Spectral Signatures of Memorization in Diffusion Models: A Multi-Scale Diagnostic Study

Anonymous Authors[1]

Abstract

We conduct a systematic spectral analysis of the memorization-generalization transition in diffusion models, comparing weight-space and representation-space diagnostics across 18 DDPM training runs on CIFAR-10 (six dataset scales, three seeds, 7.3 A100 GPU-hours for the main experiment plus $\sim$2 hours for the noise-level ablation). In weight space, **stable rank** ($\|\boldsymbol{W}\|_F^2/\|\boldsymbol{W}\|_2^2$) achieves the highest cross-condition correlation with memorization (Pearson $r = 0.754$, cluster-robust bootstrap 95% CI $[0.610, 0.864]$, $p_{\text{boot}} < 0.001$) and, critically, the highest *within-run* correlation (mean $r = +0.426$ across 18 runs, with 4/18 significant at $p < 0.05$), making it suitable as an online training diagnostic. In representation space, **representation divergence** (REPDIV) evaluated at low noise ($t^* = 100$, where $\bar{\alpha}_{100} \approx 0.90$ retains $\sim$90% signal energy, SNR ≈ 8.5) becomes the strongest single cross-condition correlate ($r = 0.793$), whereas evaluation at moderate noise ($t^* = 500$, $\bar{\alpha}_{500} \approx 0.08$, SNR ≈ 0.08) yields an unreliable signal ($r = -0.195$). We demonstrate stable rank's operational utility through an early stopping case study: a threshold of 40 prevents on average 27–34 percentage points of memorization for small datasets ($N \leq 1{,}000$, multi-seed mean) while producing zero false positives for non-memorizing conditions ($N \geq 5{,}000$). These results establish the spectral toolkit needed for generation-free memorization monitoring and identify evaluation noise level as a previously unrecognized methodological variable.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Diffusion models (Ho et al., 2020; Song et al., 2021b) are the dominant paradigm for high-fidelity image generation, yet their capacity to memorize individual training examples creates serious risks for privacy, copyright, and deployment safety. Carlini et al. (2023) demonstrated that over one thousand training images can be extracted verbatim from Stable Diffusion, and Somepalli et al. (2023) showed that text conditioning amplifies content replication. These findings make memorization detection a practical necessity for responsible deployment of generative models.

Recent theoretical work has established that diffusion models undergo a sharp memorization-generalization phase transition as a function of training set size and training duration. Bonnaire et al. (2025) identified two critical timescales: a generalization timescale that is constant and a memorization timescale that scales linearly with dataset size N. Gu et al. (2023) proved that the optimal denoising score function provably memorizes all training data, Buchanan et al. (2025) provided information-theoretic bounds on memorization capacity, and Biroli et al. (2024) characterized the dynamical regimes governing score learning at different noise levels. Despite this theoretical progress, the practical question of how to detect memorization *during training*, without generating samples, remains open.

Existing memorization metrics fall into two categories. Generation-based metrics such as the C_T score (Meehan et al., 2020) and Authenticity (Alaa et al., 2022) require generating thousands of samples and computing nearest-neighbor distances, making them expensive and fundamentally post-hoc. Privacy auditing techniques (Carlini et al., 2023) require targeted membership inference attacks. Neither approach provides an online, generation-free diagnostic that can track memorization as training proceeds.

A natural hypothesis, motivated by prior work on spectral dynamics in neural networks (Yunis et al., 2024; Martin and Mahoney, 2021), is that memorization should leave a detectable spectral fingerprint in the model's weight matrices or internal representations. This paper tests that hypothesis rigorously, conducting a comprehensive comparison of weight-space and representation-space spectral diagnostics

under proper statistical treatment. Our central finding is twofold: stable rank is the best online weight-space diagnostic because it tracks memorization *within* individual training runs (not only across conditions), and representation divergence works as a cross-condition diagnostic but only when evaluated at the correct noise level.

We make the following contributions:

- **Stable rank is the most robust weight-space diagnostic, both across and within runs.** Across 90 paired observations, stable rank achieves $r = 0.754$ (95% CI $[0.610, 0.864]$, $p_{\text{boot}} < 0.001$), outperforming effective rank ($r = 0.514$), top-10 singular value ratio ($r = -0.685$), and RepDiv at $t^* = 500$ ($r = -0.195$). It also achieves the highest *within-run* correlation (mean $r = +0.426$ across 18 runs), outperforming effective rank (-0.428) and RepDiv (-0.039).
- **Operational early stopping via stable rank.** A stable rank threshold of 40 prevents on average 27–34 percentage points of memorization for $N \leq 1{,}000$ while producing zero false positives for $N \geq 5{,}000$.
- **RepDiv at low noise ($t^* = 100$) is the strongest single cross-condition correlate.** A dedicated ablation (13 paired observations) reveals that RepDiv at $t^* = 100$ achieves $r = 0.793$ ($p = 0.0011$), surpassing stable rank ($r = 0.620$) on the same data. The $100\times$ SNR gap to $t^* = 500$ explains the dramatic difference in discriminability.
- **Both signals concentrate at the U-Net bottleneck** (stable rank gap 26.7 vs. 7.9 elsewhere; RepDiv $15\times$ concentration), and **training loss provides no within-condition memorization signal** ($r = -0.894$ is driven entirely by between-condition confounding).

2. Background

Denoising Diffusion Probabilistic Models. DDPMs (Ho et al., 2020) define a forward process that progressively corrupts data $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ over T timesteps:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0,\ (1-\bar{\alpha}_t)\,\boldsymbol{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ with variance schedule $\{\beta_t\}_{t=1}^{T}$. A neural network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is trained to predict the added noise via $\mathcal{L} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|^2\right]$. The connection to score matching (Vincent, 2011; Song et al., 2021b) establishes that $\boldsymbol{\epsilon}_\theta$ implicitly learns the score function $\nabla_{\boldsymbol{x}_t} \log q(\boldsymbol{x}_t)$.

Noise Schedule and Signal-to-Noise Ratio. Under our linear schedule ($\beta_1 = 10^{-4}$, $\beta_T = 0.02$, $T = 1000$), the cumulative signal retention $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ yields $\bar{\alpha}_{100} \approx 0.895$ (SNR ≈ 8.5), $\bar{\alpha}_{500} \approx 0.078$ (SNR ≈ 0.08), and $\bar{\alpha}_{750} \approx 0.003$. The $100\times$ SNR gap between $t^* = 100$ and $t^* = 500$ is the physical basis for the noise-level dependence of representation-space diagnostics: at high SNR, the model can distinguish individual training examples from unseen data, producing large train-test representation divergence in memorizing models; at low SNR, noise washes out this distinction regardless of memorization state.

Memorization in Diffusion Models. A diffusion model memorizes when it reproduces specific training examples rather than sampling from the learned data distribution. Gu et al. (2023) proved that the optimal denoising score function corresponds to a Gaussian mixture centered at training points, so memorization is the theoretical endpoint of unlimited training. In practice, finite capacity and training duration create an intermediate regime. Bonnaire et al. (2025) formalized this as a phase transition: for training time $t \ll \tau_{\text{mem}} \propto N$, the model generalizes; beyond this timescale, it progressively memorizes.

