}{\sum_i w_i(d)},$$

where $\lambda_i(d)$ are eigenvalues of the uncentered second moment $M_t(d)$. The unweighted fit uses $w_i = 1$; the weighted diagnostic uses $w_i = \max(\lambda_i - \beta_d, 0)$ for an estimated isotropic floor β_d. These are data-side eigenvalues, not sample-mode RFNN rates. A single fitted κ also cannot explicitly represent the changing $a_1(q_d)^2/d$ factor of the RFNN surrogate.

The released scripts fit one global κ and store six ordered pressure thresholds θ_f for memorization fractions $f \in \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75\}$, with $\widehat{\tau}_f(d) = \inf\{s : P_d(s) \geq \theta_f\}$. They scan $\log_{10}\kappa$ from -12 to -4, compute empirical-pressure threshold averages weighted by the number of observed seeds, impose monotonicity by isotonic pooling, and minimize mean squared log hitting-time error. Calibration uses available observed hitting-time means across widths; it is not restricted to widths above the VAE identity threshold. These are one fitted rate and up to six fitted threshold levels per dataset, with no per-width fitting and no anchored sigmoid response in the released implementation. A fully censored threshold receives a filled value from neighboring levels rather than an independently fitted value; it supplies no observed validation target. Figure 16 compares the calibrated and observed hitting times.



(a) CelebA.

(b) CIFAR-10.

Figure 16: Released calibrated spectral-pressure fit. Predictions and observations share calibration data, so visual agreement is in-sample fit quality rather than held-out validation. Unreached thresholds are censored; low-width points may also be limited by VAE identity preservation. The separate one-constant sample-mode diagnostic and leave-one-width-out errors are reported in Appendix G.

The scripts exclude unobserved crossing times rather than fitting a survival model; means based on the seeds that cross can be biased downward when other seeds are censored. The one-constant g-clock discussed in Section 4 and Appendix G is a distinct diagnostic, not the mechanism used to generate these released plots. Its held-out error is compared with the released fit's in-sample error, and its calibration differs across datasets. Neither fit establishes a trained-network mechanism. Finally, changing a query threshold to $1/n$ while holding the saved $n = 1000$ spectrum fixed changes only the queried fraction; it is not an experimental training-set-size sweep. Since the fraction is computed over generated samples, $1/n$ does not in general mean one generated copy or one distinct memorized training identity. A claim about how dynamics depend on n would require recomputing the subsets and spectra and retraining or reevaluating the models.

G Supporting Theory and Spectral Diagnostics

The finite-dimensional gradient-flow identities below are exact for a fixed feature operator. The spectral interpretation is a conditional approximation checked against the released spectra. Its connection to a generated-sample memorization fraction, and its transfer to trained features, remain

unproved. We use $d = d_{\text{latent}}$, $D = d_{\text{intrinsic}}$, $n = n$, $p = p$ and $\Delta_t = 1 - e^{-2t}$ throughout this appendix.

G.1 Feature operators and exact training dynamics

At a fixed $t > 0$, let $x_t^\mu = e^{-t}x^\mu + \sqrt{\Delta_t}\eta^\mu$, $y^\mu = -\eta^\mu/\sqrt{\Delta_t}$, and $\phi(x) = p^{-1/2}\tanh(Wx/\sqrt{d})$ with independent standard Gaussian entries of the frozen W. The trained readout is $s_A = A\phi$, initialized at zero. Define

$$\begin{aligned} L_n(A) &= \frac{1}{2n}\sum_{\mu=1}^{n}\mathbb{E}_\eta\|A\phi(x_t^\mu) - y^\mu\|^2, \\ U &= \frac{1}{n}\sum_\mu \mathbb{E}_\eta[\phi\phi^\top], \qquad V = \frac{1}{n}\sum_\mu \mathbb{E}_\eta[y\phi^\top]. \end{aligned} \tag{12}$$

We write U for the main text's $\boldsymbol{U}$. Distinguish the one-draw operator $U_k^{(1)}$, its average $U^{(K)} = K^{-1}\sum_k U_k^{(1)}$, and its expectation $U = \mathbb{E}_\eta U_k^{(1)}$. Their ranks obey

$$\operatorname{rank} U_k^{(1)} \le \min(p, n), \qquad \operatorname{rank} U^{(K)} \le \min(p, Kn).$$

Thus $p - n$ zero eigenvalues are guaranteed only for a single draw when $p > n$. Neither the expectation nor the released $K = 50$ spectra inherit that count; the fixed-q control uses $K = 40$. The exact conditional-covariance decomposition is

$$U = B + C, \qquad B = \frac{1}{n}\sum_\mu m_\mu m_\mu^\top, \qquad C = \frac{1}{n}\sum_\mu \operatorname{Cov}_\eta(\phi(x_t^\mu)), \qquad m_\mu = \mathbb{E}_\eta\phi(x_t^\mu). \tag{13}$$

Both terms are positive semidefinite and $\operatorname{rank} B \le n$. This supplies no constant spectral floor without additional information about C.

Bonnaire et al. (2025, Theorems 3.1–3.2 and Appendix C.5) analyze Gaussian clean-data covariance and a noise-averaged operator. Their asymptotic floor weight is $1 - (n + d)/p$ in the stated regime $p > n + d$, distinct from the single-copy rank deficit $1 - n/p$; diffusion is already part of their construction. Their covariance argument cannot simply be replaced by already-diffused atoms. Our finite-rank Gaussian mixture also needs a separate outlier argument. George et al. (2026, Appendix B) use a five-block pencil for Gaussian subspace data; its block dimensions do not count spectral bands. These results motivate the approximation below, rather than proving its four groups.

Lemma G.1 (Exact readout dynamics)**.** *Let $U = \sum_i \lambda_i v_i v_i^\top \succeq 0$, with decreasing eigenvalues. The minimum-norm optimum is $A^\star = VU^+$, and gradient flow from zero satisfies*

$$\begin{aligned} A(T) &= A^\star(I - e^{-UT}), \qquad A(T)v_i = a_i^\star(1 - e^{-\lambda_i T}), \\ a_i^\star &= \begin{cases} Vv_i/\lambda_i, & \lambda_i > 0, \\ 0, & \lambda_i = 0. \end{cases} \end{aligned} \tag{14}$$

With $\pi_i = \|a_i^\star\|^2$, the exact residual identities are

$$\|A(T) - A^\star\|_F^2 = \sum_i \pi_i e^{-2\lambda_i T}, \qquad L_n(A(T)) - L_n(A^\star) = \frac{1}{2}\sum_i \lambda_i \pi_i e^{-2\lambda_i T}. \tag{15}$$

A mode is active only when $\pi_i > 0$; full rank alone does not ensure activity.

Proof. If $u \in \ker U$, the nonnegative sum $n^{-1}\sum_\mu \mathbb{E}(u^\top\phi)^2$ is zero, hence $Vu = 0$. Therefore $A^\star U = V$. Completing the square in L_n gives the minimum-norm optimum and $\nabla L_n = AU - V$. Solving $\dot A = -(A - A^\star)U$ gives (14); orthogonality of the v_i and the completed square give (15). □

The saved-sweep code stores $W/\sqrt{d}$ and logs unnormalized feature correlations, so $\lambda_{i,\text{code}} = p\lambda_i$. Its fresh-draw loss and its expectation are

$$\ell_k(A) = \frac{1}{dn}\sum_\mu \|\sqrt{\Delta_t}A\phi(x_t^{\mu,k}) + \eta^{\mu,k}\|^2, \qquad \mathbb{E}_\eta\ell_k(A) = \frac{2\Delta_t}{d}L_n(A).$$

Fresh-noise full-batch SGD, with learning rate η_{lr} and zero momentum, has step $A_{k+1} = A_k - h(A_k U_k^{(1)} - V_k^{(1)})$, where $h = 2\Delta_t \eta_{\mathrm{lr}}/d$. Independence of the fresh draw from A_k gives the exact mean Euler recursion $\mathbb{E}A_{k+1} = \mathbb{E}A_k - h(\mathbb{E}A_k U - V)$. The default $\eta_{\mathrm{lr}} = 0.01d/\Delta_t$ therefore gives $T = 0.02k$, with flow multiplier $e^{-0.02\lambda_i k}$ approximating the exact mean multiplier $(1 - 0.02\lambda_i)^k$. A signal e-fold is $p/(0.02\lambda_{D,\mathrm{code}})$ steps: about 788 at $d = 5$ and 8663 at $d = 200$. Loss of the mean iterate is not mean stochastic loss; parameter variance, finite-step error and Monte Carlo error are separate effects. This clock does not apply to the trainable MLP sweeps.

G.2 What spectral absorption says about a train–test gap

Let $\bar{U} = \mathbb{E}_{\mathrm{pop}}[\phi\phi^\top]$, $\bar{V} = \mathbb{E}_{\mathrm{pop}}[y\phi^\top]$, and $L(A) = \frac{1}{2}\mathbb{E}_{\mathrm{pop}}\|A\phi - y\|^2$. Put $u_i(T) = 1 - e^{-\lambda_i T}$, $\bar{u}_{ij} = v_i^\top \bar{U} v_j$, $\bar{u}_i = \bar{u}_{ii}$ and, for active modes, $\gamma_i = \langle a_i^\star, \bar{V} v_i\rangle/(\lambda_i \pi_i)$.

Proposition G.2 (Exact gap and its alignment requirement)**.** *The change in generalization gap from initialization is*

$$\begin{aligned} G(T) &= [L(A(T)) - L_n(A(T))] - [L(0) - L_n(0)] \\ &= \frac{1}{2}\sum_{i,j} u_i u_j \bar{u}_{ij}\langle a_i^\star, a_j^\star\rangle - \sum_i \gamma_i \lambda_i \pi_i u_i + \frac{1}{2}\sum_i \lambda_i \pi_i (1 - e^{-2\lambda_i T}). \end{aligned} \tag{16}$$

Terms with $\pi_i = 0$ are zero. For any chosen active sample subset $\mathcal{B}$, define

$$G_{\mathcal{B}}^0(T) = \frac{1}{2}\sum_{i\in\mathcal{B}} \pi_i[\bar{u}_i u_i(T)^2 + \lambda_i(1 - e^{-2\lambda_i T})].$$

It is increasing from zero. Dropping its diagonal alignment terms costs at most $\sum_{i\in\mathcal{B}} |\gamma_i|\lambda_i\pi_i u_i(T)$. In particular, a bound by $\gamma_0 \sum_{\mathcal{B}} \lambda_i \pi_i$ requires $|\gamma_i| \le \gamma_0$, not merely $\gamma_i \le \gamma_0$.

Proof. Insert $A(T) = \sum_i u_i a_i^\star v_i^\top$ into the two quadratic losses. The empirical quadratic and linear terms combine as $\lambda_i\pi_i(u_i - u_i^2/2) = \lambda_i\pi_i(1 - e^{-2\lambda_i T})/2$. The remaining claims follow by the triangle inequality and monotonicity of each nonnegative bracket. □

Definition G.3 (Index edges and memorization interpretation)**.** For $p \ge n > d > D$, use the index groups $\mathcal{S} = \{1, \dots, D\}$, $\mathcal{N} = \{D+1, \dots, d\}$, $\mathcal{M} = \{d+1, \dots, n\}$ and $\mathcal{Z} = \{n+1, \dots, p\}$. Interpreting $\mathcal{S}, \mathcal{N}$ as population modes requires their alignment with the population target and control of the off-diagonal sum in (16). Interpreting $\mathcal{M}$ as sample-specific additionally requires small $|\gamma_i|$ on the relevant active subset. An index-based edge alone checks neither condition. If population-aligned modes are removed from $\mathcal{M}$, the remaining top eigenvalue can only decrease, so its inverse timescale can only increase.

For perfectly aligned diagonal modes ($\bar{u}_i = \lambda_i$, $\gamma_i = 1$), the diagonal gap contribution vanishes. Pure sample modes ($\gamma_i = 0$) contribute to $G_{\mathcal{B}}^0$. More generally, population-block errors must be controlled in the weighted norm $\sum_{i\in\mathcal{S}\cup\mathcal{N}} \lambda_i\pi_i(|\bar{u}_i/\lambda_i - 1| + 2|\gamma_i - 1|)$, together with the off-diagonal sum. These quantities have not been measured on the released sweep. Small tail eigenvalues also do not imply negligible target mass: $\pi_i = \|V v_i\|^2/\lambda_i^2$. For example, $\|A(T)P_{\mathcal{Z}}\|_F^2 \le T^2\|VP_{\mathcal{Z}}\|_F^2$, so a tail assessment needs target weights. We use edge times only as spectral proxies, without claiming a derived nearest-neighbor memorization threshold.

G.3 Operating point and conditional spectral scales

Lemma G.4 (Operating point and Hermite coefficients)**.** *For the synthetic construction with center norm s, within-signal variance σ_{sig}^2 and null variance $\sigma_\perp^2$, the expected per-coordinate second moment is*

$$q(d) = \Delta_t + e^{-2t}\left[\sigma_\perp^2 + \frac{s^2 + D(\sigma_{\mathrm{sig}}^2 - \sigma_\perp^2)}{d}\right]. \tag{17}$$

This is an expectation over the training-set construction; realized traces fluctuate. It decreases with d if $s^2 + D(\sigma_{\text{sig}}^2 - \sigma_\perp^2) > 0$. For $z \sim \mathcal{N}(0,1)$ define

$$a_1(q) = \mathbb{E}\,\mathrm{sech}^2(\sqrt{q}z), \qquad c_k(q) = \frac{\mathbb{E}[\tanh(\sqrt{q}z)\mathrm{He}_k(z)]}{\sqrt{k!}},$$
$$a_*(q)^2 = \sum_{k\geq 3 \text{ odd}} c_k(q)^2. \tag{18}$$

Then $c_1 = \sqrt{q}\,a_1$, a_1 decreases with q, and

$$a_1(q)^2 q + a_*(q)^2 = \mathbb{E}\tanh^2(\sqrt{q}z), \qquad a_*(q)^2 = \frac{2}{3}q^3\left[1 - 8q + \frac{266}{5}q^2 + O(q^3)\right]. \tag{19}$$

Proof. The expected trace of the clean second moment is $s^2 + D\sigma_{\text{sig}}^2 + (d-D)\sigma_\perp^2$; diffusion adds $d\Delta_t$. Stein's identity gives c_1 and Hermite Parseval gives the trace identity. Monotonicity of a_1 follows pointwise from $\mathrm{sech}^2(\sqrt{q}z)$. The Taylor series of $\tanh$ gives $c_3^2 = \frac{2}{3}q^3(1 - 8q + 50q^2 + O(q^3))$ and $c_5^2 = \frac{32}{15}q^5 + O(q^6)$; their sum gives the displayed coefficient. Numerical tables use quadrature, not this small-q expansion. □

For fixed input points, Gaussianity over a row of W gives the exact kernel

$$K_{\mu\nu} = \sum_{k \text{ odd}} c_k(q_\mu)c_k(q_\nu)\rho_{\mu\nu}^k$$
$$= a_1(q_\mu)a_1(q_\nu)\frac{x_t^\mu \cdot x_t^\nu}{d} + (K_{\text{res}})_{\mu\nu}, \tag{20}$$

where $q_\mu = \|x_t^\mu\|^2/d$, $\rho_{\mu\nu} = x_t^\mu \cdot x_t^\nu/(d\sqrt{q_\mu q_\nu})$ and $(K_{\text{res}})_{\mu\mu} = a_*(q_\mu)^2$. The identity follows from orthogonality of normalized Hermite polynomials for jointly Gaussian variables. It does not require Gaussian input data. It also does not make residual features independent: same-cluster overlaps and variation in q_μ remain in K_{res}.

Assumption G.5 (Spectral surrogate and separation)**.** To obtain the scales below, assume that replacing finite random features by a Gaussian matrix with population kernel K approximates the relevant bulk and extreme eigenvalues at their own scales. Assume separately that this approximation extends to repeated diffusion draws, that the data directions separate from the residual directions, and that empirical signal–null cross terms and heterogeneity of $a_1(q_\mu)$ can be neglected for the data-side approximation. These are additional hypotheses for the mixture, not consequences of the exact kernel identity or of weak spectral convergence alone.

Gaussian-equivalence and nonlinear spectral results motivate such replacements under their respective hypotheses (Pennington & Worah, 2017; Benigni & Péché, 2021; Goldt et al., 2022; Hu & Lu, 2023; Mei & Montanari, 2022). They do not automatically establish finite-rank outlier transfer or joint repeated-noise equivalence here. In particular, bounded point norms do not imply the input-operator control used by Louart et al. (2018); mixture and approximate-orthogonality results also impose conditions on overlaps (Liao & Couillet, 2018; Fan & Wang, 2020).

Let $\widehat{M}_0 = n^{-1}\sum_\mu x^\mu x^{\mu\top}$, $M_t = e^{-2t}\widehat{M}_0 + \Delta_t I$ and $\ell_{\min}$ be the smallest eigenvalue of its clean signal-subspace block. This empirical second moment includes center–noise cross terms. The main text's single signal atom is an isotropic-center surrogate: fixed random centers need not have equal signal eigenvalues, even in the population. Under Assumption G.5, the candidate data-side scales are

$$\lambda_{\min}^{\mathcal{S}} \approx \frac{a_1(q)^2}{d}(e^{-2t}\ell_{\min} + \Delta_t),$$
$$\lambda_\pm^{\mathcal{N}} \approx \frac{a_1(q)^2}{d}\left[e^{-2t}\sigma_\perp^2(1 \pm \sqrt{(d-D)/n})^2 + \Delta_t\right]. \tag{21}$$

The second formula uses the shifted Marchenko–Pastur approximation for the null sample covariance. Finite p adds a Wishart factor $W^\top W/p$; it cannot be discarded when locating finite-ratio edges. The single-draw null covariance instead has diffusion inside its sampled variance. Fixed-D signal directions are outliers, not a bulk; using realized geometry improves the signal calibration over replacing all signal eigenvalues by their mean. At every fixed $t > 0$, $M_t \succeq \Delta_t I$: no inverse-eigenvalue clock of this linear data covariance diverges as $d/n \to 1$.

Table 4: Released RFNN calibration: $n = 500$, $p = 64d$, $t = 0.01$, $\sigma_\perp = 0.5$, seed 42, $K = 50$. Signal minima are in code units (p times theory units); their conditional predictions use realized signal geometry. κ_{top} is measured, not fitted by an independent edge model. At $d = 5$, $p = 320 < n$, so the tail group is empty and the sample group ends at p.

d	q	a_1^2	a_*^2	signal min: measured/predicted	κ_{top}	R
5	2.76	0.180	0.0785	20.3 / 20.2	19.0	21
10	1.52	0.281	0.0447	31.9 / 31.5	16.3	34
20	0.89	0.393	0.0235	42.0 / 44.2	11.1	63
40	0.58	0.494	0.0125	53.6 / 55.9	5.85	143
100	0.39	0.586	0.0064	67.5 / 67.8	2.41	340
200	0.33	0.625	0.0047	73.9 / 75.1	1.45	426

Remark G.6 (Index groups and approximate scales). For $D < d < n < p$, partition the measured spectrum of $U^{(K)}$ at indices D, d, n. In separated configurations, the groups have candidate scales $a_1(q)^2\alpha_t^2/d$ for D signal modes, $a_1(q)^2\beta_t^2/d$ for $d - D$ noise-dimension modes, and $\kappa_{\text{top}}a_*(q)^2/n$ for the top of the $n - d$ sample group. Here $\alpha_t^2 = e^{-2t}(s^2/D + \sigma_{\text{sig}}^2) + \Delta_t$ and $\beta_t^2 = e^{-2t}\sigma_\perp^2 + \Delta_t$ are isotropic-center surrogates, so $q = \beta_t^2 + D(\alpha_t^2 - \beta_t^2)/d$. The $p - n$ tail modes vanish for a full-row-rank single-copy feature matrix but need not vanish after noise averaging. These are index counts, not a theorem identifying four spectral components. The scales require Assumption G.5, and the signal/sample labels additionally require alignment. The Hermite variance identity does not establish individual block masses for fixed W.

