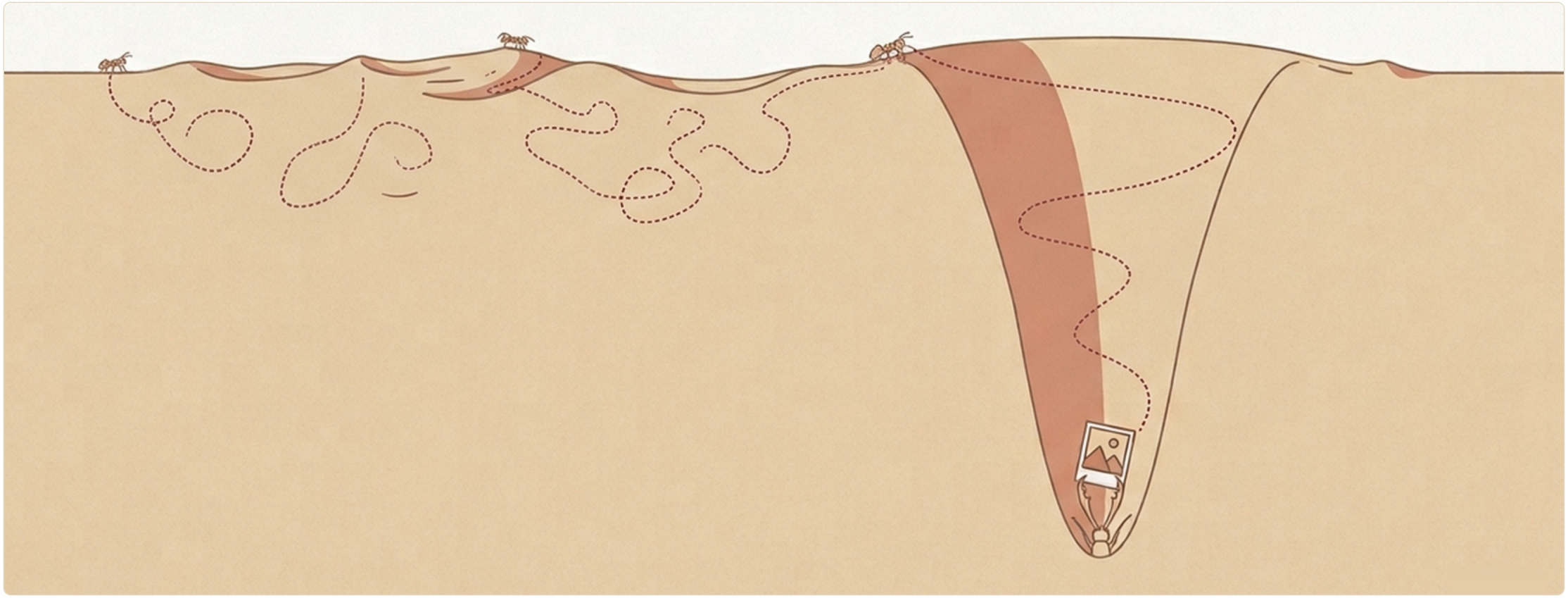
NEURIPS 2026

Cyclic Denoising Reveals Ultrastable Memories in Diffusion Models

Rishabh Sharma · Stefano Martiniani

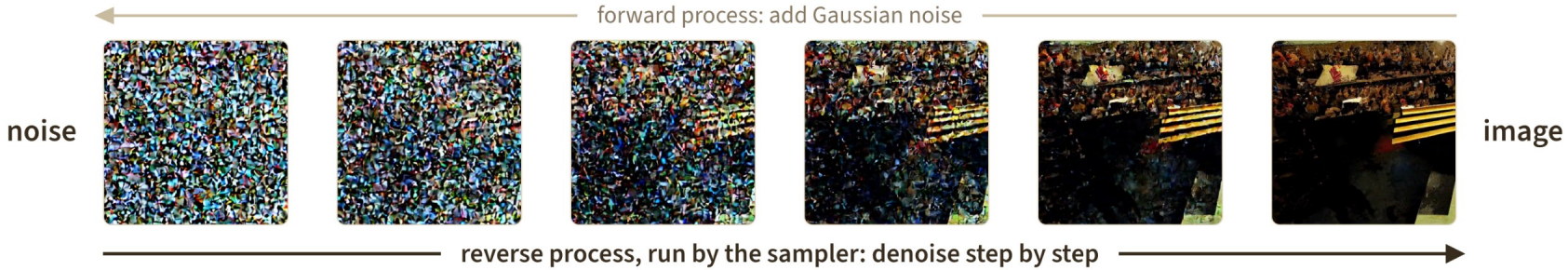
NEW YORK UNIVERSITY

SIMONS FOUNDATION



THE PROBLEM · AND THIS WORK

Can a model's own dynamics reveal what it has memorized?