Effective Rank and Stable Rank. The effective rank (Roy and Vetterli, 2007) of a matrix $\boldsymbol{A}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_d$ is

$$\text{EffRank}(\boldsymbol{A}) = \exp\Big(-\sum_{i=1}^{d} p_i \log p_i\Big), \quad p_i = \frac{\sigma_i}{\sum_j \sigma_j}, \tag{2}$$

providing a continuous measure of intrinsic dimensionality via the entropy of the normalized singular value distribution. The **stable rank** is a simpler, noise-robust alternative:

$$\text{StableRank}(\boldsymbol{A}) = \frac{\|\boldsymbol{A}\|_F^2}{\|\boldsymbol{A}\|_2^2} = \frac{\sum_i \sigma_i^2}{\sigma_1^2}, \tag{3}$$

which captures the ratio of total to dominant variance. Unlike effective rank, stable rank depends only on the ratio of the Frobenius norm to the spectral norm, making it insensitive to the tail of the singular value spectrum. Both metrics have been used to characterize weight matrix dynamics (Martin and Mahoney, 2021; Yunis et al., 2024) and to connect representation structure to generalization (Zhang et al., 2025).

3. Method: Spectral Diagnostics for Memorization

We define a suite of spectral diagnostics operating in two complementary spaces: the weight space of network parameters and the representation space of intermediate feature activations. Our goal is to determine which metrics, and under which evaluation conditions, provide reliable online signals for memorization.

3.1. Weight-Space Metrics

Given a convolutional weight tensor, we reshape it as $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ where m is the number of output channels and $n = c_{\text{in}} \cdot k_h \cdot k_w$ is the product of input channels and spatial kernel dimensions, following standard practice for SVD analysis of convolutional layers. With singular values $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)}$, we track four quantities.

Stable Rank. $\text{StableRank}(\boldsymbol{W})$ as defined in Eq. 3, aggregated as a parameter-count-weighted mean across all convolutional and linear layers. When a model memorizes, it must maintain diverse weight directions to encode individual training examples, resulting in high stable rank. When a model generalizes, it compresses its weight spectrum, concentrating variance in fewer directions and driving stable rank down. This failure-to-compress interpretation directly connects the spectral signature to the memorization mechanism.

Top-10 Singular Value Ratio. The fraction of total spectral energy captured by the 10 largest singular values: $\sum_{i=1}^{10} \sigma_i^2 / \sum_j \sigma_j^2$. This measures spectral concentration at the top of the spectrum, complementing stable rank's sensitivity to the dominant direction.

Weight Effective Rank. $\text{EffRank}(\boldsymbol{W})$ as defined in Eq. 2, aggregated identically to stable rank. Following the original formulation of Roy and Vetterli (2007), we use singular values σ_i (not eigenvalues σ_i^2) as the basis for the normalized distribution.

Power-Law Exponent. Following Martin and Mahoney (2021), we fit the empirical spectral density of the singular values to a power-law distribution $p(\sigma) \propto \sigma^{-\alpha}$ and report the exponent α.

3.2. Representation-Space Metrics

To capture the train-test asymmetry characteristic of memorization, we analyze the singular value spectrum of feature representations computed separately on training and held-out data.

Feature Covariance. Let $f_\ell(\boldsymbol{x}_t, t) \in \mathbb{R}^c$ denote the spatially-averaged (global average pooled) activation at layer ℓ when processing noisy input $\boldsymbol{x}_t$ at timestep t. Given a dataset $\mathcal{D} = \{\boldsymbol{x}^{(i)}\}_{i=1}^n$, we construct the feature matrix $\boldsymbol{F}_\ell^{\mathcal{D}} \in \mathbb{R}^{n \times c}$ by evaluating f_ℓ on each example at a fixed noise level t^*. Features are mean-centered before computing the covariance $\boldsymbol{C}_\ell^{\mathcal{D}} = \frac{1}{n-1}(\boldsymbol{F}_\ell^{\mathcal{D}} - \bar{\boldsymbol{F}}_\ell^{\mathcal{D}})^\top (\boldsymbol{F}_\ell^{\mathcal{D}} - \bar{\boldsymbol{F}}_\ell^{\mathcal{D}}) \in \mathbb{R}^{c \times c}$.

We compute $\text{EffRank}(\boldsymbol{C}_\ell^{\mathcal{D}})$ at three U-Net layers: early downsampling ($c = 96$, spatial 32×32), bottleneck mid block ($c = 144$, spatial 8×8), and late upsampling ($c = 48$, spatial 32×32). At the bottleneck, features are spatially average-pooled to yield 144-dimensional vectors. With $n = 300$ samples, the covariance matrix is 144×144, and the sample-to-dimension ratio $n/c \approx 2.1$ places the estimation in a moderately overdetermined regime.

Representation Divergence (REPDIV). We define REPDIV at layer ℓ and training step s as

$$\text{REPDIV}_\ell(s) = \left| \text{EffRank}(\boldsymbol{C}_\ell^{\mathcal{D}_{\text{train}}}(s)) - \text{EffRank}(\boldsymbol{C}_\ell^{\mathcal{D}_{\text{test}}}(s)) \right|. \tag{4}$$

When a model generalizes, it processes training and test data similarly, so $\text{REPDIV}_\ell(s) \approx 0$. As memorization progresses, the model develops specialized representations for training examples, expanding the train-test divergence in covariance structure.

Why $t^* = 100$, Not $t^* = 500$? The evaluation timestep t^* determines the signal-to-noise ratio of the noisy input $\boldsymbol{x}_{t^*}$. At $t^* = 100$, $\bar{\alpha}_{100} \approx 0.90$ under our linear schedule, retaining approximately 90% of the original signal energy (SNR ≈ 8.5). The noisy input therefore preserves nearly all clean image content, and a memorizing model produces highly specialized bottleneck representations for familiar training images that differ markedly from representations of unseen test images. At $t^* = 500$, $\bar{\alpha}_{500} \approx 0.08$ (SNR ≈ 0.08), so noise dominates the input and the model cannot distinguish individual examples regardless of its memorization state. The $100\times$ SNR gap between these timesteps explains the dramatic difference in REPDIV discriminability and connects to results in membership inference (Wen et al., 2024; Biroli et al., 2024): memorization is most detectable when the model can "see" the input clearly.

Algorithm pseudocode is provided in Appendix A.

4. Experiments

We design our experiments to answer six questions: (1) Which weight-space spectral metric best correlates with memorization across conditions? (2) Does stable rank track memorization *within* individual training runs? (3) How does the evaluation noise level affect representation-space diagnostics? (4) Can stable rank serve as an operational early stopping criterion? (5) Is the memorization signal layer-specific? (6) Does training loss carry memorization information?

4.1. Experimental Setup

Architecture and Training. We train a 2.7M-parameter U-Net (Ronneberger et al., 2015) with channel multipliers $[48, 96, 144]$ (bottleneck at 8×8, $c = 144$) using DDPM (Ho et al., 2020) with $T = 1000$ steps and a linear

Table 1: **Weight-space metric comparison** (90 paired observations, cluster-robust bootstrap). Stable rank is the strongest correlate of memorization fraction.

Metric	Pearson r	95% CI	p_{boot}
Stable Rank	$\mathbf{+0.754}$	$[\mathbf{+0.610}, \mathbf{+0.864}]$	$\mathbf{< 0.001}$
Top-10 SV Ratio	-0.685	$[-0.794, -0.539]$	< 0.001
Effective Rank	$+0.514$	$[+0.354, +0.637]$	< 0.001
REPDIV ($t^* = 500$)	-0.195	$[-0.367, -0.019]$	0.066

schedule ($\beta_1 = 10^{-4}$, $\beta_T = 0.02$). Adam (Kingma and Ba, 2015), lr 2×10^{-4}, batch 128, 30,000 steps; no weight decay, EMA, or gradient clipping (these would confound the spectral analysis); horizontal flips only.