The noise-dimension/sample gap is 0.6–1.3 decades in the reported $\sigma_\perp = 0.5$ configurations and nearly closes at $\sigma_\perp = 0.01$ (Figure 7).

G.4 Edge ratios and their measured calibration

Proposition G.7 (Spectral absorption proxies)**.** *For positive edges and a common relative parameter tolerance $\varepsilon \in (0, 1)$, the exact gradient-flow proxies are*

$$\begin{aligned} \tau_{\text{gen}} &= \frac{\log(1/\varepsilon)}{\lambda_D}, & \tau_{\text{mem}}^{\text{on}} &= \frac{\log(1/\varepsilon)}{\lambda_{d+1}}, \quad R = \frac{\tau_{\text{mem}}^{\text{on}}}{\tau_{\text{gen}}} = \frac{\lambda_D}{\lambda_{d+1}}. \\ R &\approx \frac{n}{d}\frac{a_1(q)^2}{a_*(q)^2}\frac{e^{-2t}\ell_{\min} + \Delta_t}{\kappa_{\text{top}}}, & \kappa_{\text{top}} &:= \frac{n\lambda_{d+1}}{a_*(q)^2}. \end{aligned} \tag{22}$$

Only the first line is an exact dynamical identity. The second inserts the signal approximation of (21) *and a measured residual-edge coefficient κ_{top}, so it is a decomposition rather than a parameter-free prediction of the sample edge.*

Proof. Each active mode has relative parameter residual $e^{-\lambda_i T}$. Taking the slowest signal edge and fastest sample-group edge gives the first line; substituting the edge scale and the definition of κ_{top} gives the second. □

Every active mode begins relaxing at initialization. Consequently a count of intermediate modes cannot multiply an absorption time. Added dimensions affect R through the $1/d$ normalization, q, realized signal geometry and the residual edge. Even growth of a_1^2/a_*^2 does not by itself establish increasing R. The ratio λ_d/λ_{d+1} compares the slowest entire data-side direction to the sample edge and can behave differently. A sample-block median and its top edge likewise describe different timescales.

Table 4 separates the measured ratio from the approximated signal scale. Across the released sweep the signal e-fold grows about $11\times$ and the sample-edge e-fold about $221\times$. The latter combines a roughly $17\times$ reduction of a_*^2 with a $13\times$ reduction of κ_{top}. Monotonicity of a_*^2 over this operating range is a quadrature observation, not a continuous-interval theorem. The residual top is heterogeneous at small d; attributing its entire motion to the Hermite coefficient would be incorrect. For the noise-dimension minimum, measured/predicted code eigenvalues at $d = 10, 20, 40, 100, 200$ are $4.10/3.92$, $5.16/4.71$, $4.90/4.81$, $3.57/3.66$, $2.12/2.17$. Agreement of a trace does not validate

individual spectral groups. The K-draw sample group and tail can be contiguous; at $\sigma_\perp = 0.01$ the noise-dimension/sample gap nearly closes. The four-group interpretation is useful only where separation and alignment are supported.

Remark G.8 (Fixed operating-point control). Set $\sigma_\perp^2(d) = (0.57d - 14)/(d - D)$ for $D = 5$, holding the expected clean input variance per coordinate at 0.57. Then diffusion also preserves a common q across widths. At $n = 500$, $p = 4096$, $t = 0.01$, $K = 40$ and three seeds, widths $d = 40, 60, 80, 120, 160, 240$ give $R = 92, 140, 158, 136, 119, 100$, versus $92, 145, 201, 266, 287, 291$ in the baseline. The endpoint change is $1.09\times$ rather than $3.17\times$, although the noise-dimension count grows $6.7\times$. Signal edges fall 4.3–$4.4\times$ in both sweeps, while sample edges fall $4.8\times$ at fixed q and $13.6\times$ in the baseline. This intervention supports an operating-point contribution; it also changes the null covariance, and the fixed-q curve is non-monotone.

Remark G.9 (Isotropic control). With $\sigma_\perp = \sigma_{\text{sig}} = 1$ at $n = 500$, $p = 4096$, $t = 0.01$ and three seeds, $R = 14, 30, 58, 87, 77, 60$ for $d = 5, 10, 20, 40, 80, 160$. Thus the ratio peaks near 40 and then falls. Null-coordinate modes still exist, but their variance is no longer separated from within-signal variance, and de-saturation is much weaker. This is a spectral simulation, not a trained-network memorization result.

G.5 An exact linear comparison

For the population second moment $M = \mathbb{E}[x_t x_t^\top] \succ 0$, consider $s_B(x) = Bx$ and the half-scaled score risk $\frac{1}{2}\mathbb{E}\|Bx_t - s^\star(x_t)\|^2$. Integration by parts gives $\mathbb{E}[s^\star(x_t)x_t^\top] = -I$, so from $B(0) = 0$,

$$\begin{aligned} &\dot{B} = -(BM + I), \qquad B(T) = -M^{-1}(I - e^{-MT}), \\ &\mathbb{E}\|B(T)x_t - s^\star(x_t)\|^2 = R_{\text{irr}} + \operatorname{tr}(M^{-1}e^{-2MT}), \end{aligned} \tag{23}$$

where $R_{\text{irr}} = \mathbb{E}\| - M^{-1}x_t - s^\star(x_t)\|^2$. The identity follows by solving the linear differential equation and orthogonality of the best linear residual to x_t. In the population mixture, the null coordinates factor as a Gaussian with variance $\beta_t^2 = e^{-2t}\sigma_\perp^2 + \Delta_t$ and exact score $-x_{t,\perp}/\beta_t^2$. They contribute $(d-D)\beta_t^{-2}e^{-2\beta_t^2 T}$ to the score error, each with the same rate. For an empirical training set, noise-averaged denoising training replaces M by $\widehat{M} = e^{-2t}n^{-1}\sum_\mu x^\mu x^{\mu\top} + \Delta_t I$ in the dynamics. This illustrates parallel absorption and the positive diffusion floor; it does not derive a nonlinear sample-specific spectrum or near-duplicate generation.

G.6 Real-data spectral diagnostics and transfer limits

The real-data calculations are evaluations of a frozen $\tanh$ surrogate at $t = 0.1$, not identities for the trained depth-five MLP. Two diagnostics use different residual scales. For the first, set $q_d = \operatorname{tr} M_t/d$ from the frozen latent spectrum and, for $0 \le \rho \le 1$, define the coherent Hermite residual

$$g(q, \rho) = \sum_{k \ge 3 \text{ odd}} c_k(q)^2 \rho^k, \qquad \widehat{\tau}_f(d) = \frac{Cn}{g(q_d, 1 - \Delta_t/q_d)}. \tag{24}$$

The series is an exact homogeneous Gaussian kernel expression, increasing in ρ with $g(q, 1) = a_*^2$. Using it as a per-sample eigenvalue or a memorization clock requires additional localization, noise-covariance and sampler assumptions, which are not established here. Thus (24) is a one-constant diagnostic. On all but one width, fit $\log C$ as the median of $\log[\tau_f(d)g(q_d, 1 - \Delta_t/q_d)/n]$, then evaluate absolute log error on the held-out width, using the empirical hitting times of Table 3. Table 5 reports the average held-out error; the fitting intervals include widths below the identity threshold.

Table 5: Mean absolute log error for the 5% hitting time. The one-constant clocks use leave-one-width-out evaluation; the released calibrated predictor is an in-sample comparison, not the same validation protocol.

Widths	coherent g-clock	$\tau \propto d$	released predictor
CelebA, $d \ge 50$ (11 widths)	0.407	0.125	0.339
CIFAR-10, $d \ge 60$ (11 widths)	0.151	0.198	0.146

The g-clock captures $1.78\times$ of the $3.9\times$ slowdown over CelebA $50 \to 200$, and its fitted C differs by $2\times$ between datasets. A dimension-only clock performs better on the stated CelebA interval. These results delimit its explanatory value and do not establish cross-dataset prediction.

The second diagnostic uses pure a_*^2 time rescaling: relative to width d_0, assume

$$\frac{\widehat{\tau}_f(d)}{\tau_f(d_0)} = \frac{a_*(q_{d_0})^2}{a_*(q_d)^2}, \qquad \widehat{F}_d(T) = F_{d_0}\left(T\,\frac{a_*(q_d)^2}{a_*(q_{d_0})^2}\right). \tag{25}$$

This holds the remaining dynamics and response curve fixed; it is a modeling choice, distinct from the coherent g-clock. The archived CelebA spectral summaries report clean traces $252.5, 255.2, 260.4$, effective dimensions $89.5, 90.3, 87.9$, and $q = 1.473, 1.342, 1.247$ at $d = 160, 180, 200$. At $d \geq 180$, the median of the lowest 30% of eigenvalues is within 2.4×10^{-4} of Δ_t, supporting a floor fraction of at least 15%. Here the released effective-dimension estimator uses the participation ratio above an estimated floor; it equals the clean-covariance participation ratio only where that floor equals Δ_t. These summaries indicate small added-coordinate variance; flat trace and participation ratio do not establish an unchanged live spectrum. This comparison uses archived spectral summaries rather than a new spectral measurement.

Table 6: Pure q-rescaling diagnostic relative to $d = 160$. Final fractions are at 5M steps. Observed 5% ratios use mean hitting times in Table 3; CIFAR-10 at $d = 260$ has substantial seed variation.

Dataset, widths	quantity	surrogate	observed
CelebA, $160 \to 180$	5% hitting-time ratio	1.11	$1.08/0.92 = 1.17$
	final fraction	0.595	0.509
CelebA, $160 \to 200$	5% hitting-time ratio	1.21	$1.16/0.92 = 1.26$
	final fraction	0.559	0.448
CIFAR-10, $160 \to 260$	5% hitting-time ratio	1.54	$2.52/1.04 = 2.42$
	final fraction	0.472	0.125

The rescaling gives the observed direction but a smaller magnitude. On CelebA it accounts for about 0.55–0.82 of the log-slowdown across the reported thresholds and $\log(0.559/0.608)/\log(0.448/0.608) = 0.28$ of the final-fraction log-decline. On CIFAR-10 the 5% values give $\log(1.54)/\log(2.52/1.04) = 0.49$, and the final-fraction comparison gives $\log(0.472/0.600)/\log(0.125/0.600) = 0.15$. These are descriptive ratios, not estimates of a causal share explained. The remaining differences lie outside this surrogate.

All these calculations use an isotropic $\tanh$ layer with unit pre-activation gain. The deployed GELU MLP uses learned features, time embeddings and momentum across diffusion times. Even its initialization has a different operating point: with fan-in $d + 64$, the expected first-layer variance is approximately $(e^{-2t}\|z\|^2 + d\Delta_t + 32)/(3(d+64))$, nearly flat across the reported widths. A change in gain alters both the variance and the correlation between noisy views; equal q alone is not sufficient to preserve the coherent kernel. Neither an individual-training-point eigenvector assignment nor a map from spectral absorption to decoded nearest-neighbor events is proved. The trained-network measurements in Section 4 therefore remain necessary: the surrogate supports one mechanism in frozen features, while the empirical delay is the main evidence in learned representations.

Table 7: Signal-edge calibration for the sweep in Figure 2. Edge values are in saved-code units, p times the eigenvalues of (3); R is unit invariant. The conditional prediction uses realized signal geometry. Additional widths and coefficients are in Table 4.

d_{latent}	q	signal minimum, measured	conditional prediction	measured R
10	1.52	31.9	31.5	34
40	0.58	53.6	55.9	143
100	0.39	67.5	67.8	340
200	0.33	73.9	75.1	426

Extended diagnostic supplement

This supplement provides the extended RFNN width and noise sweeps, VAE training and reconstruction checks, per-width real-data trajectories, the spatial-DiT pilot, and spectral-predictor diagnostics supporting the preceding appendices.

H GUIDE TO THE EXTENDED DIAGNOSTICS

The supplement follows the experimental progression of the main paper. Appendix I gives implementation details, and Appendix J presents the RFNN width and noise sweeps. Appendix K contains the trainable synthetic-MLP memorization and score diagnostics. For the image experiments, Appendix L gives VAE training and reconstruction checks, and Appendix M gives the per-width memorization and FID trajectories. The spatial-DiT pilot is in Appendix N; the full frozen-latent spectrum predictor diagnostics are in Appendix O.

I EXPERIMENTAL DETAILS

Synthetic latent data. The synthetic experiments draw a low-dimensional signal subspace with dimension $d_{\text{intrinsic}}$ and embed it in an observed latent space of dimension d_{latent}. Unless noted otherwise, the signal coordinates are sampled from an anisotropic Gaussian mixture and the remaining coordinates receive independent null noise. The nonlinear random-feature experiments use the same data design but study the spectrum of the feature kernel rather than generated samples. The trainable synthetic MLPs minimize denoising score matching with the conditional noise target. The analytic population GMM score is used only for evaluation; it is not a training target.

Synthetic experiment details. For Experiment 1, we fix $d_{\text{intrinsic}} = 5$ and sweep d_{latent}, so the exact null-buffer size is $d_{\text{latent}} - d_{\text{intrinsic}}$. For Experiment 2, we fix the observed latent space at $d_{\text{latent}} = 20$ and sweep $d_{\text{intrinsic}}$, so the exact null-buffer size decreases until it vanishes at $d_{\text{intrinsic}} = d_{\text{latent}}$. The default RFNN data are anisotropic GMM samples with $\sigma_{\text{signal}} = 1.0$, $\sigma_{\perp} = 0.5$, center scale 3.0, $n = 500$, diffusion time $t = 0.01$, and K Monte Carlo diffusion-noise samples used to estimate $\boldsymbol{U}$ ($K = 50$ for the $p = 64d_{\text{latent}}$ raw-data spectra, $K = 500$ for the $p = d_{\text{latent}} + n + 300$ set; see each run's `config.json`). A single-draw estimate has rank at most $\min(p, n)$, while the K-draw average has rank at most $\min(p, Kn)$ and can be full rank. The noise-averaged tail is therefore not an exact null space. The Gaussian robustness panels remove mixture centers and either sweep $\sigma_{\perp} \in \{0.1, 0.3, 0.5\}$ at fixed $\sigma_{\text{signal}} = 1.0$, or sweep $\sigma_{\text{signal}} \in \{0.5, 1.0, 1.5\}$ at fixed $\sigma_{\perp} = 0.3$. The synthetic MLP uses the same $d_{\text{intrinsic}} = 5$, $n = 500$, $\sigma_{\perp} = 0.5$ latent-distribution family as the main RFNN sweep, trains a fixed-hidden-width score MLP for 5M steps over seeds 42–46, and evaluates sample recall by the nearest-neighbor ratio on generated synthetic samples.

Width controls. The RFNN spectra are reported under three width rules: $p = 64d_{\text{latent}}$, $p = d_{\text{latent}} + n + 300$, and fixed $p = 1800$ with fixed total null energy. The first setting checks a high-feature-count regime. The second grows width just enough to cover the n data-supported directions plus a $d_{\text{latent}} + 300$ margin, so $\psi_p = p/d_{\text{latent}}$ approaches 1 at large d_{latent} and the noise-dimension block broadens accordingly. The fixed-width/fixed-energy control removes the two main confounds in the d_{latent} sweep: increasing model width and increasing total injected null variance.

Memorization metric. A generated sample $\hat{x}$ is counted as memorized when the ratio of its distances to its nearest and second-nearest training points, $\|\hat{x} - \text{NN}_1\|/\|\hat{x} - \text{NN}_2\|$, is below $1/3$, the nearest-neighbor ratio criterion of Bonnaire et al. (2025) (see also Somepalli et al., 2023). Synthetic experiments apply it in latent space; image experiments apply it in pixel space to decoded samples. The relevant comparison set is the training subset used by the diffusion model, not the full VAE training set. Reported curves are the mean across five seeds with confidence bands unless the caption states that the figure is a deterministic VAE diagnostic.

VAE reconstruction identity threshold. Before training diffusion, we pass the same diverse 1k subset through each frozen VAE and apply the pixel-space nearest-neighbor ratio to the immediate reconstructions. This is not diffusion memorization. It measures whether the VAE bottleneck preserves enough sample identity for any downstream pixel-space memorization test to be meaningful. We define the operational identity threshold as the smallest width at which 95% of these reconstructions pass the ratio test: approximately 70 for CelebA and 140 for CIFAR-10. This is not an estimate of manifold intrinsic dimension. Decoder changes can affect the metric even above this threshold.

Real-data VAEs. CelebA and CIFAR-10 experiments first train one VAE per latent dimension on the full available training split, not on the 1k diffusion subset. The VAE is a five-stage convolutional encoder/decoder with channel widths $(32, 64, 128, 256, 512)$. The final checkpoints use the ResNet-style variant: stride-two convolutional downsampling, GroupNorm/SiLU residual blocks, transpose-convolution upsampling, and a Tanh output. CelebA uses 21k training and 6k validation images from the local CelebA-HQ tar split at 32×32 grayscale resolution; CIFAR-10 uses the standard 50k/10k train/test split at 32×32 RGB resolution. The objective combines reconstruction with a KL penalty to the standard normal prior. In the capacity-loss form used by the final checkpoints, $\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{recon}} + \gamma|\mathrm{KL} - C(t)|$, with $C(t)$ warmed up early in training; the logs also record β, reconstruction MSE/L1, KL, learning-rate reductions, and early stopping. This is why the appendix reports VAE reconstruction memorization before any diffusion result: if encode/decode alone cannot preserve sample identity, a pixel-space memorization test on generated samples is not yet meaningful.

Real-data latent diffusion. After selecting the best VAE checkpoint by validation objective, we encode the same farthest-first diverse 1k subset for each dataset and train a latent-space MLP score model on those codes. The final reported runs use a depth-five MLP with hidden width 1024, sinusoidal time features, SGD with learning rate 10^{-3} and momentum 0.80, batch size 512, and a 5M-step horizon over five seeds. The hidden width is fixed across d_{latent}, so increasing the latent dimension changes the input/output dimension and the amount of latent buffer, but not the dominant hidden-layer capacity. Training loss is logged every 10k steps; memorization and FID are evaluated every 100k steps using 10k generated samples, with the same 1k training subset as the FID reference.

Spatial-DiT pilot. The appendix also includes a CelebA pilot that replaces the vector VAE/MLP pipeline with a spatial 4×4 VAE and an unconditional DiT-style score model. In that setting the plotted d_{latent} is a channel count, so a single step in d_{latent} changes the latent state by 16 scalar coordinates. This makes the sweep harder to tune and interpret than the vector-latent MLP experiments: changing the channel count changes the scalar latent dimension, the VAE decoder, the first and last DiT projections, and the second-order cross-position/cross-channel covariance structure that the transformer sees. The DiT pilot therefore tests whether the buffer idea leaves a trace in a more realistic image architecture, not whether the clean vector-latent theory transfers without modification.

Optimization failures and clipped diagnostics. Nonfinite runs are treated as failed optimization runs, not as evidence of zero memorization. Gradient clipping was useful diagnostically because it prevented some high-dimensional CIFAR-10 runs from producing NaNs, but it also suppressed memorization almost entirely. The clean CIFAR-10 aggregate therefore uses unclipped replacement seeds with the same SGD hyperparameters rather than mixing clipped and unclipped regimes.