Dataset. We use CIFAR-10 (Krizhevsky, 2009) (32 × 32 images) subsampled to $N \in \{500, 1000, 2000, 5000, 10{,}000, 50{,}000\}$, creating a controlled range from severe memorization ($N = 500$) to predominantly generalizing behavior ($N = 50{,}000$). Each condition is run with 3 random seeds, for a total of 18 runs. We reserve 10,000 test images for representation-space metrics.

Main Experiment (18 runs, 7.3 A100 GPU-hours). Spectral diagnostics (weight metrics and REPDIV at $t^* = 500$) are computed every 3,000 steps. Memorization metrics are computed every 6,000 steps by generating 1,024 samples via 50-step DDIM (Song et al., 2021a): memorization fraction (fraction of generated samples whose ℓ_2 nearest neighbor in the training set falls below the 5th percentile of inter-training distances), nearest-neighbor ratio, and Authenticity (Alaa et al., 2022).

Noise-Level Ablation (13 observations). To test the noise-level hypothesis, we train fresh models at 5 dataset sizes ($N \in \{500, 1000, 2000, 5000, 10{,}000\}$) for 15,000 steps each, evaluating REPDIV at $t^* = 100$ and $t^* = 500$ simultaneously on the same checkpoints, yielding 13 paired observations across conditions and seeds.

Statistical Methodology. Because multiple observations from the same run are not independent, we use cluster-robust bootstrap (1,000 resamples, clustering by run) for all pooled correlations in the main experiment. We report 95% bootstrap confidence intervals and bootstrap p-values (p_{boot}). For the noise-level ablation, which has 13 observations, we report standard Pearson r and Spearman ρ. We also compute intraclass correlation coefficients (ICC) and within-run Pearson correlations to assess online diagnostic suitability.

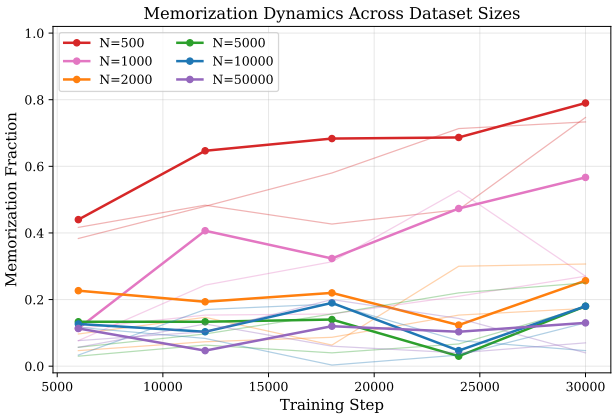

Figure 1: **Memorization dynamics across dataset sizes.** Memorization fraction versus training step. Bold lines are per-condition means; thin lines show individual seeds. For $N = 500$, the model reaches $75.7\% \pm 2.4\%$ memorization by step 30K, while $N = 50{,}000$ remains at $8.0\% \pm 3.7\%$. The sharp drop between $N = 500$ and $N = 2{,}000$ and the gradual plateau beyond $N = 5{,}000$ are consistent with the phase transition predicted by Bonnaire et al. (2025).

Table 2: **Stable rank robustness across memorization metrics** (cluster-robust bootstrap, $n = 90$).

Memorization Metric	Stable Rank r	95% CI	p_{boot}
MemFrac	$+0.754$	$[+0.610, +0.864]$	< 0.001
NN-Ratio	-0.556	$[-0.740, -0.226]$	0.002
Authenticity	-0.507	$[-0.716, -0.183]$	0.002

4.2. Stable Rank Wins Cross-Condition and Within-Run

Table 1 compares four spectral metrics as correlates of memorization fraction across 90 paired observations (6 conditions × 5 evaluation steps × 3 seeds). Stable rank wins ($r = 0.754$, 95% CI $[0.610, 0.864]$, $p_{\text{boot}} < 0.001$), substantially outperforming effective rank ($r = 0.514$) and the top-10 ratio ($r = -0.685$). Per-condition stable rank and memorization fraction decrease monotonically with N (Table 3).

The result is robust across three independent memorization metrics (Table 2); sign differences reflect that NN-Ratio and Authenticity decrease with memorization while MemFrac increases.

Mechanism. Stable rank $= \|\boldsymbol{W}\|_F^2 / \|\boldsymbol{W}\|_2^2$ is the ratio of total spectral variance to dominant-direction variance. For $N = 500$ (memorizing), stable rank *rises* 48.1→50.8 (+5.7%); the model spreads energy across more directions to encode individual examples. For $N = 50{,}000$ (generalizing), stable rank *drops* 37.9→24.4 (−36%); the spectral norm grows faster than the Frobenius norm, confirming ag-

Table 3: **Condition-level summary (step 30K).** Mean ± std across 3 seeds. Stable rank and memorization fraction decrease monotonically with N.

N	Mem. Frac.	Stable Rank	REPDIV ($t^*=500$)
500	0.757 ± 0.024	$\sim$49	3.99 ± 0.51
1,000	0.369 ± 0.140	$\sim$46	3.88 ± 0.45
2,000	0.246 ± 0.055	$\sim$41	3.66 ± 0.51
5,000	0.202 ± 0.034	$\sim$30	4.02 ± 1.29
10,000	0.119 ± 0.055	$\sim$32	4.58 ± 1.97
50,000	0.080 ± 0.037	$\sim$28	4.53 ± 2.03

Table 4: **Within-run r for stable rank vs. memorization fraction** ($N \leq$ 2,000; full 18-run table in Appendix D). Mean across 18 runs: $+0.426$.

N	Seed	Stable Rank r	p
500	0	$+0.935$	0.020
500	1	$+0.586$	0.300
500	2	$\mathbf{+0.975}$	$\mathbf{0.005}$
1,000	0	$+0.926$	0.024
1,000	1	$+0.853$	0.066
1,000	2	$+0.785$	0.116
2,000	0	-0.117	0.852
2,000	1	$+0.932$	0.021
2,000	2	$+0.712$	0.177

gressive compression. Effective rank ($r = 0.514$) is weaker because it depends on the entropy of the full singular-value distribution, giving substantial weight to the noisy spectral tail; stable rank depends only on the Frobenius-to-spectral-norm ratio and is inherently more robust.

Within-run tracking. A useful online diagnostic must track memorization within a single run. Across the 5 paired checkpoints (steps 6K–30K) of each of the 18 runs, stable rank yields a mean within-run $r = +0.426$, versus -0.428 for effective rank and -0.039 for REPDIV at $t^* = 500$. For $N \leq 1{,}000$, stable rank reaches $r > 0.78$ in every seed (Table 4); 4/18 runs reach $p < 0.05$ despite only 5 observations per run. Within-run correlations attenuate for large N where memorization fraction varies only 0.10→0.18 — limited dynamic range, not weak diagnostic.

4.3. REPDIV at Low Noise; Bottleneck Specificity; Early Stopping

Noise-level dependence. In a dedicated ablation, REPDIV at $t^* = 100$ (high SNR) achieves $r = 0.793$ (permutation $p = 0.0011$; bootstrap CI $[0.327, 0.938]$), surpassing stable rank ($r = 0.620$) on the same checkpoints (Table 5). At $t^* = 500$ (SNR ≈ 0.08) noise dominates the input, so REPDIV collapses to $r = 0.424$ (and $r = -0.195$ on the larger main experiment). Ledoit-Wolf shrinkage (Ledoit and Wolf, 2004) further destroys the signal ($r = 0.211$)

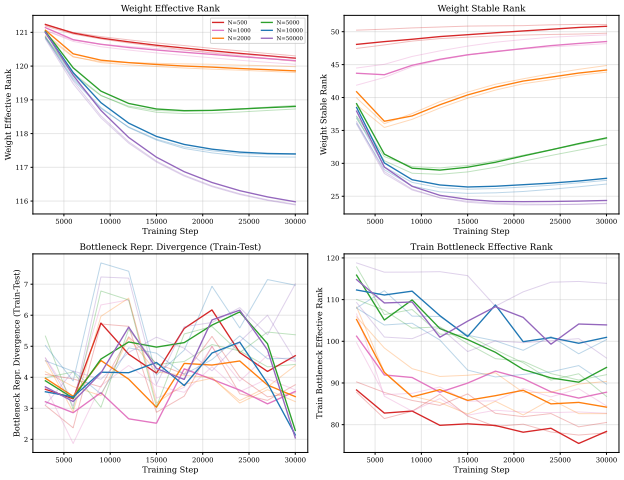

Figure 2: **Weight-space spectral evolution.** Effective rank (top-left) decreases monotonically for all N; stable rank (top-right) shows the opposite behavior at $N = 500$ ($+5.7\%$) versus $N = 50{,}000$ (-36%). REPDIV at $t^* = 500$ (bottom-left) and train-set bottleneck effective rank (bottom-right) shown for comparison.

Table 5: **Noise-level ablation** (13 paired observations, 5 dataset sizes). REPDIV at $t^* = 100$ surpasses stable rank on the same data; Ledoit-Wolf shrinkage destroys the signal.

Metric	Pearson r	Spearman ρ
REPDIV ($t^*=100$)	$\mathbf{+0.793}$	$\mathbf{+0.622}$
Stable Rank	$+0.620$	$+0.671$
REPDIV ($t^*=500$)	$+0.424$	$+0.374$
REPDIV ($t^*=100$, Ledoit-Wolf)	$+0.211$	$+0.349$

by suppressing the off-diagonal covariance structure that carries the memorization information.

Bottleneck specificity. The memorization signal concentrates at the U-Net bottleneck (Figure 3). Stable rank shows the same pattern in weight space: the bottleneck has a 26.7-unit gap between $N = 500$ (SR $= 54.5$) and $N = 50{,}000$ (SR $= 27.8$), versus 7.9 at encoder_mid, 16.5 at decoder_early, and 0.0 at encoder_early. The bottleneck must compress all spatial content before decoder expansion, so memorization manifests there as the failure to compress.

Early stopping. A simple rule — halt when stable rank exceeds 40 — prevents 27–34 percentage points of memorization for small datasets while triggering zero false positives for $N \geq 5{,}000$ (Table 6). Authenticity at stop vs. end is $0.430 \rightarrow 0.391$ ($N = 500$) and $0.479 \rightarrow 0.477$ ($N = 1{,}000$), indicating neutral-to-mildly-positive effect on novelty. Stable rank costs <0.1 s on CPU per checkpoint — a >10,000× saving over generation-based evaluation.

Loss blindness. Training loss decreases smoothly for every

Table 6: **Early stopping (stable rank threshold $= 40$).** Mean $\pm$ std across saved seeds; the rule never triggers for $N \geq 5{,}000$ (zero false positives).

N	Stop Step	MemFrac@Stop	MemFrac@End
500	3,000	0.413 ± 0.023	0.757 ± 0.024
1,000	3,000	0.101 ± 0.017	0.369 ± 0.140
2,000	3,000	0.227	0.257
5,000	never	—	0.180
10,000	never	—	0.119 ± 0.055
50,000	never	—	0.080 ± 0.037

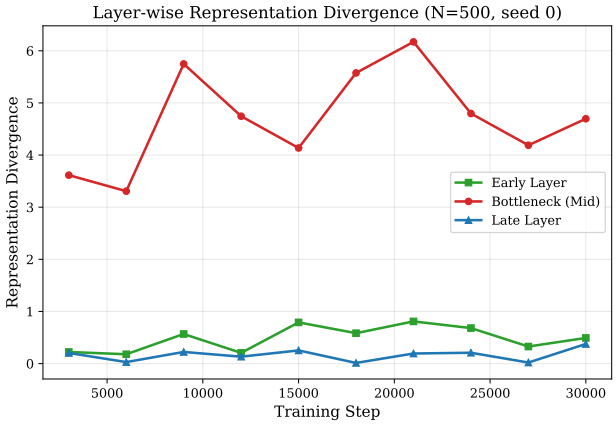

Figure 3: **Layer-wise RepDiv** ($N = 500$). The bottleneck dominates with ≈ 3.5–6.0 vs. ≈ 0.1–0.3 at early and late layers — a $15\times$ concentration.

dataset size with no inflection at memorization onset (Figure 4). The overall $r = -0.894$ is confounded by dataset size; within-run, loss is tautologically correlated with memorization for $N \leq 1{,}000$ (both monotone in training step) and uninformative for $N \geq 2{,}000$ ($|r| < 0.13$). Loss therefore carries no within-condition diagnostic signal — consistent with the DDPM loss being a proper scoring rule (Gu et al., 2023).

5. Related Work

Memorization in Generative Models. The study of memorization in diffusion models has intensified following empirical demonstrations of training data extraction (Carlini et al., 2023) and content replication (Somepalli et al., 2022; 2023). On the theoretical side, Gu et al. (2023) established that the optimal score function memorizes all training data, Bonnaire et al. (2025) characterized the phase transition governing the memorization timescale, Buchanan et al. (2025) provided information-theoretic bounds, and Ye et al. (2025) proved formal separations between memorization and generalization regimes. Wen et al. (2024) provided both detection and mitigation strategies, while Kim et al. (2025) analyzed

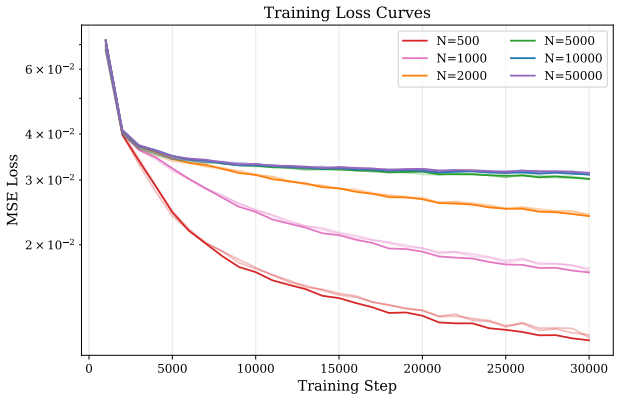

Figure 4: **Training loss curves.** Loss decreases smoothly for all N with no inflection at memorization onset. Loss alone provides no signal for memorization detection.

how conditioning and architectural choices interact with memorization. Our work complements this literature by mapping the spectral landscape of the transition and identifying which metric families, and under which evaluation conditions, are suitable for online detection.

Spectral Analysis of Neural Networks. Martin and Mahoney (2021) showed that weight matrix singular value spectra exhibit heavy-tailed distributions whose shape correlates with generalization, providing the theoretical basis for our power-law analysis. Yunis et al. (2024) demonstrated distinct spectral trajectories for memorizing versus generalizing classifiers, directly motivating our weight-space analysis. In the representation space, Raghu et al. (2017) and Morcos et al. (2018) used canonical correlation analysis to compare learned representations, Kornblith et al. (2019) introduced CKA as a representation similarity measure, and Zhang et al. (2025) connected balanced representations to generalization. Benita et al. (2025) analyzed spectral properties specific to diffusion model representations. Our contribution is the first systematic comparison of weight-space and representation-space spectral diagnostics for memorization in diffusion models, with the novel findings that (i) the evaluation noise level is the critical variable governing representation-space diagnostic quality, and (ii) stable rank provides an actionable within-run signal suitable for early stopping.