J RFNN Spectral Diagnostics

The RFNN experiments are spectral diagnostics, not sampled diffusion runs. They therefore do not have an image-level memorization percentage. Their purpose is to check whether the kernel spectrum separates into the predicted signal, noise-dimension, sample, and tail index groups under the same data choices used by the trainable experiments.

Experiment 2, the source-dimensionality sweep, uses the updated synthetic parameters from the final rerun: d_{latent} is fixed at 20, $d_{\mathrm{intrinsic}}$ ranges over $\{2, 5, 8, 12, 16, 20, 25, 30, 35, 40\}$, and the whole sweep is repeated under $p = 64d_{\mathrm{latent}}$, $p = d_{\mathrm{latent}} + n + 300$, and fixed-$p$ width controls. This experiment is the reverse of Experiment 1, the main d_{latent} sweep. It asks whether the noise-dimension block shrinks when more of the fixed observed latent space is occupied by true source

variation, and whether over-intrinsic source structure merely compresses into the fixed observed space rather than creating additional observed signal modes.

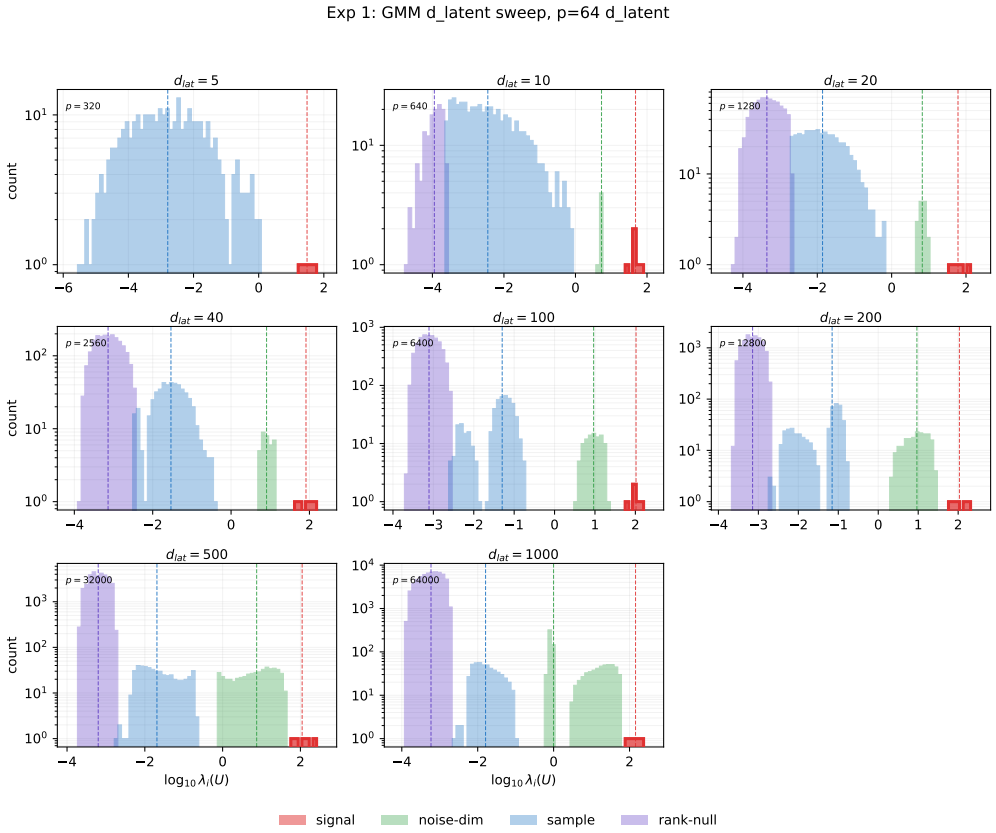


Figure 17: Experiment 1 RFNN four-bulk spectra for the latent-dimension sweep with $p = 64d_{\text{latent}}$. Here $d_{\text{intrinsic}} = 5$, $\sigma_{\text{signal}} = 1.0$, $\sigma_{\perp} = 0.5$, GMM center scale is 3.0, and the swept variable is $d_{\text{latent}} \in \{5, 10, 20, 40, 100, 200, 500, 1000\}$. Each panel plots the eigenvalue histogram after assigning eigenmodes to the signal, noise-dimension, sample, and tail index groups. The panels with $d_{\text{latent}} \geq n$ lie outside the $p > n > d$ regime, where an $n-d$ sample-group count applies. This highly overcomplete feature setting checks that the expected spectral ordering is visible when the RFNN has many more features as d_{latent} grows. Since the synthetic $d_{\text{intrinsic}} = 5$ is known exactly, the green block grows in the exact increasing-buffer regime $d_{\text{latent}} - d_{\text{intrinsic}}$, with little irregular variation beyond the deterministic change in block size. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

Feature-scale caveat for RFNN spectra. The nonlinear feature map is $\tanh$, whose response depends on the input scale and operating point q. The location and separation of the colored index groups therefore depend on the signal and null variances, not only their dimensions. In the separate $\sigma_{\perp} = 0.01$ runs, the noise-dimension/sample gap nearly closes. Four visibly separated components are not a universal consequence of adding coordinates.

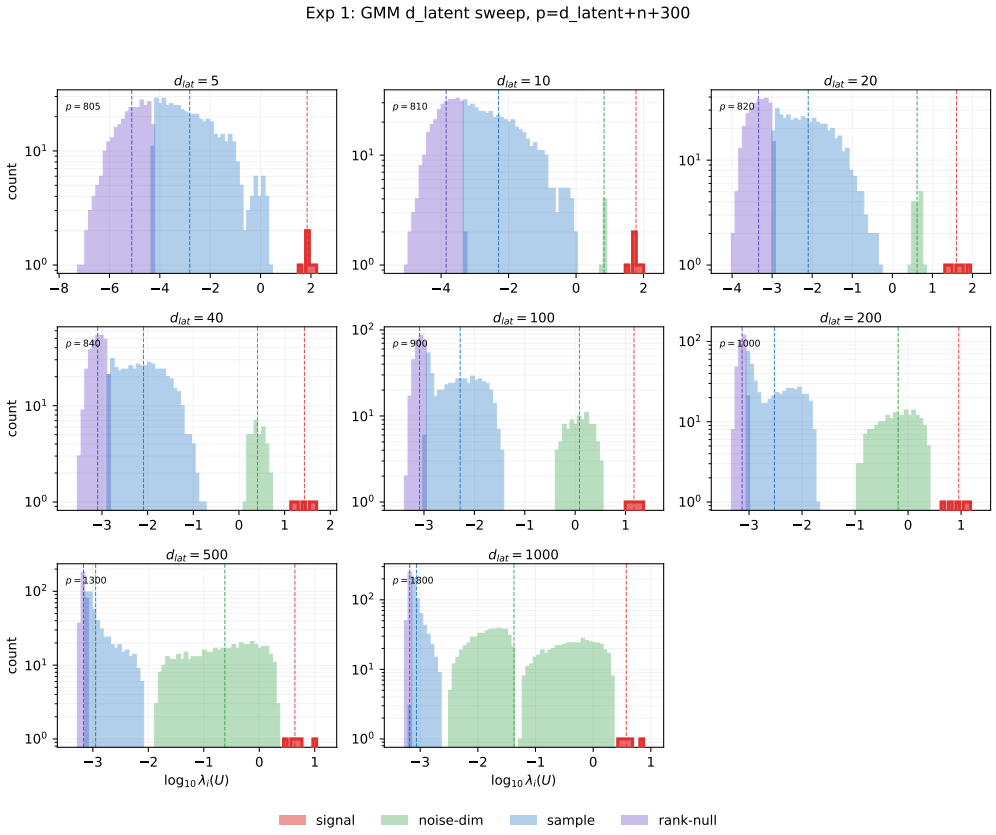


Figure 18: Experiment 1 RFNN four-bulk spectra for the latent-dimension sweep with $p = d_{\text{latent}} + n + 300$. The data sweep is the same as Figure 17: fixed $d_{\text{intrinsic}} = 5$, fixed $\sigma_\perp = 0.5$, and increasing d_{latent}. This width grows just enough to represent the latent, sample, and tail index groups, so it checks that the four-bulk ordering does not require the much wider $p = 64 d_{\text{latent}}$ feature map. Separation remains visible in parts of the sweep, but both the aspect ratios and the nonlinear feature operating point change with d_{latent}. The same near-monotone growth of the green block with the exact buffer size $d_{\text{latent}} - d_{\text{intrinsic}}$ appears even in this controlled-width setting. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

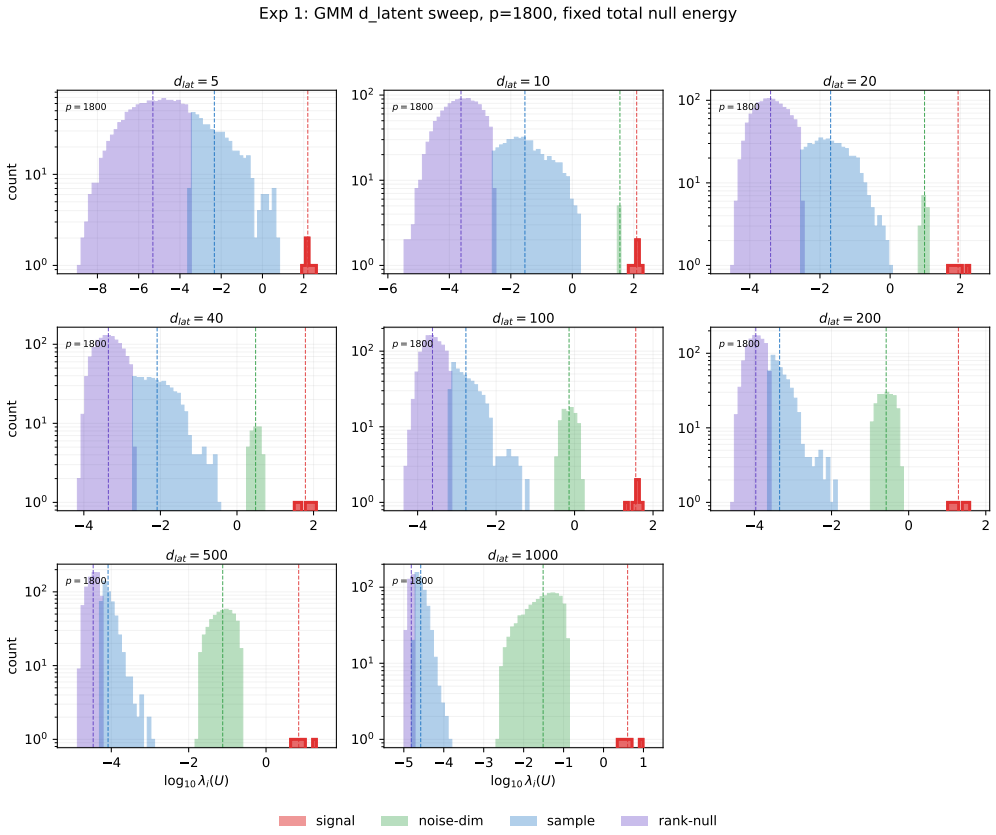


Figure 19: Experiment 1 RFNN four-bulk spectra for the latent-dimension sweep with fixed $p = 1800$ and fixed total null energy. The swept variable is still d_{latent}, but $\sigma_\perp^2$ is rescaled as $E_{\text{null}}/(d_{\text{latent}} - d_{\text{intrinsic}})$ when $d_{\text{latent}} > d_{\text{intrinsic}}$, so that the total null variance does not grow with latent dimension. This is the strictest control because increasing d_{latent} no longer increases either the number of random features or the total variance injected into null coordinates. This control does not fix the feature operating point q: distributing the same null energy over more coordinates changes the input scale entering $\tanh$. The remaining separation therefore does not isolate an explanation based only on mode counts. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

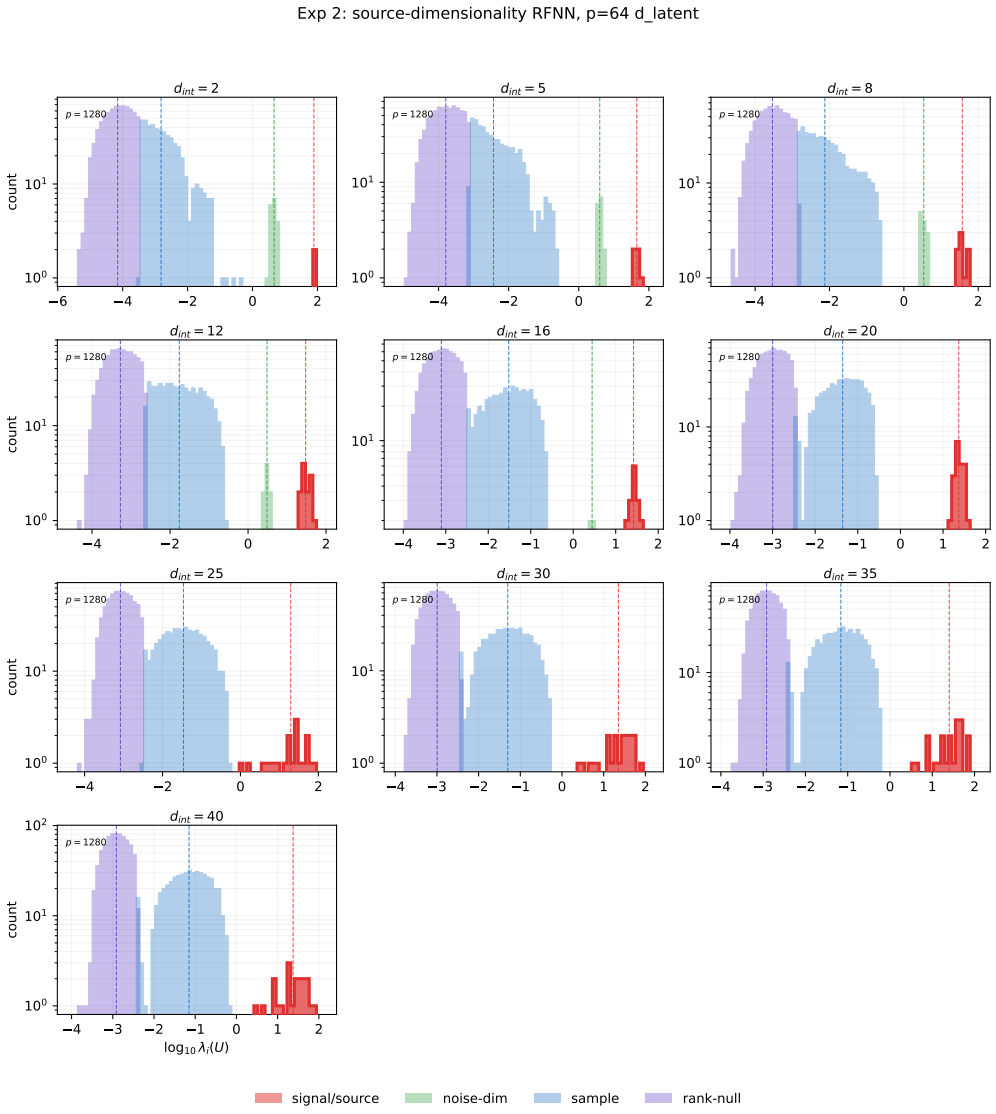


Figure 20: Experiment 2 RFNN source-dimensionality sweep with a fixed observed latent space under the $p = 64d_{\text{latent}}$ width rule. The learned latent dimension is held fixed at $d_{\text{latent}} = 20$, $\sigma_{\text{signal}} = 1.0$, $\sigma_{\perp} = 0.5$, and GMM center scale 3.0, while the intrinsic dimension of the source distribution sweeps $d_{\text{intrinsic}} \in \{2, 5, 8, 12, 16, 20, 25, 30, 35, 40\}$. When $d_{\text{intrinsic}} < d_{\text{latent}}$, the remaining observed coordinates form a genuine null buffer; as $d_{\text{intrinsic}}$ approaches d_{latent}, that buffer shrinks and the signal and sample regions become less separated. Because $d_{\text{latent}} = 20$ is fixed and known, this is the exact decreasing-buffer regime: the green block shrinks as $d_{\text{latent}} - d_{\text{intrinsic}}$ and disappears at $d_{\text{intrinsic}} = d_{\text{latent}}$. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

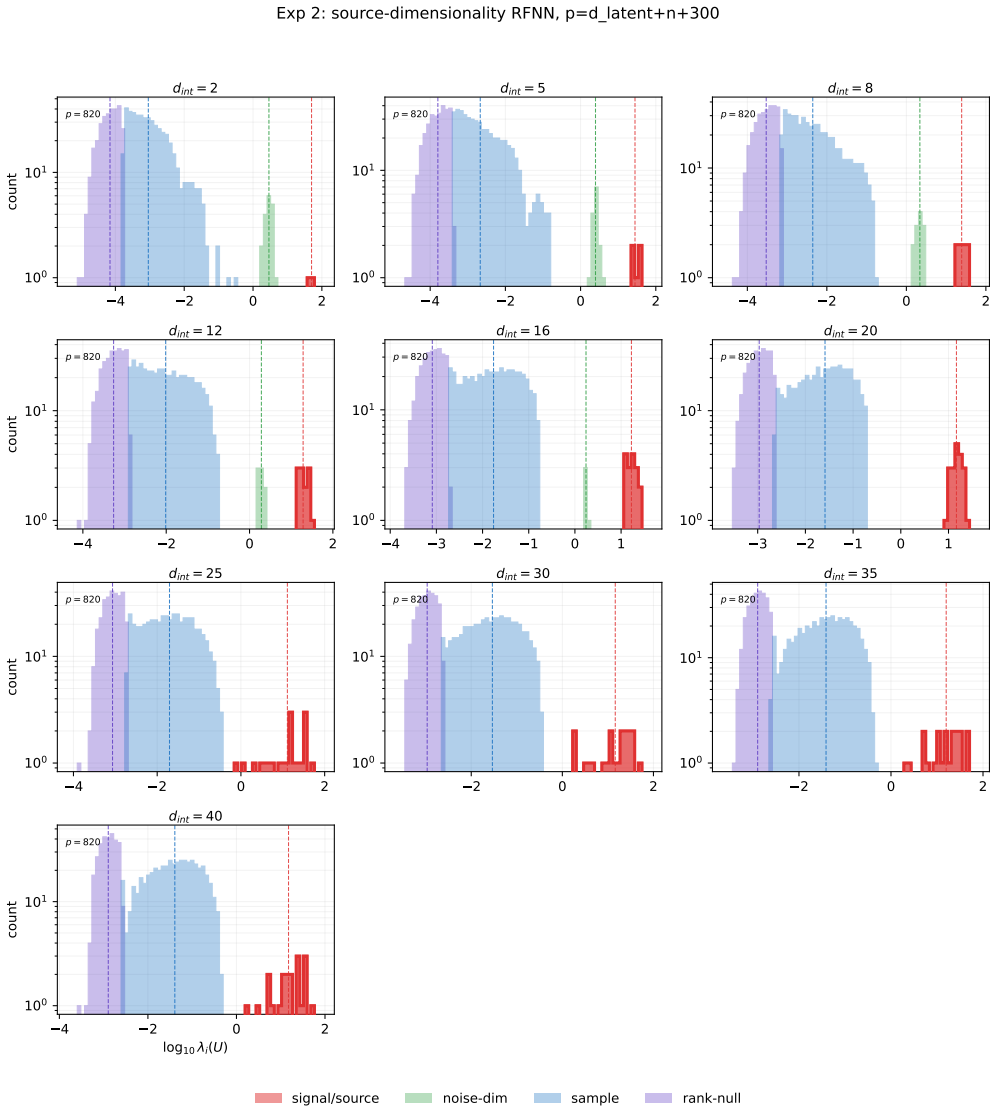


Figure 21: Experiment 2 RFNN source-dimensionality sweep with $p = d_{\text{latent}} + n + 300$. This panel uses the same fixed observed latent space as Figure 20 and the same $d_{\text{intrinsic}}$ sweep, but removes the highly overcomplete feature-count scaling. The same loss of separation as $d_{\text{intrinsic}}$ approaches d_{latent} shows that the effect is not specific to the widest RFNN parameterization. The smooth decrease of the null-buffer block with the exact gap $d_{\text{latent}} - d_{\text{intrinsic}}$ is the key diagnostic. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

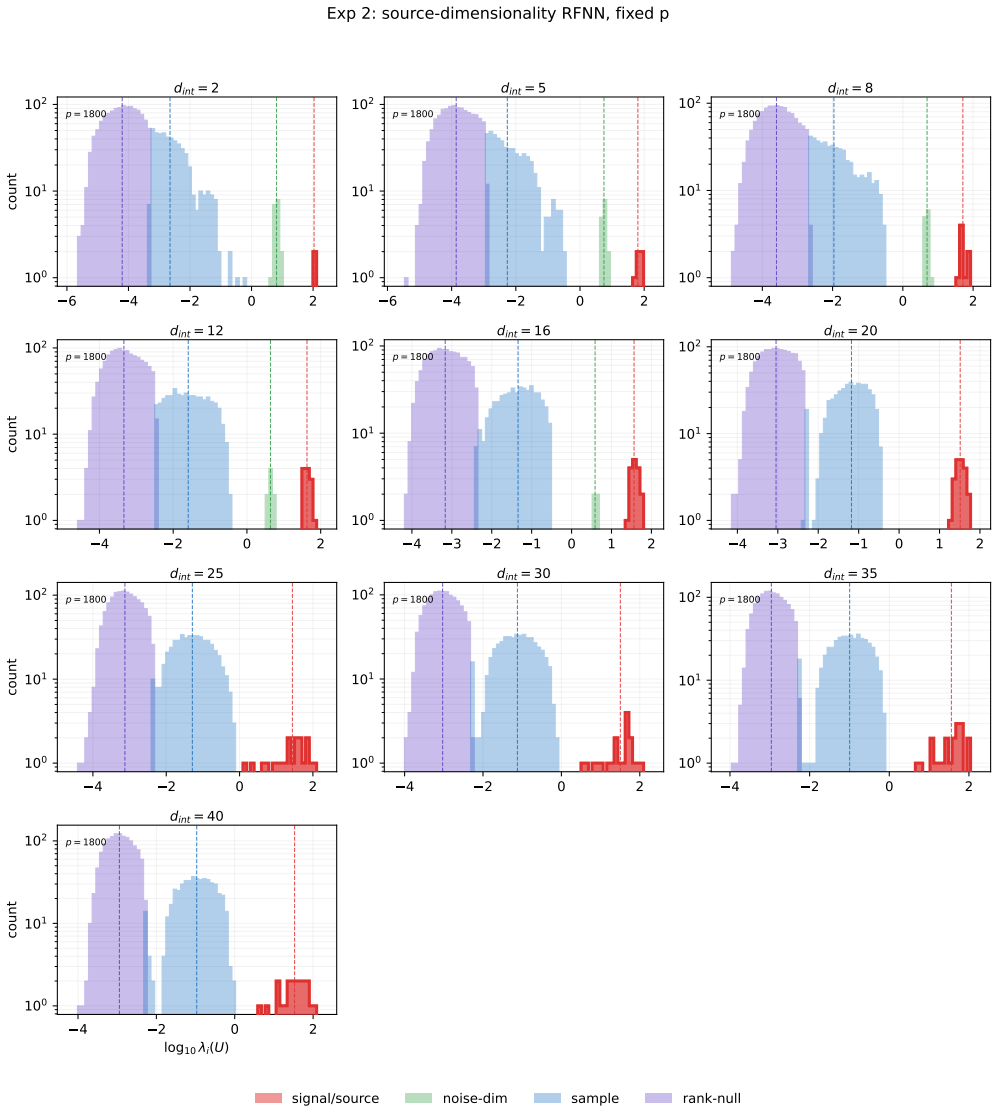


Figure 22: Experiment 2 RFNN source-dimensionality sweep with fixed width. For this control $d_{\text{latent}} = 20$ is fixed and $d_{\text{intrinsic}}$ is swept as in the other Experiment 2 panels, while RFNN width is held fixed instead of growing with d_{latent} or $d_{\text{intrinsic}}$. For $d_{\text{intrinsic}} > d_{\text{latent}}$, the source structure is compressed into the fixed observed latent space, so the high-$d_{\text{intrinsic}}$ regime should be read as an over-intrinsic stress test rather than a literal null-buffer regime. The fixed-width control makes the narrowing of the buffer visible without simultaneously changing RFNN capacity. The relevant regime is therefore decreasing until the exact boundary $d_{\text{intrinsic}} = d_{\text{latent}}$, followed by compressed over-intrinsic behavior rather than a new buffer. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

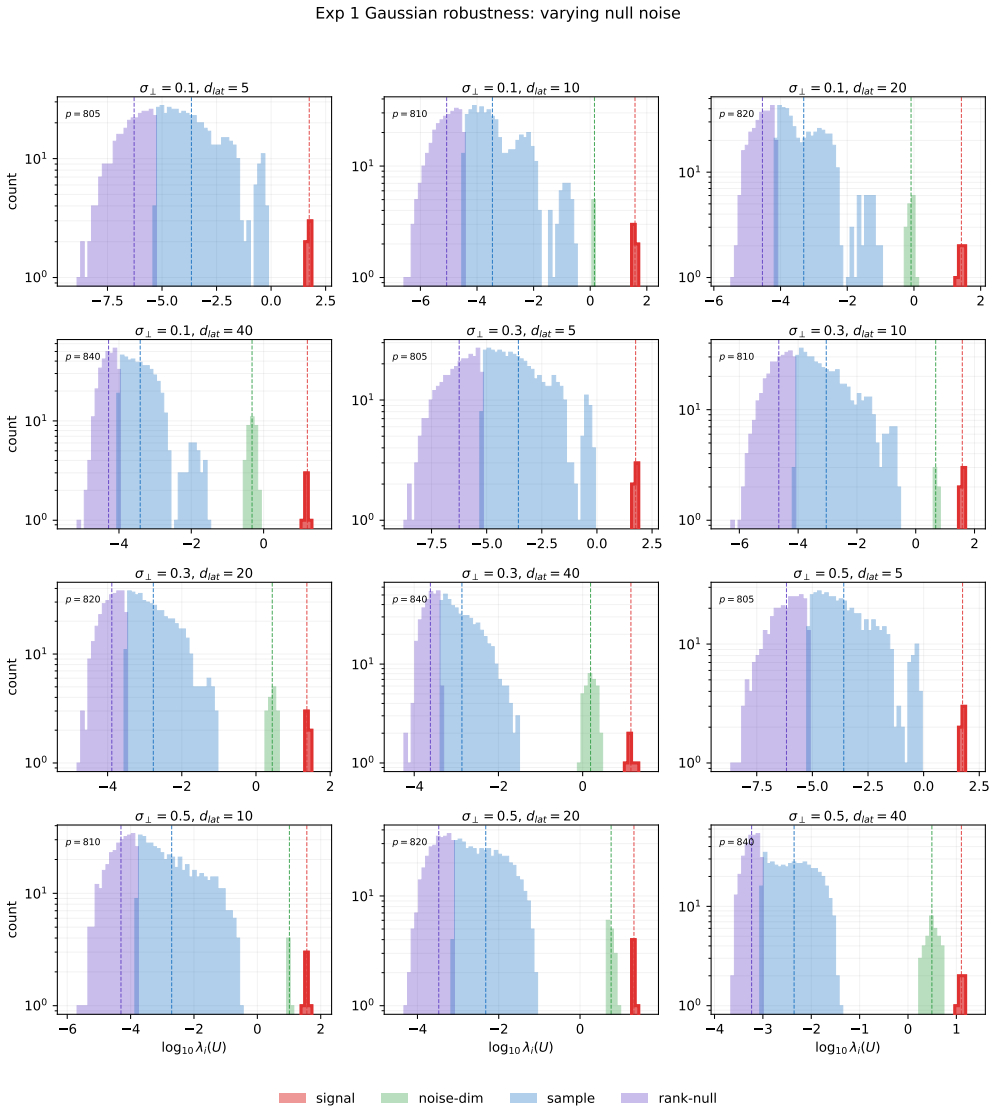


Figure 23: Experiment 1 RFNN latent-dimension sweep on pure anisotropic Gaussian data while varying the null-noise scale. Rows use $\sigma_\perp \in \{0.1, 0.3, 0.5\}$ at fixed $d_{\text{intrinsic}} = 5$, $\sigma_{\text{signal}} = 1.0$, and swept d_{latent}. This test removes mixture structure from the synthetic data and checks that the four-bulk decomposition is not an artifact of the default GMM. Increasing null noise shifts the latent-null region and changes the gap to the sample region, as expected from the buffer interpretation. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

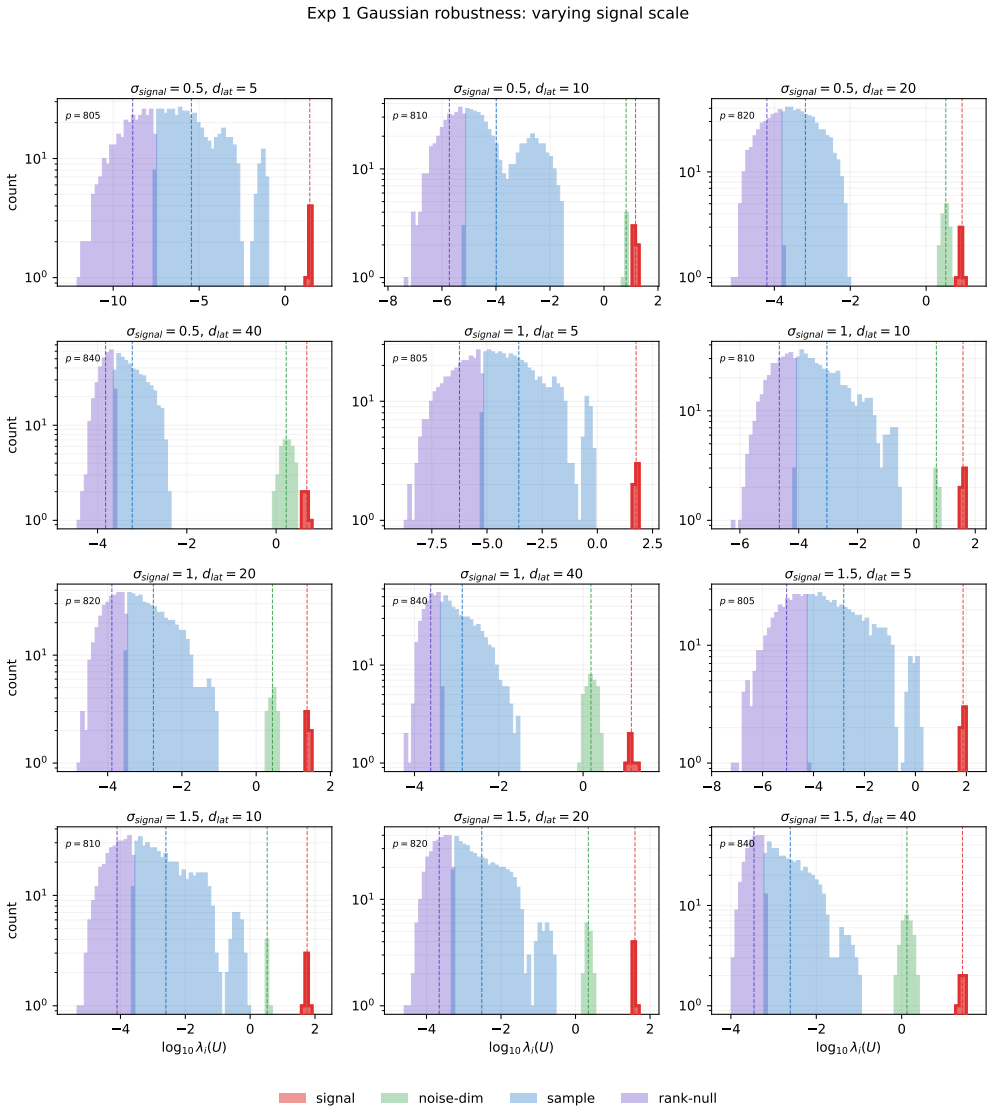


Figure 24: Experiment 1 RFNN latent-dimension sweep on pure anisotropic Gaussian data while varying the signal scale. Rows use $\sigma_{\text{signal}} \in \{0.5, 1.0, 1.5\}$ at fixed $d_{\text{intrinsic}} = 5$, $\sigma_{\perp} = 0.3$, and swept d_{latent}. The signal block moves with the strength of the structured coordinates, while the sample and null regions remain governed by the latent buffer and noise statistics. This confirms that the visible bulk ordering responds to the intended data-level parameters. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

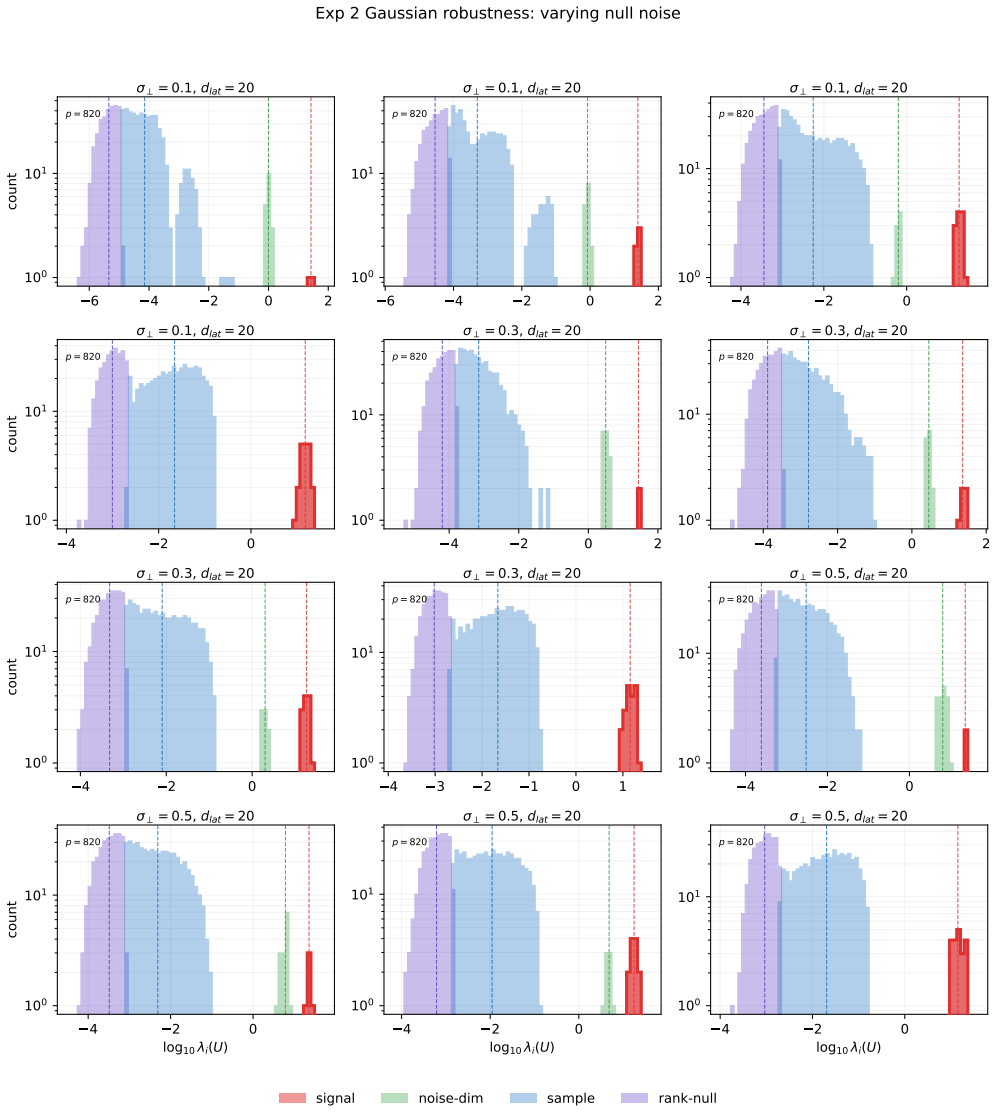


Figure 25: Experiment 2 RFNN source-dimensionality sweep on pure anisotropic Gaussian data while varying the null-noise scale. Rows use $\sigma_\perp \in \{0.1, 0.3, 0.5\}$ at fixed $d_{\text{latent}} = 20$ and swept $d_{\text{intrinsic}}$. The same buffer-collapse pattern appears when $d_{\text{intrinsic}}$ approaches the fixed observed latent dimension, and the amount of separation between null and sample regions changes with the null-noise scale. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

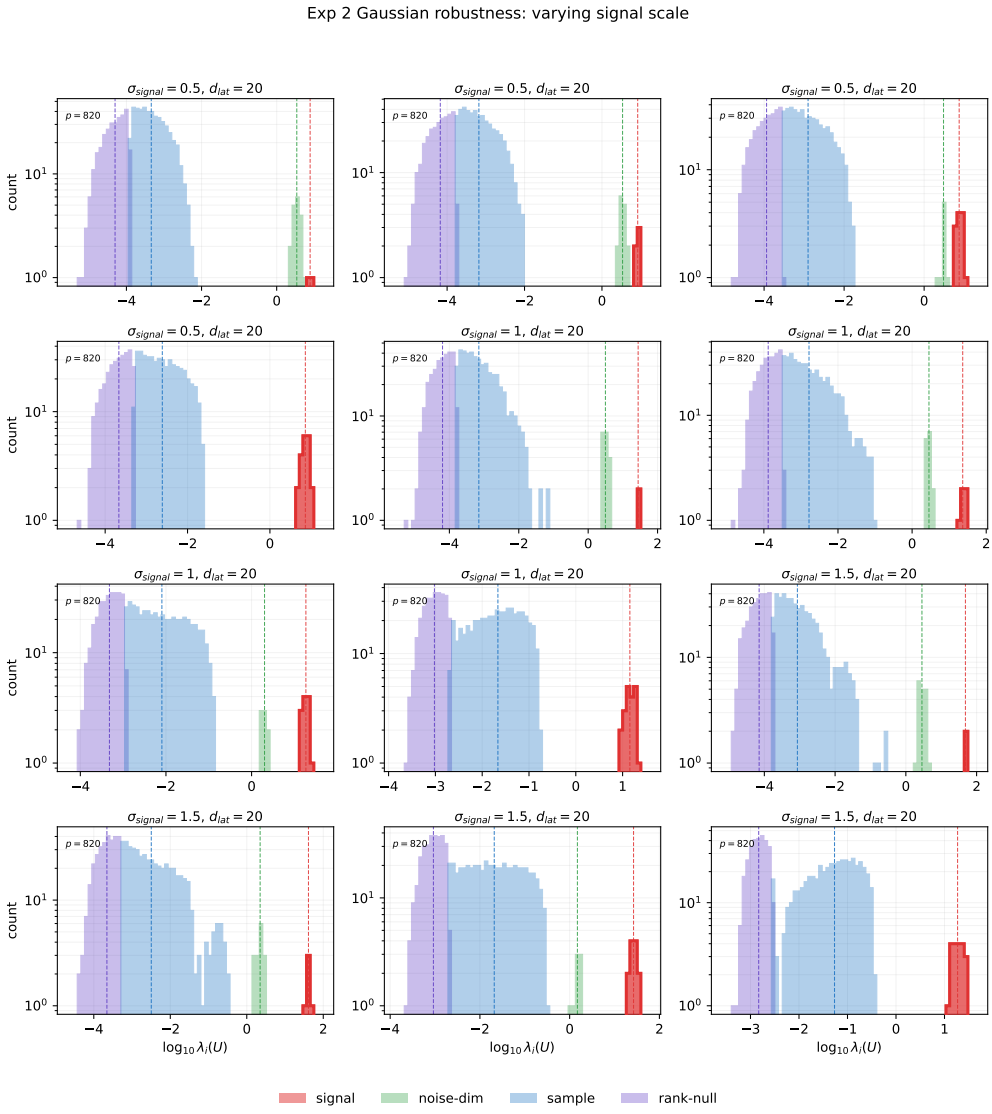


Figure 26: Experiment 2 RFNN source-dimensionality sweep on pure anisotropic Gaussian data while varying the signal scale. Rows use $\sigma_{\text{signal}} \in \{0.5, 1.0, 1.5\}$ at fixed $d_{\text{latent}} = 20$, $\sigma_{\perp} = 0.3$, and swept $d_{\text{intrinsic}}$. Larger signal scale sharpens the leading signal region, but the qualitative transition as $d_{\text{intrinsic}}$ approaches d_{latent} remains governed by the amount of available null-buffer dimension. Colors are assigned by sorted eigenvalue-index blocks rather than fitted clusters: red denotes signal/source modes, green the observed latent-null/noise-dimension block, blue sample-specific modes, and purple tail indices. These index assignments do not establish eigenvector alignment. Noise averaging can make the tail positive. Dashed vertical lines mark the median log-eigenvalue of each colored block.