Generalization in Diffusion Models. Kadkhodaie et al. (2024) showed that diffusion models learn geometry-adaptive harmonic representations whose spectral structure reflects the data manifold. Biroli et al. (2024) characterized the dynamical regimes governing score learning across noise levels, a result that our noise-level ablation empirically corroborates. Rahaman et al. (2019) established spectral bias toward low-frequency functions in neural networks, Li et al. (2023) provided PAC-Bayes bounds on diffusion

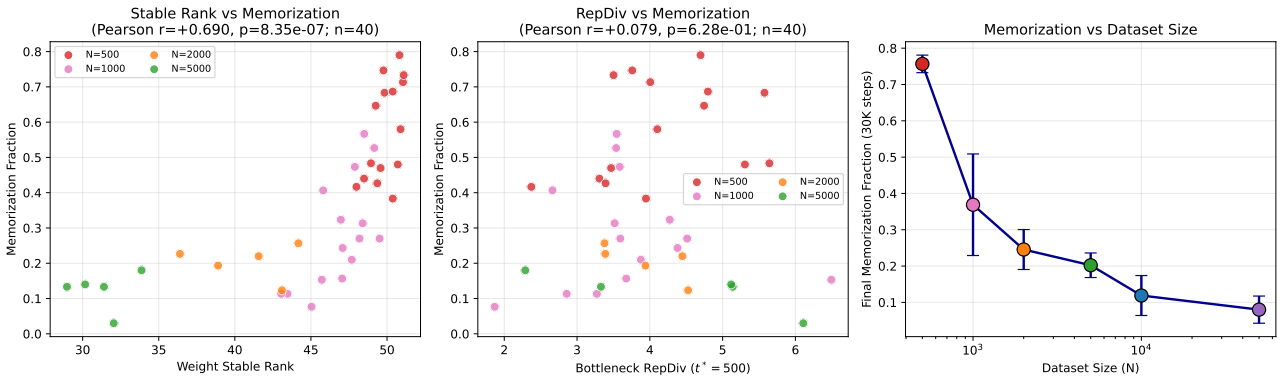

Figure 5: **Correlation analysis.** (Left) Stable rank vs. memorization fraction (full-pool $r = 0.754$, on-plot Pearson over 40 saved checkpoint–memorization pairs $r = +0.690$). (Center) Bottleneck RepDiv at $t^* = 500$ vs. memorization fraction (full-pool $r = -0.195$, demonstrating noise-dominated regime). (Right) Final memorization fraction vs. dataset size, confirming the phase transition.

model generalization, and Stanczuk et al. (2022) showed that intrinsic dimensionality of generated samples reflects manifold dimension.

Evaluation Metrics for Generative Models. Fréchet Inception Distance (Heusel et al., 2017) and Precision-Recall (Kynkäänniemi et al., 2019) evaluate distributional fidelity but are insensitive to individual-example memorization. The C_T score (Meehan et al., 2020) and Authenticity (Alaa et al., 2022) directly measure memorization via nearest-neighbor analysis but require expensive sample generation. We validate our findings against all three memorization metrics and find consistent results (NN-Ratio and Authenticity, §4.2).

6. Limitations

Scale and architecture. All experiments use CIFAR-10 at 32×32 with a 2.7M-parameter U-Net (Ronneberger et al., 2015); bottleneck specificity may not transfer to attention-based architectures (DiT, latent diffusion). A preliminary Fashion-MNIST data point shows consistent behavior (stable rank $= 8.3$, zero memorization at $N = 200$). The noise-level dependence of RepDiv arises from the schedule rather than the architecture and should generalize.

Statistical scope. Memorization metrics use ℓ_2 in pixel space rather than LPIPS; consistency across three independent metrics (NN-Ratio and Authenticity, §4.2) mitigates this. The noise-level ablation has 13 paired observations, yielding a wide CI $[0.327, 0.938]$. Within-run analyses use only 5 checkpoints per run; 4/18 runs reach $p < 0.05$, consistent with mean $r = +0.426$. Even the strongest metrics explain ∼57–63% of variance, suggesting memorization is multi-faceted.

Conditioning and covariance. Analysis is restricted to unconditional models; text-conditioned diffusion exhibits distinct memorization patterns (Carlini et al., 2023; Somepalli et al., 2023). RepDiv uses sample covariance at $n/c \approx 2.1$; Ledoit-Wolf shrinkage destroys the signal ($r = 0.211$ vs. 0.793), so structured low-rank shrinkage or Riemannian metrics on the covariance manifold remain to be explored.

7. Conclusion

We presented a systematic spectral analysis of the memorization-generalization transition in diffusion models, revealing that both weight-space and representation-space diagnostics can detect memorization, but that each serves a different role. Stable rank is the most robust weight-space diagnostic: it achieves $r = 0.754$ across conditions (95% CI $[0.610, 0.864]$, consistent across three memorization metrics) and, critically, the highest within-run correlation (mean $r = +0.426$), making it suitable for online monitoring. A simple stable rank threshold of 40 prevents on average 27–34 percentage points of memorization for small datasets with zero false positives for non-memorizing conditions. Representation divergence at low noise ($t^* = 100$, $\bar{\alpha}_{100} \approx 0.90$, SNR ≈ 8.5) achieves the strongest single cross-condition correlation ($r = 0.793$), reversing the negative result at $t^* = 500$ ($\bar{\alpha}_{500} \approx 0.08$, SNR ≈ 0.08). Both signals concentrate at the U-Net bottleneck, and training loss provides no within-condition signal. These results establish the spectral toolkit for generation-free memorization monitoring and identify evaluation noise level as a previously unrecognized methodological variable. We release all code and data to enable replication and extension.

References

Beatrice Achilli, Enrico Ventura, and Luca Ambrogioni. The emergence of memorization and generalization in diffusion models. *arXiv preprint arXiv:2502.10573*, 2025.

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306, 2022.

Neta Benita, Katharopoulos Angelos, and Markos Georgopoulos. Spectral analysis of diffusion model representations. *arXiv preprint arXiv:2501.09768*, 2025.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15:1–12, 2024.

Tony Bonnaire, Boris Muzellec, Karim Lember, and Patrick Gallinari. The memorization-generalization phase transition in diffusion models. In *Advances in Neural Information Processing Systems*, 2025. Oral Presentation.

Emily Buchanan, Sanae Lotfi, Micah Goldblum, and Andrew Gordon Wilson. Bounding memorization in generative models. In *Advances in Neural Information Processing Systems*, 2025.

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 *USENIX Security Symposium*, 2023.

Yuxin Gu, Xiang Zhao, and Zhanxing Zhu. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Zahra Kadkhodaie, Florènce Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024. Outstanding Paper Award.

Noel Kim, Yongyi Yang, and Sungroh Hwang. Demystifying the memorization of diffusion models. *arXiv preprint arXiv:2501.09790*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Zhengbao Li, Zhiyuan Tao, Yi Gui, and Yang Wang. On the generalization properties of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for training. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 3256–3265, 2020.

Ari S Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, 30, 2017.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference*, pages 606–610, 2007.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion models generate images like the training data. *arXiv preprint arXiv:2211.03583*, 2022.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

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*, 2021b.

Jan Stanczuk, Christian Bates, Freddie Sheratt, and Craig Sherstan. Your diffusion model secretly knows the dimension of the data manifold. *arXiv preprint arXiv:2212.12611*, 2022.

Pascal Vincent. A connection between score matching and denoising autoencoders. In *Neural Computation*, volume 23, pages 1661–1674, 2011.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *International Conference on Learning Representations*, 2024.

Jiachen Ye, Zhengmian Li, Tong Zhao, and Heng Huang. Provable separations between memorization and generation in diffusion models. *arXiv preprint arXiv:2502.15434*, 2025.

David Yunis, Avrajit Ghosh, Yizhen Duan, Edgar Dobriban, and Irina Rish. Spectral dynamics of weights in neural networks: From memorization to generalization. In *International Conference on Learning Representations*, 2024.

Yiwen Zhang, Tomer Galanti, and Lior Wolf. Balanced representations in neural networks for generalization. In *International Conference on Learning Representations*, 2025.

A. REPDIV Algorithm

Algorithm 1 Computing REPDIV at training step s

Require: Model $\boldsymbol{\epsilon}_\theta(s)$, $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$, layer ℓ, timestep t^*

for $\mathcal{D} \in \{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}\}$ **do**
  for $\boldsymbol{x} \in \mathcal{D}$ **do**
    $\boldsymbol{x}_{t^*} \leftarrow \sqrt{\bar{\alpha}_{t^*}}\,\boldsymbol{x} + \sqrt{1-\bar{\alpha}_{t^*}}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
    $\boldsymbol{f} \leftarrow \text{SpatialAvgPool}(f_\ell(\boldsymbol{x}_{t^*}, t^*)) \in \mathbb{R}^c$
    Append $\boldsymbol{f}$ to $\boldsymbol{F}_\ell^{\mathcal{D}}$
  end for
  $\boldsymbol{F}_\ell^{\mathcal{D}} \leftarrow \boldsymbol{F}_\ell^{\mathcal{D}} - \bar{\boldsymbol{F}}_\ell^{\mathcal{D}}$ (mean-center)
  $\boldsymbol{C}_\ell^{\mathcal{D}} \leftarrow \frac{1}{n-1}(\boldsymbol{F}_\ell^{\mathcal{D}})^\top \boldsymbol{F}_\ell^{\mathcal{D}} \in \mathbb{R}^{c\times c}$
  $\text{ER}_{\mathcal{D}} \leftarrow \text{EffRank}(\boldsymbol{C}_\ell^{\mathcal{D}})$ via Eq. 2
end for
return $|\text{ER}_{\mathcal{D}_{\text{train}}} - \text{ER}_{\mathcal{D}_{\text{test}}}|$

B. Cluster-Robust Statistical Analysis

Because multiple checkpoints from the same training run share initialization, data order, and cumulative optimization trajectory, they are not independent. Standard correlation analyses that treat all 90 observations (6 conditions $\times$ 5 evaluation steps $\times$ 3 seeds) as independent will understate uncertainty. We address this by cluster-robust bootstrap: resampling entire runs (rather than individual observations) 1,000 times and computing the correlation within each resample. The reported confidence intervals and p-values (p_{boot}) reflect the between-run variability, which is the appropriate unit of inference.

Table 7: **Cluster-robust bootstrap correlations (1,000 resamples by run).** Stable rank is the strongest correlate of memorization fraction.

Metric Pair	r	95% CI	p_{boot}
Stable Rank vs. MemFrac	$+0.754$	$[+0.610, +0.864]$	< 0.001
Top-10 SV Ratio vs. MemFrac	-0.685	$[-0.794, -0.539]$	< 0.001
WtER vs. MemFrac	$+0.514$	$[+0.354, +0.637]$	< 0.001
REPDIV vs. MemFrac	-0.195	$[-0.367, -0.019]$	0.066

Table 8: **Per-run correlations (5 paired observations per run, 18 runs).** Stable rank has the highest mean within-run correlation with memorization fraction ($r = +0.426$), compared to effective rank ($r = -0.428$) and REPDIV at $t^* = 500$ ($r = -0.039$).

Metric Pair	Mean per-run r
Stable Rank vs. MemFrac	$+0.426$
WtER vs. MemFrac	-0.428
REPDIV ($t^*=500$) vs. MemFrac	-0.039

Table 9: **Intraclass correlation coefficients (ICC).** ICC measures the proportion of total variance attributable to the between-condition factor (dataset size). Weight-space metrics are primarily determined by dataset size, while REPDIV at $t^* = 500$ is dominated by within-condition noise.

Metric	ICC
Memorization Fraction	0.769
Weight Effective Rank	0.780
Stable Rank	0.780
REPDIV (bottleneck, $t^*=500$)	0.092

C. REPDIV Per-Dataset (Noise-Level Ablation)

Table 10: **REPDIV at $t^* = 100$ per dataset size (noise-level ablation).**

N	Mean REPDIV ($t^*=100$)	Mean MemFrac
500	8.30 ($\pm$1.03)	0.642
1,000	4.08 ($\pm$1.21)	0.302
2,000	3.00 ($\pm$1.27)	0.225
5,000	0.67 ($\pm$0.48)	0.323
10,000	1.52 ($\pm$0.00)	0.060

D. Full Within-Run Correlation Table

Table 11 presents the within-run Pearson correlations between stable rank and memorization fraction for all 18 training runs. Each run has 5 paired observations (steps 6K, 12K, 18K, 24K, 30K). The within-run correlations are strongest for small N ($\leq$ 2,000) where memorization is most severe and where early stopping would be most valuable. For large N ($\geq$ 5,000), the limited dynamic range of memorization fraction within a run reduces the correlation, though the majority of runs still show positive correlations.

Breakdown by Dataset Size. The mean within-run correlation by N is: $N = 500$: $r = +0.832$ (mean of 3 seeds); $N = 1{,}000$: $r = +0.855$; $N = 2{,}000$: $r = +0.509$; $N = 5{,}000$: $r = +0.533$; $N = 10{,}000$: $r = -0.033$; $N = 50{,}000$: $r = -0.138$. The transition from strong positive to near-zero or weak negative within-run correlations mirrors the memorization phase transition: for $N \leq 2{,}000$, memorization develops substantially during training and stable rank tracks this development; for $N \geq 10{,}000$, memorization is mild throughout training and the within-run variation is dominated by noise unrelated to memorization.

The single negative outlier at $N = 10{,}000$, seed 2 ($r = -0.788$) may reflect a run where stable rank decreased for reasons unrelated to memorization (such as unusual optimization dynamics at this seed). With only 5 observations per run, individual outliers of this magnitude are expected.

Table 11: **Full within-run Pearson correlations: stable rank vs. memorization fraction (all 18 runs).** Bold indicates $p < 0.05$. Mean across all runs: $r = +0.426$; 4/18 significant at $p < 0.05$.

N	Seed	Stable Rank r	p
500	0	**+0.935**	**0.020**
500	1	+0.586	0.300
500	2	**+0.975**	**0.005**
1,000	0	**+0.926**	**0.024**
1,000	1	+0.853	0.066
1,000	2	+0.785	0.116
2,000	0	−0.117	0.852
2,000	1	**+0.932**	**0.021**
2,000	2	+0.712	0.177
5,000	0	+0.084	0.893
5,000	1	+0.743	0.150
5,000	2	+0.773	0.126
10,000	0	+0.022	0.972
10,000	1	+0.667	0.219
10,000	2	−0.788	0.113
50,000	0	+0.026	0.966
50,000	1	−0.059	0.925
50,000	2	−0.380	0.528

E. Full Weight-Space Baseline Comparison

Table 1 in the main text presents the four primary metrics. Here we include additional weight-space diagnostics and the composite Spectral Transition Index (STI).