K SYNTHETIC MLP DIAGNOSTICS

The synthetic MLP figures use five seeds, a 5M-step training horizon, and the nearest-neighbor memorization ratio described in Appendix I. This appendix repeats the main fixed-width sweep with the larger latent range $d_{\text{latent}} \in \{5, 8, 10, 12, 15, 20, 25, 30, 35, 40, 60, 80, 100, 120, 140, 160, 180, 200, 220, 240\}$. We report memorization percentage over training steps so that the onset, growth, and saturation of memorization remain visible.

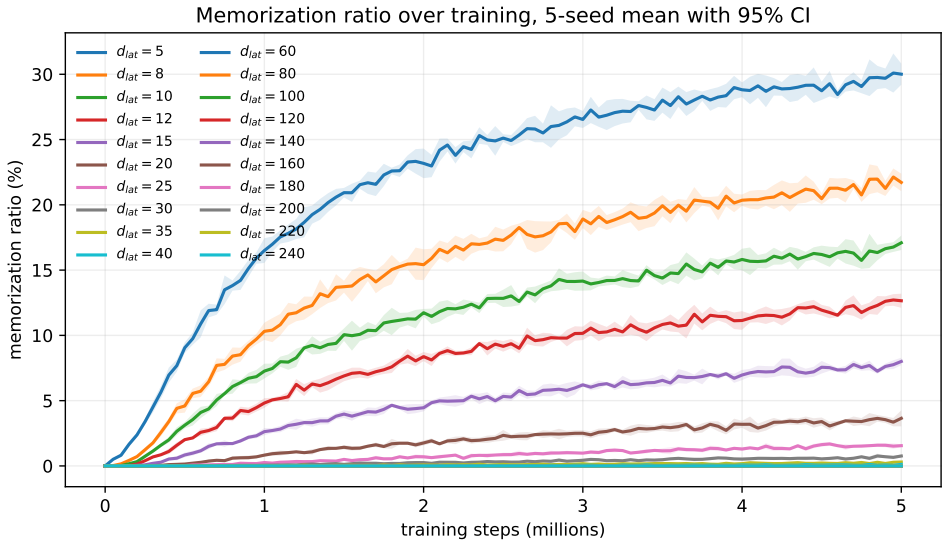
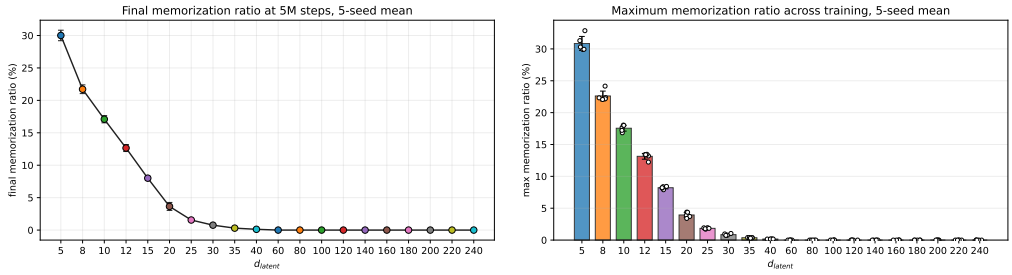


Figure 27: Synthetic trainable MLP memorization trajectories across latent dimension. The curves show when each model begins to reproduce individual training samples according to the nearest-neighbor ratio, while retaining the full shape of the rise rather than collapsing it to a single threshold. Larger latent dimensions show a later rise in memorization. This observation does not establish sequential absorption of feature groups: all modes of the frozen-feature model learn in parallel, and the trainable MLP can change its features. The extended sweep shows that this continues through $d_{\text{latent}} = 240$ under the fixed-width MLP control. The sweep is almost ordered by the exact synthetic buffer size $d_{\text{latent}} - d_{\text{intrinsic}}$: after the no-buffer $d_{\text{latent}} = d_{\text{intrinsic}} = 5$ case, increasing d_{latent} produces a steadily lower and later memorization regime rather than large dimension-to-dimension variation.

Interpreting threshold-free memorization curves. Single hitting times such as τ_{mem} are convenient summaries, but they hide whether a run crosses a threshold sharply, gradually, or only transiently. The full memorization-fraction curves are therefore the primary diagnostic for the trainable MLP experiments. Thresholded hitting times are used later only in the spectral-predictor analysis, where multiple fixed memorization levels are compared against predicted iso-memorization times.



(a) Final memorization at 5M steps.

(b) Peak memorization during training.

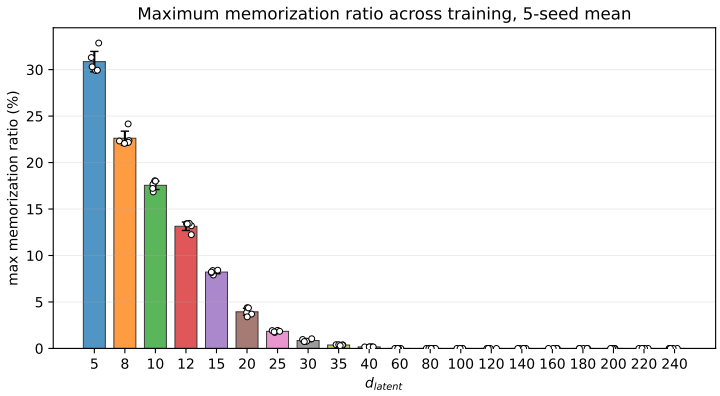


Figure 28: Synthetic MLP final and peak memorization summaries. The final memorization fraction measures the end-of-training state, while the peak memorization fraction captures transient sample recall that may not persist until 5M steps. The seed-level histogram exposes how much of the latent dimensional trend is stable across random seeds rather than being driven by a single unusually memorizing run. The final and peak summaries both support the same decreasing regime with increasing d_{latent}, including the high-dimensional extension beyond the main-text $d_{\text{latent}} \leq 40$ view.

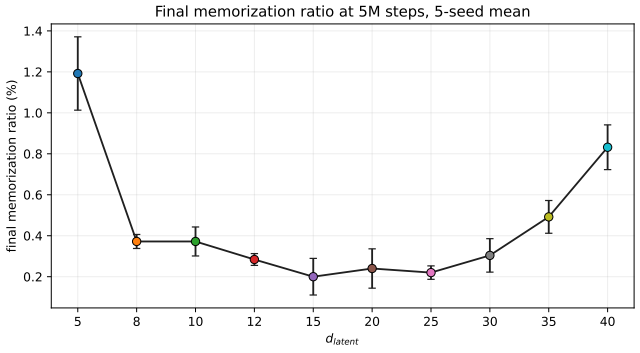
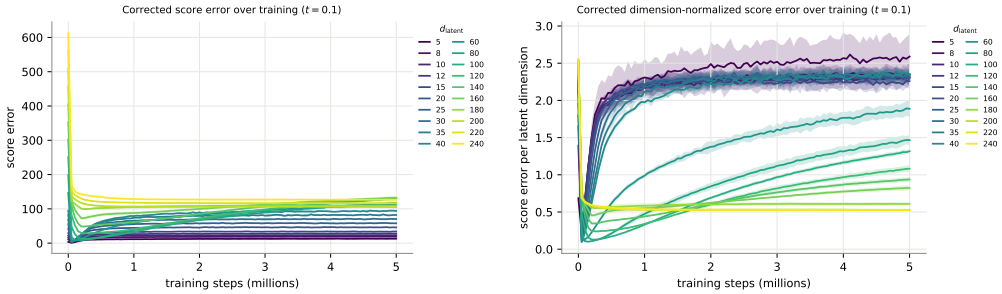


Figure 29: Width-scaled control for Figure 1: the same sweep with hidden width $8d_{\text{latent}}$ (40 to 320) instead of a fixed 256, five seeds, 5M steps. Memorization sits at the 1% level and rises again for $d_{\text{latent}} \geq 30$ as the network grows, so a width-scaled sweep is a capacity sweep and does not isolate the latent dimension. This is why the main text holds the width fixed.

Table 8: Dimension-controlled memorization test for the fixed-width synthetic MLP (5M steps, five seeds, 5000 samples each; same configuration as Figure 1). The nearest-neighbor ratio test of Appendix I is applied in the full latent space and after projecting generated and training points onto the true $d_{\text{intrinsic}} = 5$ signal coordinates. Controls: fresh draws from the population; exact copies of a training point plus the sampler's terminal noise in every coordinate; and signal copies, which add the same terminal noise to a training point's signal coordinates and redraw its null coordinates. The added terminal-noise variance is 0.0269 per coordinate. α is the projection coefficient of the learned score onto the exact empirical score of the training set in the signal coordinates at $t = 0.1$ (0: population score; 1: exact fit of the training set).

	$d_{\text{latent}} = 5$	10	20	40
model samples, full-space test	29.9%	16.9%	3.6%	0.1%
model samples, signal-projected test	29.9%	16.8%	5.8%	1.2%
fresh draws, signal-projected test	0.4%	0.4%	0.4%	0.4%
exact copies, full-space test	57%	69%	84%	97%
signal copies, full-space test	57%	0.8%	0.0%	0.0%
signal copies, signal-projected test	57%	57%	58%	58%
α (signal coordinates), 0.25M steps	0.54	0.37	0.22	0.09
α (signal coordinates), 5M steps	0.75	0.54	0.39	0.31



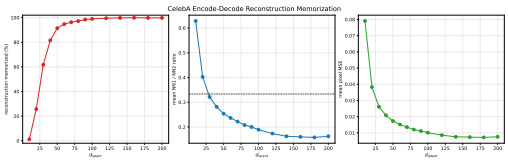
(a) Score error, all coordinates.

(b) Score error normalized per latent dimension.

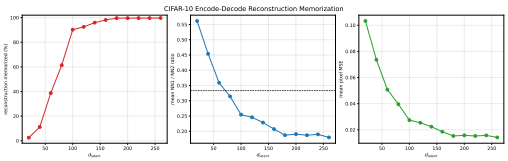
Figure 30: Synthetic MLP score-error diagnostics: population score error at $t = 0.1$ recovered from saved held-out denoising losses using the corrected analytic GMM score, five-seed mean with 95% CI, over the complete fixed-width sweep $d_{\text{latent}} = 5$–240 (the main-text panel of Figure 1 covers $d_{\text{latent}} \leq 40$). The left panel sums over all latent coordinates; the right panel divides by d_{latent}, which removes the mechanical scaling with d_{latent}. This distinction matters because a larger latent space has more coordinates to predict even if the average coordinate-wise error is comparable. Score error is the appropriate quality metric in this synthetic setting because the target distribution is specified analytically in latent space rather than through decoded images. Beyond $d_{\text{latent}} \approx 60$ the per-coordinate error has not returned to its small-d_{latent} plateau by the 5M-step horizon: it ends at 2.59 for $d_{\text{latent}} = 5$ but 0.53 for $d_{\text{latent}} = 240$, so the widest latents have not reached the small-dimension plateau within this horizon. The DSM recovery is exact in expectation; its finite evaluation sample leaves a cross term, so these curves are recovered estimates rather than direct checkpoint score evaluations.

L REAL-DATA VAE DIAGNOSTICS

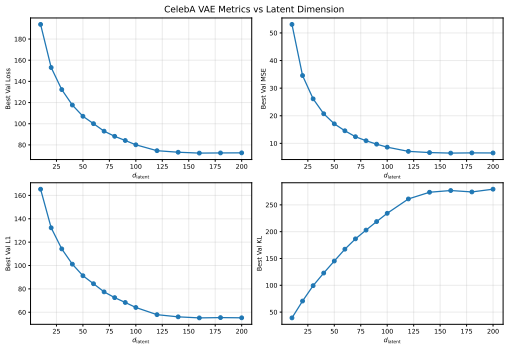
The real-data experiments separate two questions. First, the VAE must preserve enough image identity that nearest-neighbor memorization can be detected after encoding and decoding. Second, the diffusion model must have enough capacity and stable optimization to actually memorize the encoded subset. For this reason, the VAE figures appear before the diffusion figures: they identify widths where reconstruction loses sample identity. This operational threshold depends on the VAE and metric and does not determine the intrinsic dimension of the image manifold.



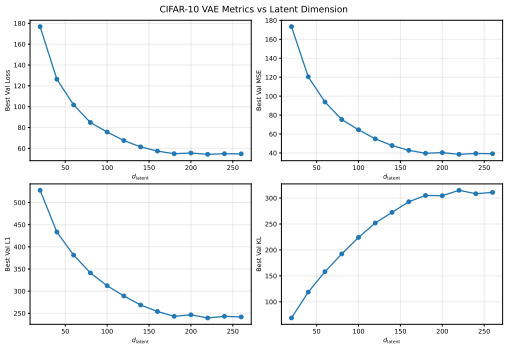
(a) CelebA reconstruction memorization.



(b) CIFAR-10 reconstruction memorization.



(c) CelebA VAE loss terms.



(d) CIFAR-10 VAE loss terms.

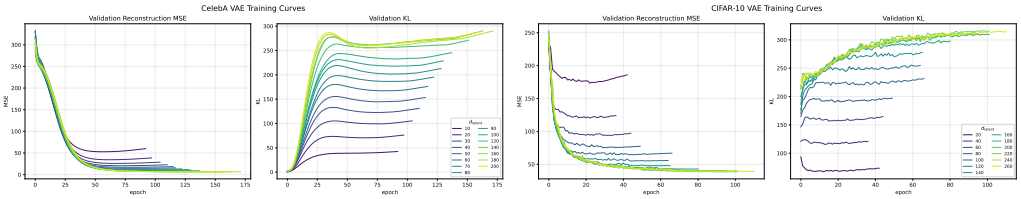
Figure 31: VAE diagnostics across latent dimension for CelebA and CIFAR-10. The reconstruction-memorization panels encode the 1k diffusion subset through the frozen VAE and decode it immediately, then apply the same pixel-space KNN ratio used for generated samples. This measures whether the VAE bottleneck itself preserves enough sample identity for memorization to be detectable. The loss panels show the corresponding validation reconstruction and KL behavior across dimensions for VAEs trained on the full training split. Low-dimensional VAEs blur or average away identity, so weak diffusion memorization at small d_{latent} can reflect a lossy representation rather than an intrinsically nonmemorizing diffusion model. The operational 95% reconstruction-identity thresholds are near $d_{\text{latent}} = 70$ and 140. They do not establish intrinsic manifold dimension, and reconstruction changes remain a potential confound above them.



Figure 32: CelebA VAE encode/decode contact sheet. The rows show original images and their final checkpoint reconstructions across latent dimensions, making visible the point at which identity-level facial features become stable under the VAE bottleneck. The contact sheet shows how the preservation of sample identity changes with VAE width; it is not an intrinsic-dimension estimate. This qualitative check is necessary because the downstream KNN memorization metric can only detect sample recall when reconstructions preserve enough pixel-level identity.



Figure 33: CIFAR-10 VAE encode/decode contact sheet. The rows show original images and final checkpoint reconstructions. CIFAR-10 has much more class and texture variation than the CelebA subset, so visual reconstruction quality is an important check before interpreting downstream diffusion memorization. The higher-dimensional VAEs preserve more object identity and local structure, making pixel-space nearest-neighbor tests more meaningful. As with CelebA, these checks characterize the representation and its operational identity threshold rather than manifold intrinsic dimension.

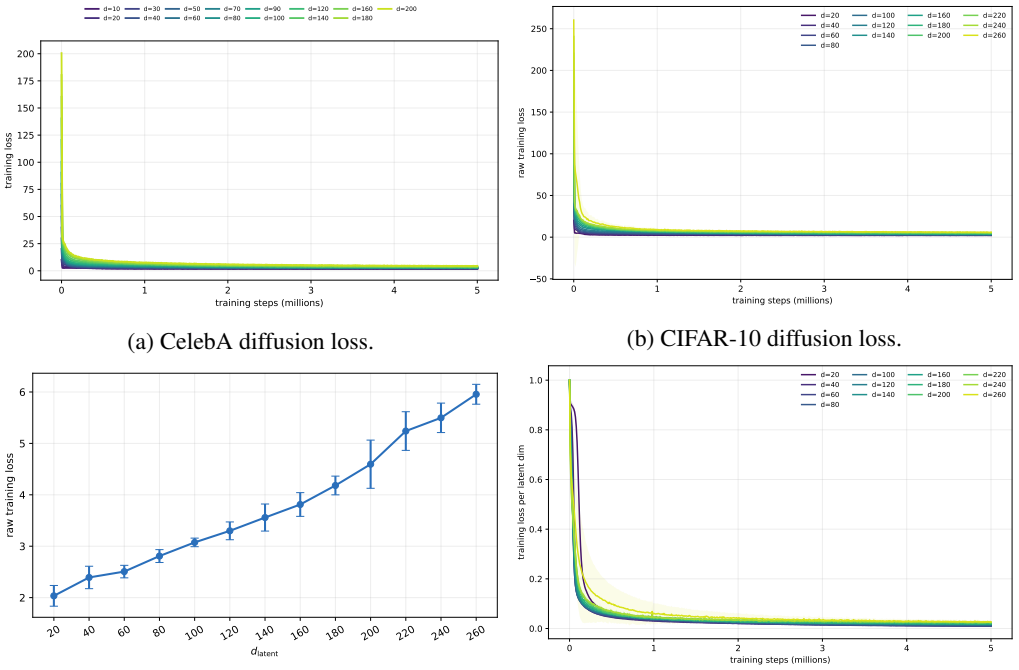


(a) CelebA VAE training curves.

(b) CIFAR-10 VAE training curves.

Figure 34: VAE training curves for CelebA and CIFAR-10. The reported VAE checkpoints are selected by validation objective after the reconstruction and KL objectives have largely stabilized, so the downstream latent-diffusion trends are not simply caused by comparing undertrained encoders at different latent dimensions. Together with Figures 31–33, these diagnostics justify treating high-dimensional VAE latents as a meaningful substrate for pixel-space memorization tests.

M Real-Data Diffusion Diagnostics



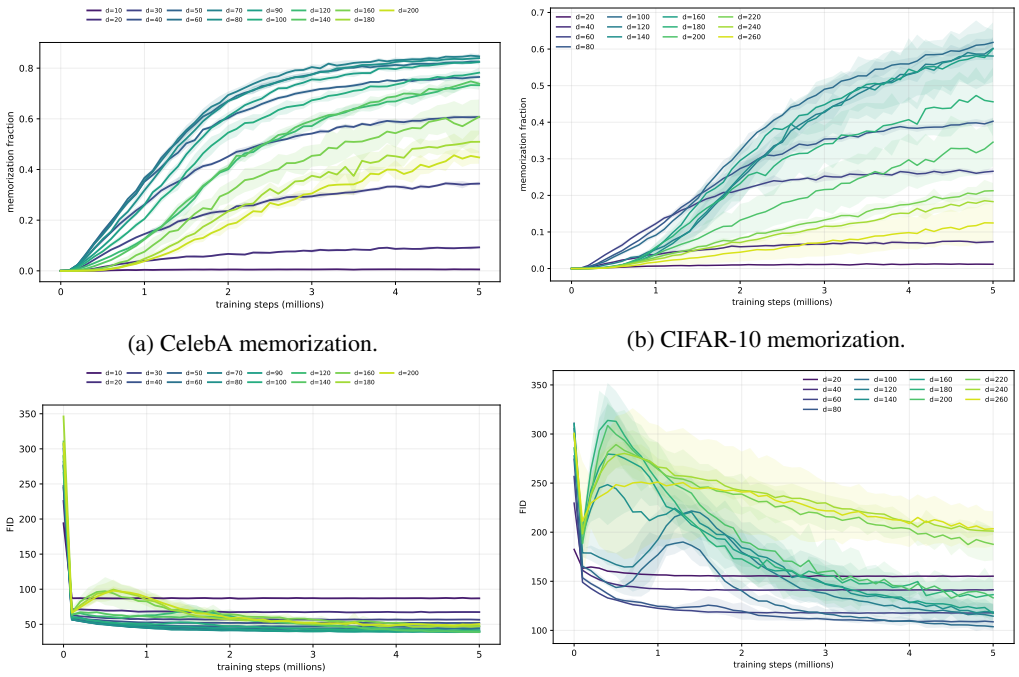
(a) CelebA diffusion loss.

(b) CIFAR-10 diffusion loss.

(c) CIFAR-10 final loss by d_{latent}.

(d) CIFAR-10 score loss per dimension.

Figure 35: Optimization diagnostics for the real-data diffusion runs, shown before the memorization and FID panels because they establish that the models are actually training. The loss curves track whether differences in memorization and FID coincide with gross optimization failure, while the CIFAR-10 per-dimension score loss removes the trivial scaling of the objective with latent dimension. Gradient clipping was used only as a diagnostic intervention: it prevented some unstable runs from diverging, but it also strongly suppressed memorization and therefore changed the effective training regime. The aggregate CIFAR-10 results are reported using stable unclipped runs so that optimizer changes are not conflated with latent-dimensional effects.



(a) CelebA memorization.

(b) CIFAR-10 memorization.

(c) CelebA FID.

(d) CIFAR-10 FID.

Figure 36: Real-data diffusion trajectories with five-seed confidence bands. The memorization panels report the fraction of generated samples that pass the pixel-space nearest-neighbor ratio against the diffusion training subset. The FID panels report image-quality evolution over the same optimization horizon. Reading the two rows together separates quality improvement from direct sample recall: a model can improve FID before it begins to reproduce exact training examples, and a high-dimensional latent can delay memorization even when the generated samples are already becoming visually plausible. After the VAE identity threshold, the broad trend is toward later and lower memorization at larger widths, although CIFAR-10 hitting times are not strictly monotone. FID improves at intermediate widths and worsens at the widest latents; this does not identify the cause of the quality loss.

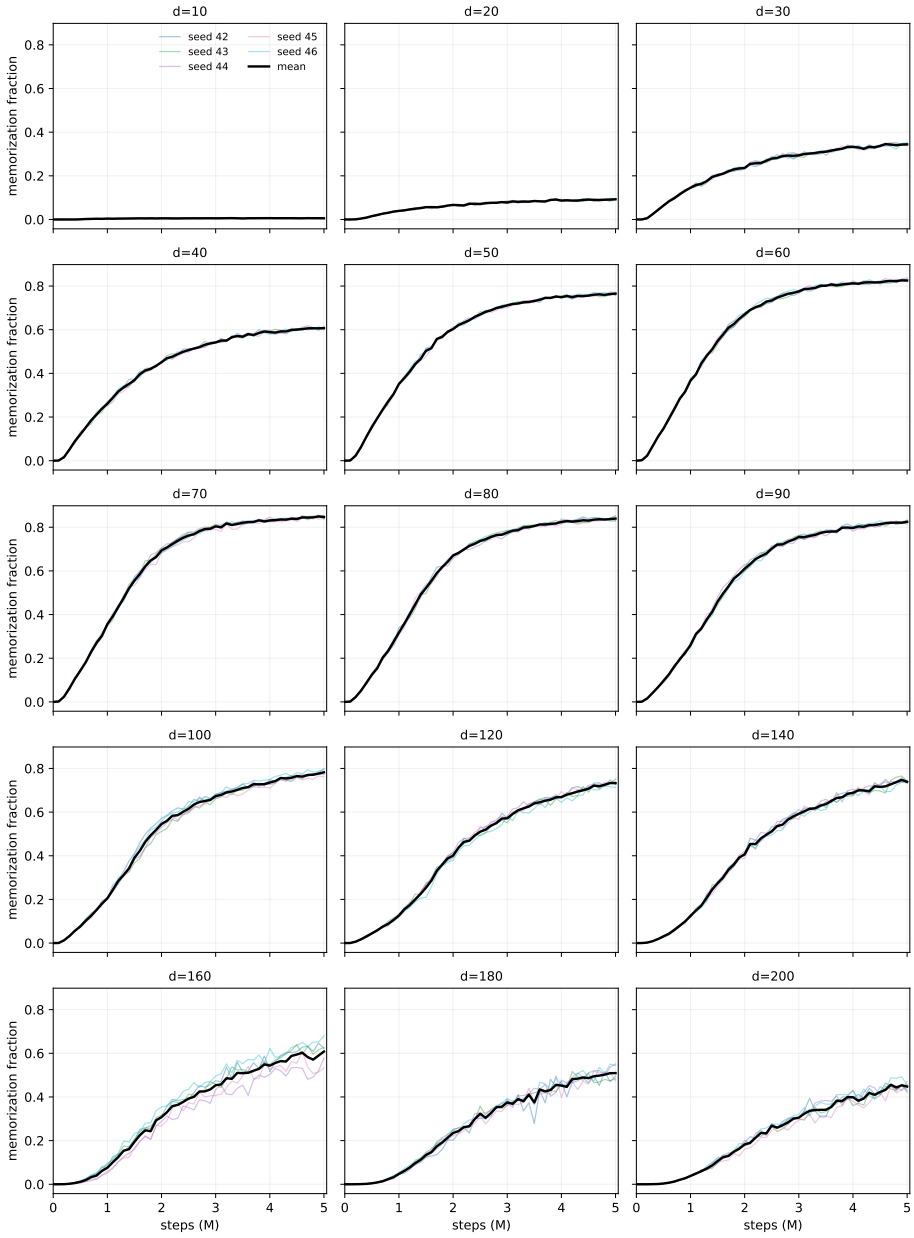


Figure 37: CelebA per-dimension memorization trajectories. Each panel isolates one latent dimension and shows the seed-level evolution of KNN memorization over the 5M-step run. This view is important because the aggregate curve can hide whether a dimension memorizes consistently across seeds or only through a few high-memorization runs. At very small d_{latent}, the simple latent-buffer prediction need not apply cleanly: reconstruction loss may prevent the decoder from preserving enough sample-specific detail for the pixel-space test. This can also contribute to low-width errors in the calibrated spectral fit. Once past that bottleneck, the panels show a largely ordered increasing-then-decreasing memorization regime rather than unsystematic variation across adjacent latent dimensions.

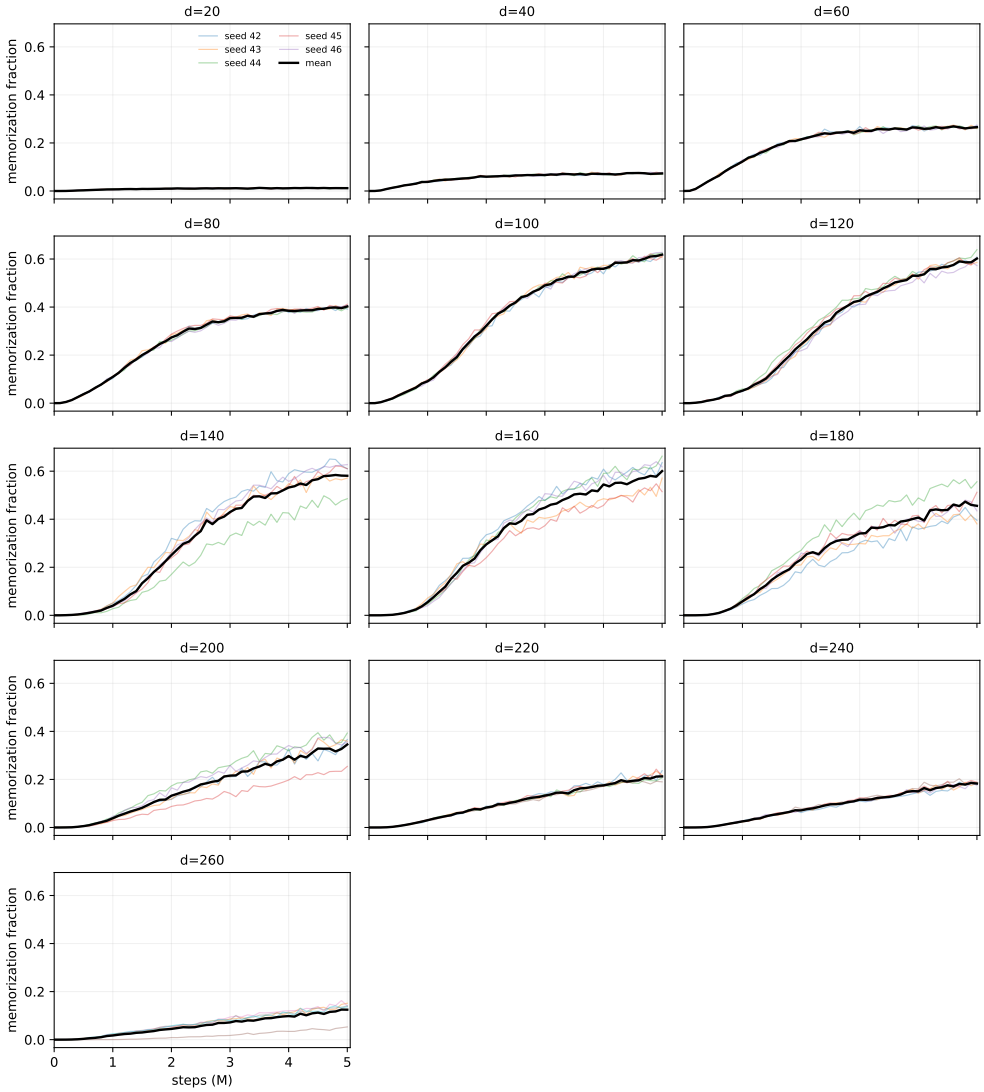


Figure 38: CIFAR-10 per-dimension memorization trajectories. The larger view exposes which latent dimensions memorize consistently across seeds and which remain limited by representation or optimization. As with CelebA, the low-dimensional regime mixes two effects: a smaller latent buffer and a VAE bottleneck that may be too lossy for pixel-space KNN memorization to detect exact sample identity. These comparisons depend on VAE identity preservation. Above the operational identity threshold, memorization is broadly delayed as width grows, but the hitting-time trend is not strictly monotone.

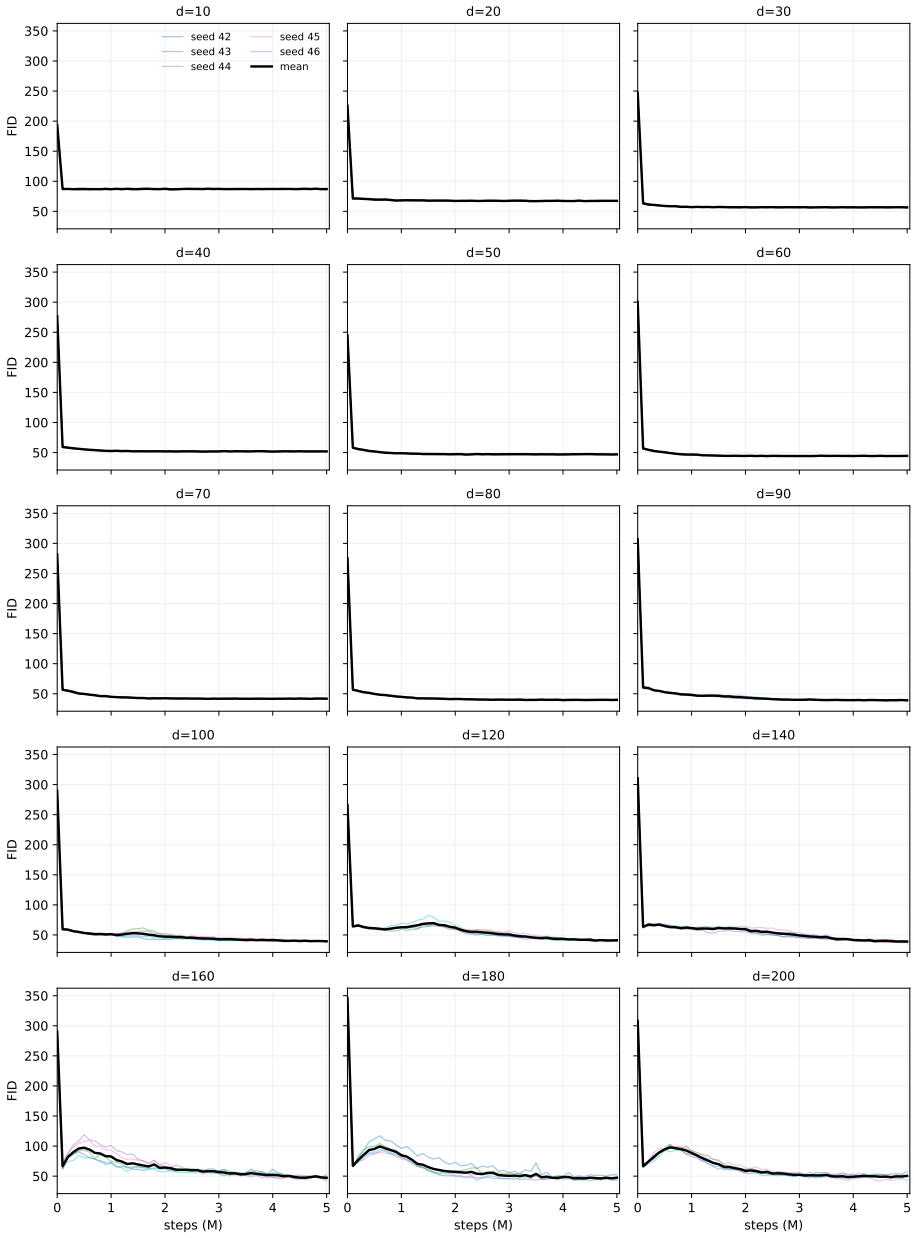


Figure 39: CelebA per-dimension FID trajectories. These panels use the same dimension-by-dimension organization as Figure 37 but track sample quality rather than memorization. The comparison shows where increasing d_{latent} improves representational fidelity and where it mainly changes the timing of sample recall. FID is best at intermediate widths and worsens again at the widest latents. This is an empirical quality trend, not an estimate of intrinsic dimension or a proof of an overwide-score mechanism. In particular, a dimension with better FID is not automatically more memorizing; the two quantities are coupled by the VAE representation and diffusion optimization but measure distinct behaviors.

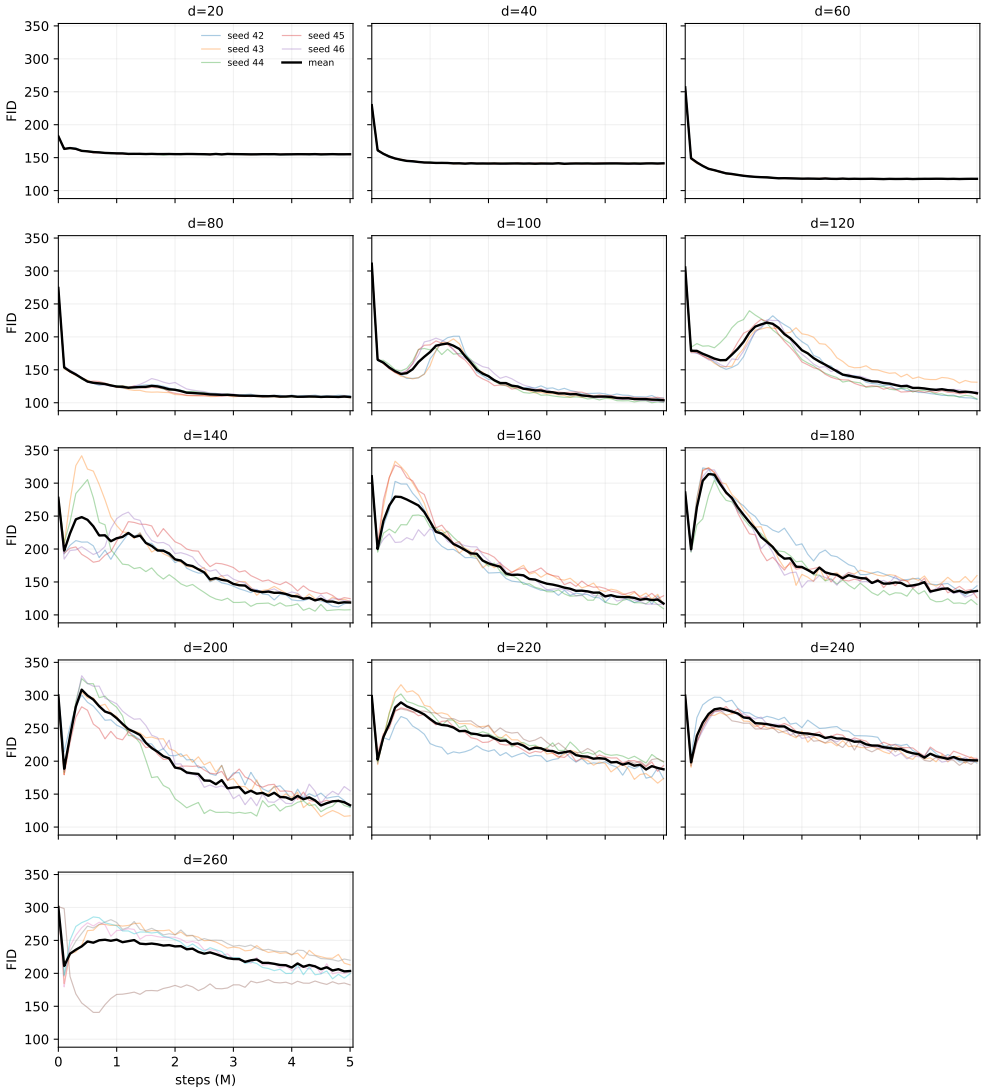
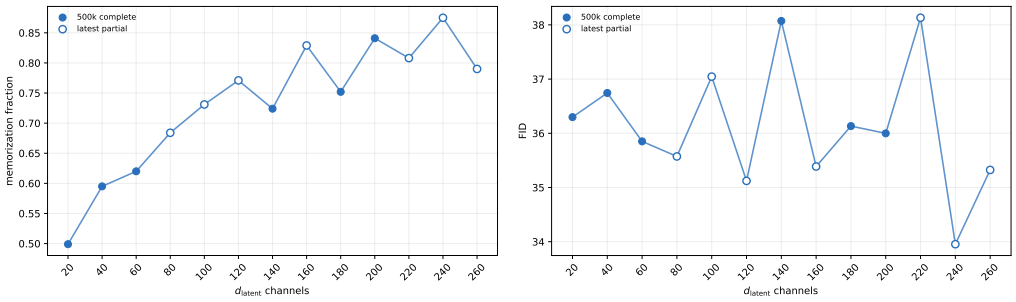


Figure 40: CIFAR-10 per-dimension FID trajectories. Plotted separately from the memorization trajectories, these curves show that sample quality and sample recall are not interchangeable metrics. Some latent dimensions improve FID before reaching high KNN memorization, while others memorize more strongly without proportionally improving image quality. The broad trend is still a smooth quality regime: FID decreases as the VAE bottleneck is relieved and increases again in the high-d_{latent} excess-latent regime.