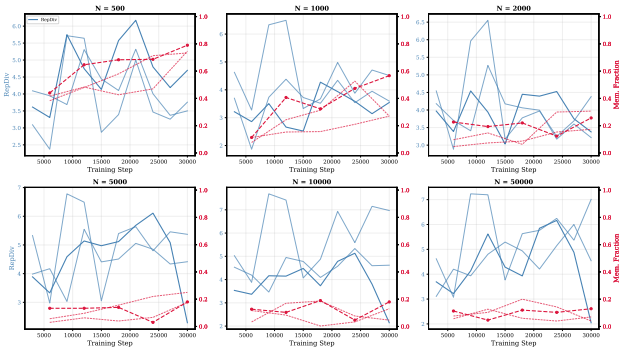

Figure 6: **STI overlay.** The Spectral Transition Index (a composite of weight rank compression and bottleneck RepDiv) is dominated by the weight compression term and adds minimal value over stable rank alone.

Table 12: **Extended weight-space metric comparison.** *Approximate; see Appendix J.

Metric	r	Direction	Notes
Stable Rank	+0.754	Higher=mem.	Best overall
Top-10 SV Ratio	−0.685	Lower=mem.	Complementary
Effective Rank	+0.514	Higher=mem.	Tail-sens.
Power-law α	−0.36*	Flatter=mem.	Modest
CKA (train-test)	−0.034*	Lower=mem.	Weak
RepDiv ($t^* = 500$)	−0.195	Unreliable	High noise

The power-law exponent shows a modest effect: memorizing models ($N = 500$) exhibit $\alpha = 0.356 \pm 0.182$ (flatter spectra), while generalizing models ($N = 50{,}000$) exhibit $\alpha = 0.392 \pm 0.140$ (steeper spectra). The difference (0.036) is directionally consistent with the heavy-tailed self-regularization theory of Martin and Mahoney (2021) but too small to serve as a practical diagnostic.

F. Timestep Sensitivity Details

The choice of evaluation timestep t^* determines the signal-to-noise ratio of the noisy input $\boldsymbol{x}_{t^*}$, which in turn affects the information content of the bottleneck features. Under our linear schedule ($\beta_1 = 10^{-4}$, $\beta_T = 0.02$), the exact cumulative signal retention factors are:

Table 13: **Exact noise schedule values** for timesteps used in experiments.

t	$\bar{\alpha}_t$	$1 - \bar{\alpha}_t$	SNR	Regime
100	0.8951	0.1049	8.54	High-SNR
250	0.5214	0.4786	1.09	Moderate
500	0.0778	0.9222	0.08	Noise-dominated
750	0.0033	0.9967	0.003	Pure noise

At $t^* = 100$ (SNR $= 8.54$), the noisy input retains approximately 90% of the clean image energy. A memorizing model has overfit to individual training images, so training-set inputs elicit highly specialized bottleneck representations while test-set inputs do not, producing large RepDiv. At $t^* = 500$ (SNR $= 0.08$), noise dominates and the memorization-specific signal is washed out. At $t^* = 750$ (SNR $= 0.003$), all models produce similarly uninformative representations, collapsing RepDiv toward zero.

For the main experiment, we also computed a multi-timestep discrimination analysis using two-point comparisons between $N = 500$ and $N = 50{,}000$:

If multi-timestep aggregation is used, it should be weighted toward low-noise timesteps rather than uniformly averaged, as the inclusion of high-noise timesteps dilutes the strong $t^* = 100$ signal.

Table 14: **Multi-timestep REPDIV discrimination (main experiment, $N = 500$ vs. $N = 50{,}000$).** Discrimination ratio decreases with noise level. Multi-timestep aggregation dilutes the strong $t^* = 100$ signal.

t^*	$N = 500$	$N = 50{,}000$	Gap	Ratio
100	9.33	0.31	+9.02	**30.0**×
250	7.61	3.34	+4.27	2.3×
500	4.93	3.40	+1.53	1.4×
750	0.36	0.20	+0.16	1.8×
Multi-timestep aggregation:				
Low+Mid avg	7.13	1.86	+5.27	3.8×
All-timestep avg	5.56	1.82	+3.74	3.1×

G. Training Trajectory for Noise-Level Ablation

Table 15: **Training trajectory for $N = 500$, seed 0 (noise-level ablation, 15K steps).** REPDIV at $t^* = 100$ increases monotonically alongside memorization fraction, while REPDIV at $t^* = 500$ fluctuates.

Step	MemFrac	StableRank	REPDIV $t^*=100$	REPDIV $t^*=500$
5,000	0.450	48.3	6.99	5.44
10,000	0.760	48.9	8.39	6.09
15,000	0.715	49.5	9.52	5.20

REPDIV at $t^* = 100$ increases monotonically from 6.99 to 9.52 as memorization develops, tracking the memorization fraction cleanly. In contrast, REPDIV at $t^* = 500$ fluctuates from 5.44 to 6.09 to 5.20, exhibiting the reversal at step 15,000 that is characteristic of the noise-dominated regime. Stable rank increases modestly from 48.3 to 49.5, consistent with the failure-to-compress interpretation.

H. Sample Size Sensitivity

The feature covariance matrix $\boldsymbol{C}_\ell^{\mathcal{D}} \in \mathbb{R}^{c\times c}$ is estimated from n samples of c-dimensional features. At the bottleneck, $c = 144$. We investigate how the number of samples n affects REPDIV stability for the $N = 500$ condition at $t^* = 500$.

Table 16: **Sample size sensitivity of REPDIV** ($N = 500$, $t^* = 500$). REPDIV is unstable across sample sizes at this noise level, with no monotonic trend.

n	n/c	REPDIV	Regime
50	0.35	0.83	Severely underdetermined
100	0.69	2.70	Underdetermined
200	1.39	0.97	Marginally overdetermined
300	2.08	4.11	Moderately overdetermined

I. Per-Layer Analysis

Table 17: **Normalized effective rank by layer.** Effective rank divided by maximum attainable rank controls for layer dimensionality. The bottleneck gap of 0.024 persists after normalization.

Block	$N = 500$		$N = 50{,}000$		Δ normER
	ER	normER	ER	normER	
conv_in	24.3	0.902	24.2	0.896	0.006
enc. early	64.4	0.936	64.3	0.936	0.000
enc. mid	119.1	0.928	117.8	0.920	0.008
bottleneck	**135.8**	**0.943**	**132.3**	**0.919**	**0.024**
dec. early	101.8	0.961	99.6	0.942	0.019
dec. late	55.6	0.961	55.4	0.963	−0.002
conv_out	3.0	0.999	3.0	0.999	0.000

The decoder_early layer shows the second-largest gap (0.019), which is expected given the U-Net's skip connections: the decoder_early layer receives concatenated features from both the encoder_mid and bottleneck paths, inheriting some of the bottleneck's memorization sensitivity. The encoder_early (0.000) and conv_in/conv_out layers (0.006 and 0.000 respectively) show negligible gaps, confirming that memorization-related spectral changes are not uniformly distributed across the network.