N SPATIAL-DIT CELEBA PILOT

The main real-data experiments use vector latents and fixed-width MLP score models because that is the cleanest way to isolate the role of latent dimension. We also ran a CelebA pilot with a spatial VAE and a DiT-style score network to see what changes when the architecture looks more like a modern latent image generator. The VAE maps each 32×32 image to a $4 \times 4 \times d_{\mathrm{latent}}$ latent grid. Thus the channel count shown on the horizontal axis is not the total latent dimension: $d_{\mathrm{latent}} = 20$ already means 320 scalar latent coordinates, and $d_{\mathrm{latent}} = 260$ means 4160 scalar latent coordinates. This is the main reason the spatial sweep is less clean than the vector-latent MLP sweep. A one-channel change in the plotted variable adds sixteen coordinates, and the covariance structure now includes both channel and spatial-position interactions.

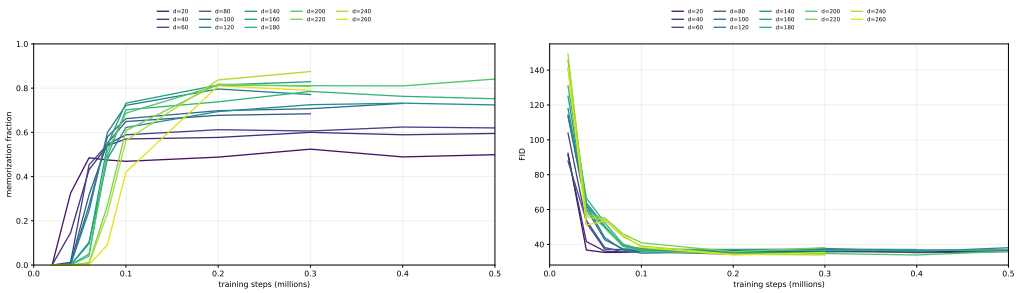
For the broad pilot, we train a DiT-S-style unconditional trunk with hidden size 384, depth 12, 6 attention heads, MLP ratio 4, batch size 1024, learning rate 10^{-4}, EMA decay 0.9999, and gradient clipping at 1.0. The model is evaluated at 20k, 40k, 60k, 80k, 100k, 200k, 300k, 400k, and 500k steps using 1k generated samples, pixel-space nearest-neighbor memorization against the same diverse 1k subset, and FID against that subset. Because this pilot uses a different architecture, EMA, gradient clipping, and a spatial latent representation, we treat it as a diagnostic rather than as part of the main causal evidence. The pilot does not reproduce the clean high-width decline in memorization of the vector MLP. It limits architectural transfer of the empirical claim; the changed representation and optimizer prevent attribution to one cause.



(a) Latest memorization by spatial channel count.

(b) Latest FID by spatial channel count.

Figure 41: CelebA spatial-DiT pilot summary across channel counts. Each plotted d_{latent} corresponds to a $4 \times 4 \times d_{\text{latent}}$ latent grid, so the true scalar latent dimension is $16d_{\text{latent}}$. The broad seed-42 sweep therefore changes latent size in large jumps compared with the vector MLP experiments. FID is reasonably stable across many widths, while memorization is high and does not show the clean high-dimensional decline seen in the fixed-vector MLP. This does not rule out a buffer effect in spatial architectures; rather, it shows that channel count, spatial covariance, VAE quality, EMA, clipping, and DiT optimization all enter before the scalar buffer can be read off as a one-dimensional sweep.



(a) Memorization over training.

(b) FID over training.

Figure 42: Spatial-DiT training trajectories for the broad CelebA pilot. The memorization panel uses the same decoded pixel-space nearest-neighbor ratio as the main real-data experiments. The FID panel uses the same 1k diverse subset as the reference. Unlike the vector MLP curves, these trajectories are not a clean function of the plotted channel count, because each width change also changes a 4×4 spatial representation and its second-order covariance structure.

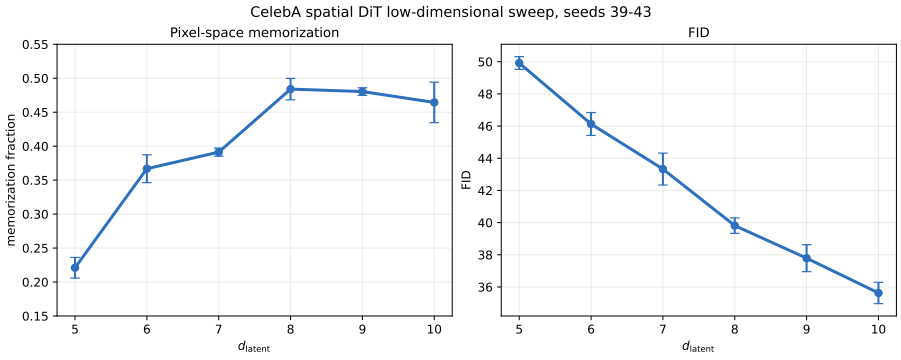


Figure 43: Low-dimensional spatial-DiT check on CelebA, averaged over seeds 39–43. The channel range $d_{\text{latent}} = 5$–10 corresponds to 80–160 scalar latent coordinates. Memorization rises from about 22% at $d_{\text{latent}} = 5$ to roughly 48% by $d_{\text{latent}} = 8$–10, while FID improves from about 50 to about 36. This is negative transfer evidence for an architecture-independent width rule: the trend differs from the fixed-vector-MLP result. The sweep does not identify which architectural or optimization change causes it.

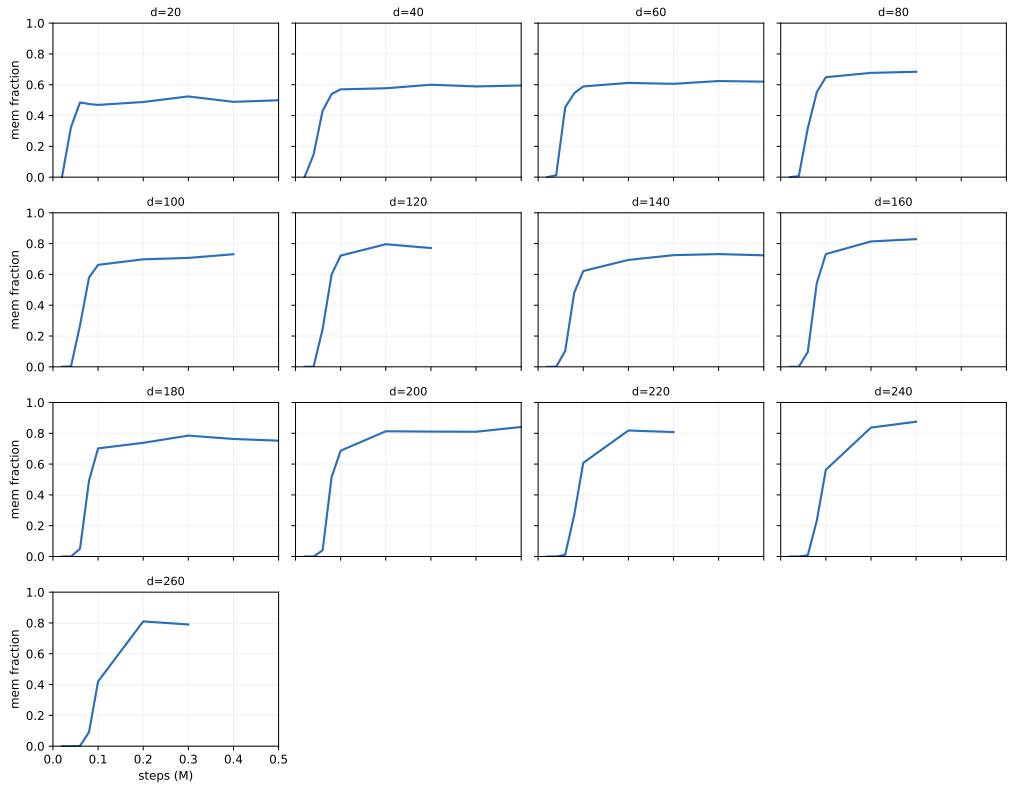


Figure 44: Per-channel spatial-DiT memorization trajectories. These panels are included to make clear that the broad pilot is not a direct replacement for the MLP evidence. Adjacent widths can behave differently because the architecture sees a structured grid rather than an unordered vector of latent coordinates.

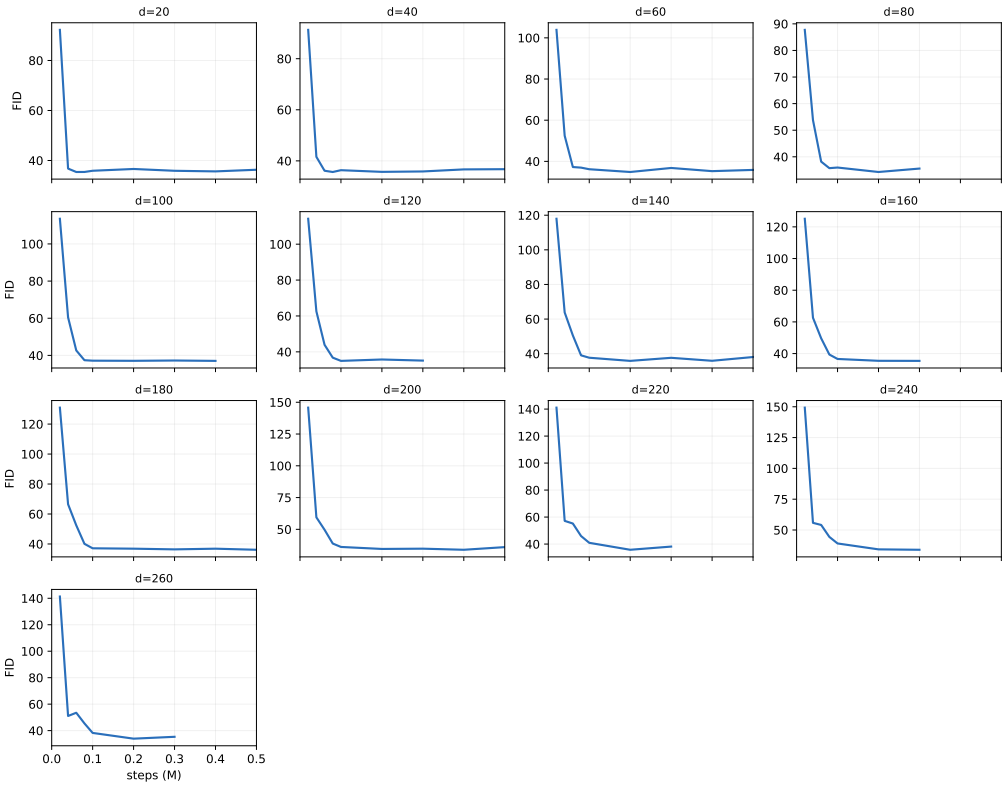
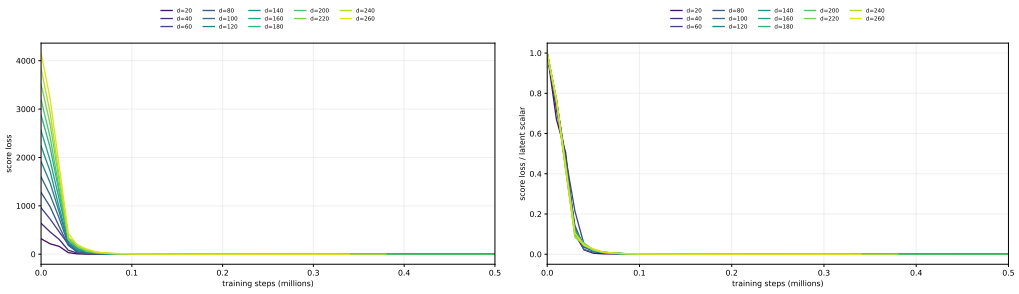


Figure 45: Per-channel spatial-DiT FID trajectories. Some spatial widths improve quality without reducing memorization, so sample quality and direct recall remain separate diagnostics in the DiT pilot as in the vector-latent MLP experiments.



(a) Score loss.

(b) Score loss per scalar latent coordinate.

Figure 46: Spatial-DiT optimization diagnostics. The per-scalar loss divides by $16d_{\text{latent}}$ to remove the trivial increase in the number of latent coordinates. These curves help distinguish ordinary training instability from genuine changes in memorization or FID.

O Data-Spectrum Memorization Predictor Diagnostics

The released spectral predictor uses frozen VAE encoder means, but is calibrated against observed diffusion hitting times. It does not inspect score-model weights and is not an uncalibrated forecast from data alone.

For each width d, let $Z_d \in \mathbb{R}^{n\times d}$ contain encoder means for the same diverse subset ($n = 1000$). At $t = 0.1$ define

$$M_t(d) = e^{-2t} Z_d^\top Z_d/n + (1 - e^{-2t}) I_d, \qquad P_d(s) = \frac{\sum_i w_i(d)(1 - e^{-\kappa\lambda_i(d)s})^2}{\sum_i w_i(d)}.$$

Here M_t is an uncentered second moment, and its eigenvalues are data-side quantities rather than sample-mode RFNN rates. The unweighted fit uses $w_i = 1$; the weighted diagnostic uses $w_i = \max\{\lambda_i - \beta_d, 0\}$ for an estimated isotropic floor. A single κ does not explicitly represent the varying $a_1(q_d)^2/d$ factor of the RFNN surrogate.

The released implementation fits one global rate κ and stores six ordered pressure thresholds θ_f, for $f \in \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75\}$, with $\widehat{\tau}_f(d) = \inf\{s : P_d(s) \geq \theta_f\}$. It scans $\log_{10}\kappa \in [-12, -4]$, averages empirical pressures with weights given by the number of observed seeds, imposes monotonicity by isotonic pooling, and minimizes squared log hitting-time error. These are one fitted rate and up to six fitted threshold levels per dataset, not a three-parameter anchored sigmoid. Fully censored thresholds receive filled values from neighboring levels and do not provide observed validation targets. Calibration uses all available uncensored widths, including widths below the VAE identity threshold. Unobserved crossings are excluded; means over crossing seeds can be biased downward when other seeds are censored. The overlays and hitting-time comparisons below are in-sample fits, not held-out validation. The separate one-constant g-clock and its leave-one-width-out checks are discussed in Appendix G.

Training-set-size sensitivity. Changing n would change $Z_d^\top Z_d/n$ if the subset were recomputed. The historical diagnostic in Figure 47 instead holds the saved $n = 1000$ spectra fixed and compares a fixed fraction $f = 0.01$ with a moving threshold $f = 1/n$. The latter is a threshold sensitivity diagnostic; since the measured fraction is over generated samples, it does not count distinct copied training identities or identify the first absolute copied sample. Neither panel is an empirical training-set-size sweep.

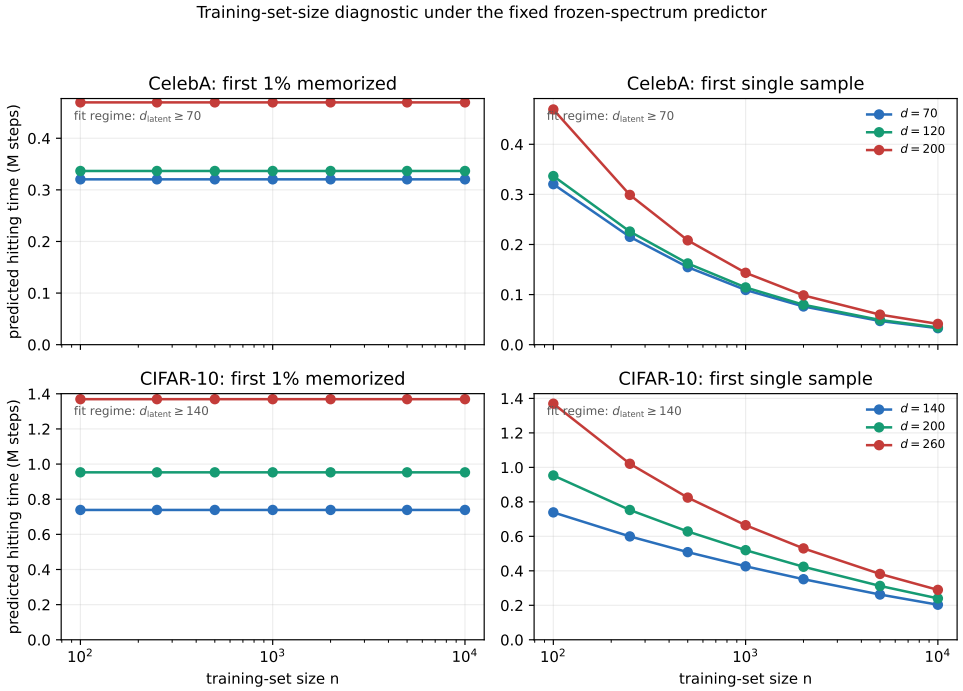
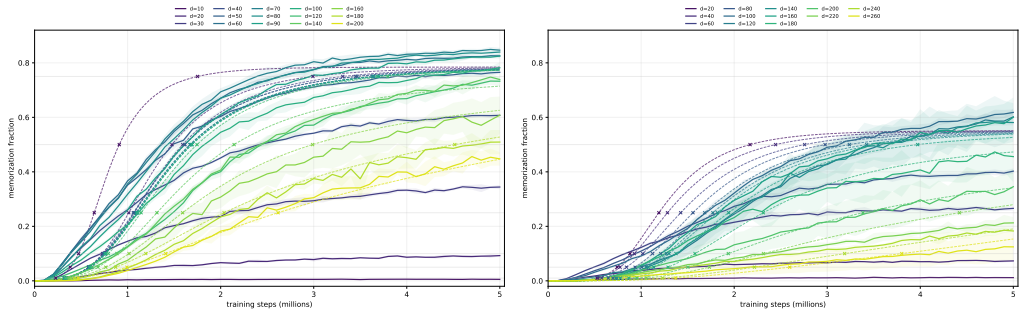


Figure 47: Historical threshold-sensitivity illustration at fixed saved $n = 1000$ spectra. This legacy panel is not a validation of the current released calibration; its fit-regime labels refer to that earlier illustration. Left: the time to a fixed predicted fraction $f = 0.01$ is therefore unchanged. Right: the query threshold is changed to $f = 1/n$, which can shift its predicted crossing time without changing any training dynamics. The legacy label "first single sample" should be read as the queried fraction $1/n$, not a count of copies or distinct training identities. These curves are not used for the main quantitative claims. A real training-set-size comparison requires new subsets, spectra, and diffusion evaluations.



(a) CelebA aggregate overlay.

(b) CIFAR-10 aggregate overlay.

Figure 48: Calibrated spectral-pressure overlays on empirical memorization curves. One rate constant and up to six monotone thresholds are fit per dataset using observed hitting times across all available widths. The displayed agreement is in-sample fit quality. Low-width observations can also be limited by VAE identity preservation; the reconstruction threshold does not identify intrinsic manifold dimension.

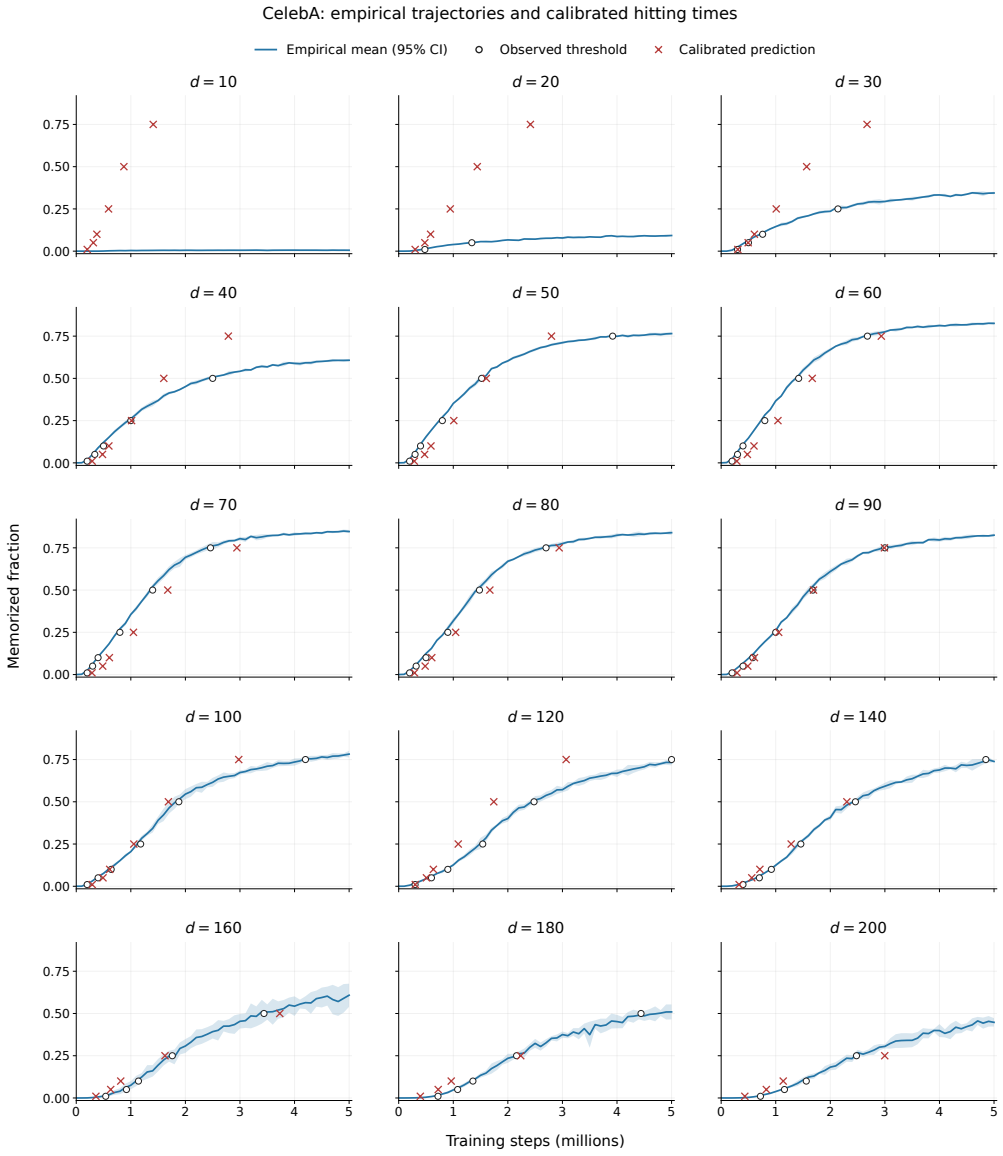


Figure 49: CelebA spectral-pressure fits by width. These panels compare calibrated threshold crossings with empirical memorization trajectories and expose censored or delayed levels. The same observations contribute to calibration, so agreement is not held-out validation. Low-width differences can be affected by VAE identity loss.

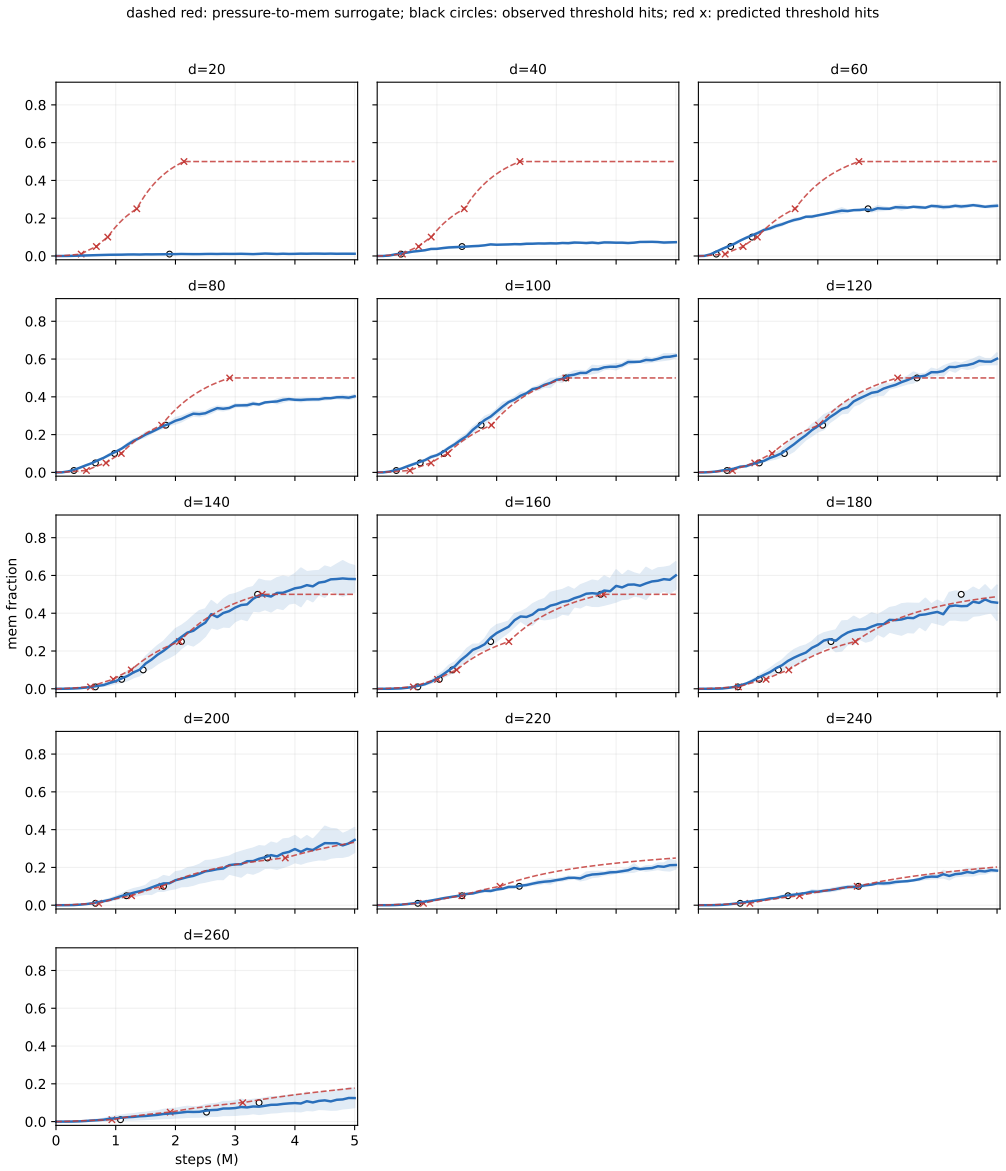
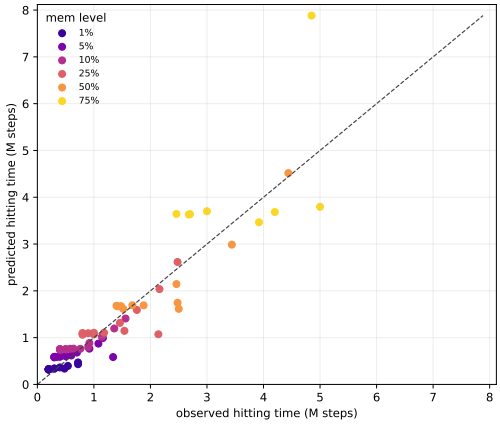
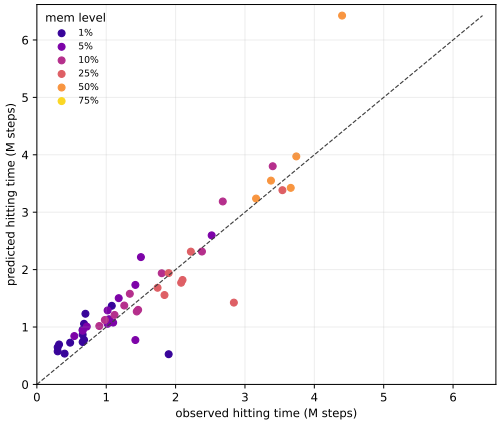


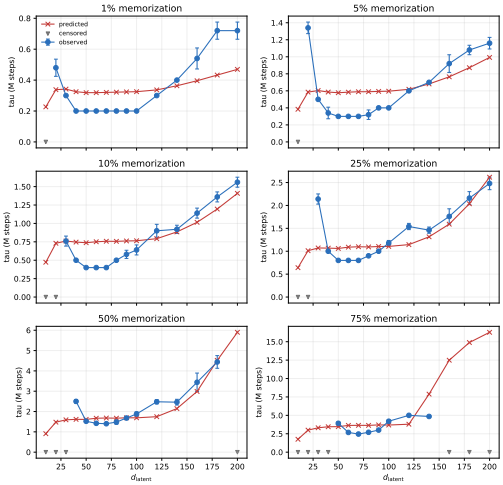
Figure 50: CIFAR-10 spectral-pressure fits by width. The global rate and observed monotone threshold levels are calibrated against hitting times across widths. The displayed predicted curve interpolates the pressure-to-fraction threshold anchors; it is not a fitted sigmoid. These are descriptive in-sample fits, and the observed post-threshold timing is not strictly monotone. VAE identity preservation and censoring affect interpretation.



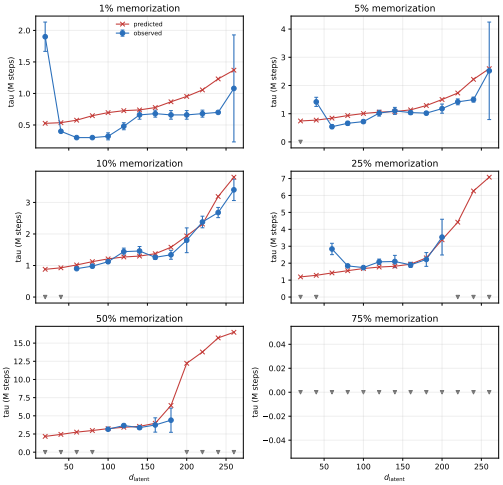
(a) CelebA predicted vs. observed hitting times.



(b) CIFAR-10 predicted vs. observed hitting times.

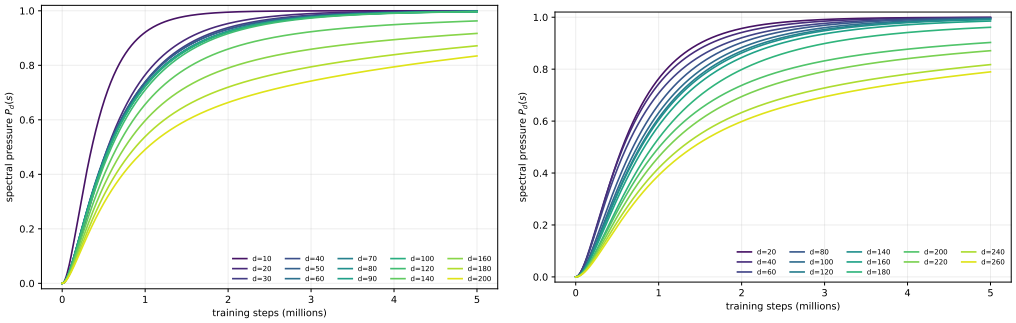


(c) CelebA hitting times by d_{latent}.



(d) CIFAR-10 hitting times by d_{latent}.

Figure 51: Hitting-time summaries of the released spectral-pressure fit. The predicted-versus-observed panels and width trends use observations that also enter calibration, so agreement measures in-sample fit quality rather than independent predictive validation. Missing crossings are censored and excluded from fitting; means over only crossing seeds can underestimate timing. Held-out diagnostics for the separate one-constant clock appear in Appendix G.



(a) CelebA unweighted pressure.

(b) CIFAR-10 unweighted pressure.

(c) CelebA excess-weighted pressure.

(d) CIFAR-10 excess-weighted pressure.

Figure 52: Calibrated pressure curves. Unweighted pressure treats all eigenvalues of the diffused latent second moment equally; excess-weighted pressure emphasizes eigenvalues above an estimated isotropic floor. These curves summarize data-side mode absorption. Identifying them with sample-specific RFNN learning requires additional assumptions, and neither pressure fit establishes the mechanism of the trained MLP.

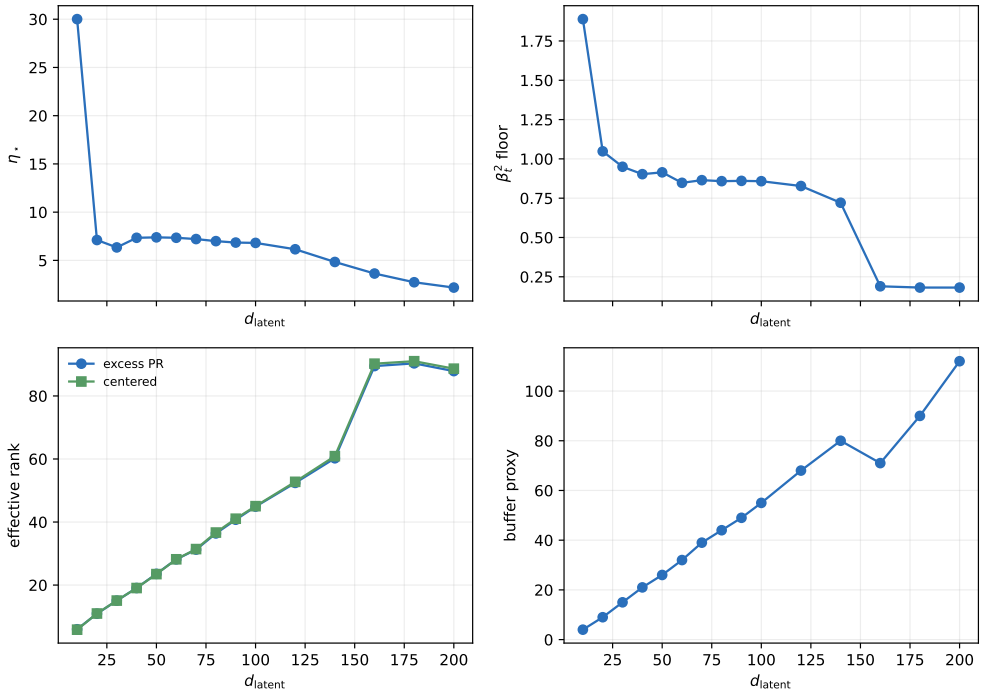


Figure 53: CelebA frozen-latent spectral summaries: eigenvalue statistics, effective rank, residual-mass proxies, and $q_d = \operatorname{tr} M_t/d$. These are computed from the encoded subset before diffusion training. The spectra characterize the representation; the memorization calibration separately uses diffusion outcomes. Their trends do not establish intrinsic dimension or validate the fitted mechanism.

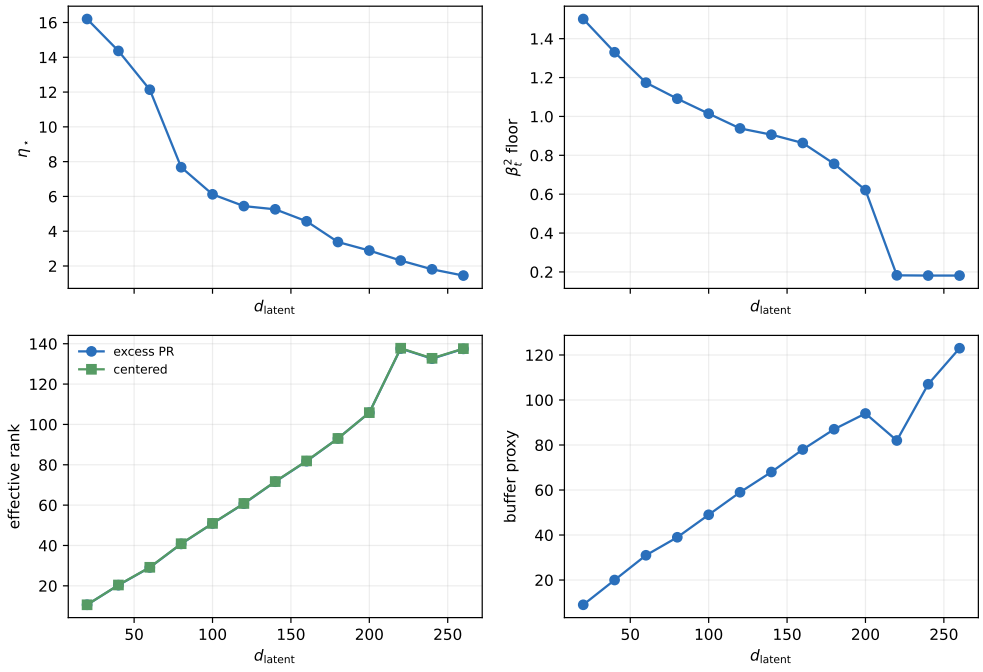


Figure 54: CIFAR-10 frozen-latent spectral summaries. These uncentered second-moment statistics describe the encoded training subset before diffusion training. Comparisons with CelebA show how representation statistics vary across datasets and widths; they do not isolate the addition of independent null coordinates or establish an intrinsic-dimension threshold.