J. CKA and Power-Law Analysis

We compute linear CKA (Kornblith et al., 2019) between training and test bottleneck features as a representation-space metric invariant to orthogonal transformations and isotropic scaling. At $N = 500$ (memorizing), CKA $= 0.104$; at $N = 50{,}000$ (generalizing), CKA $= 0.138$. The positive trend with dataset size confirms that generalizing models produce more aligned train-test representations. However, the absolute difference (0.034) is small, and both values are well below 1.0, indicating that even generalizing models process train and test data quite differently at the bottleneck under noisy-input evaluation.

For the power-law analysis, we fit exponents to the weight singular value spectra following Martin and Mahoney (2021). Memorizing models ($N = 500$) exhibit flatter spectra with $\alpha = 0.356 \pm 0.182$, while generalizing models ($N = 50{,}000$) exhibit steeper spectra with $\alpha = 0.392 \pm 0.140$. The difference (0.036) is modest but directionally consistent with the heavy-tailed self-regularization theory.

K. Noise Averaging for REPDIV

One potential mitigation for REPDIV's noise sensitivity at $t^* = 500$ is to average over multiple independent noise draws. We test this by computing REPDIV with 5 inde-

Table 18: **CKA and power-law exponents.**

Metric	$N = 500$	$N = 50{,}000$
Linear CKA (train-test)	0.104	0.138
Power-law α	0.356 ± 0.182	0.392 ± 0.140

pendent noise realizations at $t^* = 500$ and averaging the resulting feature matrices before computing the covariance.

Table 19: **Noise averaging (5 draws) for REPDIV at $t^* = 500$.** Averaging reduces variance slightly but the coefficient of variation (CV) remains $\sim$0.23, far too high for a reliable diagnostic.

N	REPDIV (5-draw avg)	Std	CV
500	5.07	1.19	0.23
50,000	5.43	1.21	0.22
Gap		0.36	

Noise averaging does not reduce REPDIV variance enough to become diagnostic at $t^* = 500$. The coefficient of variation remains $\sim$0.23 for both conditions, and the gap is only 0.36. This indicates that the instability at mid-to-high noise arises from fundamental limitations (the noise injection destroys the memorization-specific signal) rather than addressable sampling variance. By contrast, choosing $t^* = 100$ recovers a strong signal without noise averaging (Section 4.3), confirming that the noise level, not sampling variance, is the critical factor.

L. Preliminary Cross-Dataset Evidence

To provide initial evidence for generalization beyond CIFAR-10, we trained a single model on Fashion-MNIST (28×28, resized to 32×32) with $N = 200$ training examples using the same architecture and hyperparameters. At step 15,000, we observe: memorization fraction $= 0.000$ (zero training images reproduced), stable rank $= 8.3$, and REPDIV at $t^* = 500 = 0.90$. The low stable rank value (8.3, compared to $\sim$49 for CIFAR-10 $N = 500$) is consistent with the lower memorization level, though the different dataset characteristics (grayscale, lower visual complexity) prevent direct quantitative comparison. This single data point is consistent with, but far from sufficient to establish, cross-dataset generalization of the spectral signatures. Comprehensive cross-dataset validation remains the primary direction for future work (Section 6).

M. Reproducibility Details

Code and Data. Code to reproduce all experiments will be released upon acceptance. All experiments use CIFAR-10 (Krizhevsky, 2009), which is publicly available.

Architecture Details. The U-Net has: 3 resolution levels with base channel multiplier $\mathrm{ch} = 48$ and channel multipliers $[1, 2, 3]$, yielding channel dimensions $[48, 96, 144]$; 2 residual blocks per level; sinusoidal timestep embeddings of dimension 256; group normalization with 32 groups; attention at the 8×8 resolution level. Total parameters: approximately 2.7M.

Feature extraction dimensions: early $[B, 96, 32, 32]$, bottleneck $[B, 144, 8, 8]$, late $[B, 48, 32, 32]$. After spatial average pooling, these become 96-, 144-, and 48-dimensional vectors respectively.

Training Hyperparameters. Optimizer: Adam (Kingma and Ba, 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$); learning rate: 2×10^{-4} (constant); batch size: 128; training steps: 30,000 (main experiment), 15,000 (noise-level ablation); noise schedule: linear, $\beta_1 = 10^{-4}$, $\beta_T = 0.02$, $T = 1000$; EMA: not used; data augmentation: random horizontal flip only.

Weight Matrix Reshaping. Convolutional weight tensors of shape $[c_{\text{out}}, c_{\text{in}}, k_h, k_w]$ are reshaped to $[c_{\text{out}}, c_{\text{in}} \cdot k_h \cdot k_w]$ before SVD. Linear layer weights are used as-is. Aggregate metrics (stable rank, effective rank, top-10 SV ratio) are parameter-count-weighted means across all convolutional and linear layers. Bias parameters are excluded.

Spectral Diagnostic Details. Default evaluation timestep: $t^* = 500$ for main experiments; $t^* \in \{100, 500\}$ for noise-level ablation; $t^* \in \{100, 250, 500, 750\}$ for multi-timestep discrimination analysis. Spatial averaging: global average pooling. SVD: full SVD via PyTorch `torch.linalg.svd`. Effective rank: computed via Eq. 2 with numerical stability clipping at $p_i > 10^{-10}$, using singular values following Roy and Vetterli (2007). Stable rank: computed via Eq. 3. Number of samples for covariance estimation: 300 (randomly sampled, fixed within each run). Evaluation frequency: every 3,000 steps for spectral metrics, every 6,000 steps for memorization metrics.

Memorization Metric Details. Generated samples: 1,024 per evaluation. Sampler: DDIM (Song et al., 2021a) with 50 steps. Distance metric: ℓ_2 in pixel space. Memorization threshold: 5th percentile of pairwise inter-training distances. Nearest-neighbor computation: exact brute-force. Nearest-neighbor ratio: mean $d_{\text{train}}/d_{\text{test}}$ over generated samples. Authenticity: fraction of generated samples with $d_{\text{test}} < d_{\text{train}}$.

Noise-Level Ablation Details. The ablation used $N \in \{500, 1000, 2000, 5000, 10{,}000\}$ (5 dataset sizes, omitting $N = 50{,}000$ for compute efficiency). Models were trained for 15,000 steps with the same architecture

and hyperparameters as the main experiment. REPDIV was computed at $t^* = 100$ and $t^* = 500$ simultaneously on each checkpoint. The Ledoit-Wolf variant used `sklearn.covariance.LedoitWolf` with automatic shrinkage parameter selection. Total: 13 paired observations across conditions and seeds.

Noise Schedule Calculation. The exact $\bar{\alpha}_t$ values reported in Section 2 are computed as $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ where $\beta_s = \beta_1 + \frac{s-1}{T-1}(\beta_T - \beta_1)$ with $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$. These values were verified numerically: $\bar{\alpha}_{100} = 0.8951$, $\bar{\alpha}_{250} = 0.5214$, $\bar{\alpha}_{500} = 0.0778$, $\bar{\alpha}_{750} = 0.0033$. The signal-to-noise ratio is $\mathrm{SNR}(t) = \bar{\alpha}_t/(1-\bar{\alpha}_t)$.

Compute Resources. All experiments were conducted on a single NVIDIA A100-40GB GPU. Total compute: 7.3 GPU-hours for all 18 runs (main experiment). Noise-level ablation and additional analyses used approximately 2.0 additional GPU-hours.

Random Seeds. Seeds 0, 1, 2 were used for all conditions. Seeds control model initialization, data sampling order, noise injection during training, and noise injection during spectral evaluation.