Diffusion models are trained on web-scale data and can memorize individual training images, raising concerns about **privacy, copyright and intellectual property**.

PRIOR GENERATE-AND-FILTER ATTACKS

Prompts from known or suspected training captions

Large-scale generation

Post-hoc similarity search or membership inference

THIS WORK · CYCLIC DENOISING

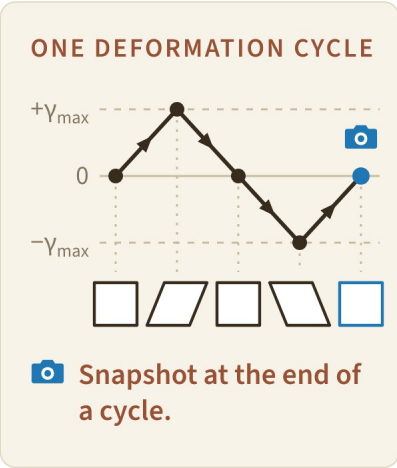
Dynamical probe of the learned generative landscape

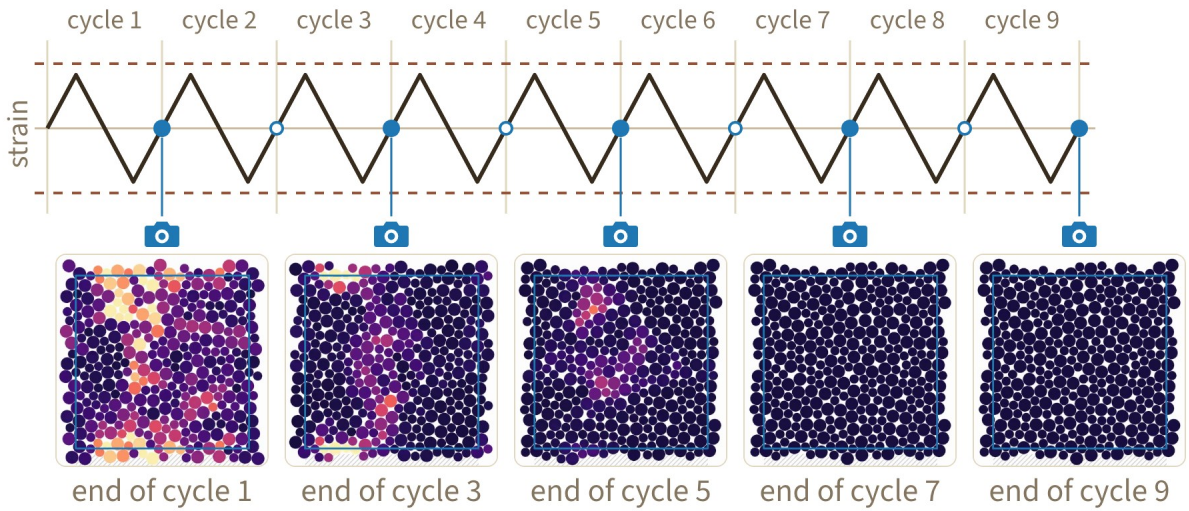
Extraction attack with **sampler-level control** and no prior knowledge of training data, captions or prompts

Exposes regions of the learned distribution that remain largely inaccessible to standard sampling

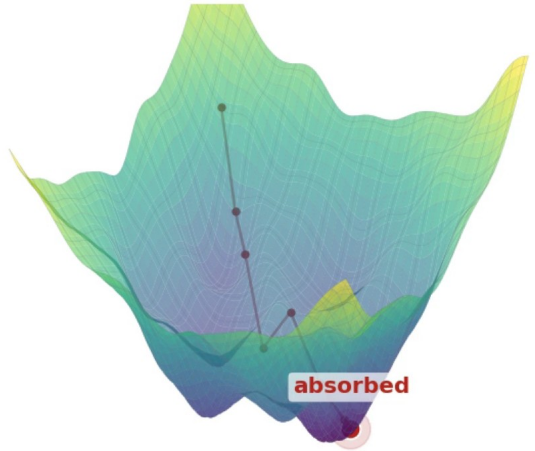
INSPIRATION · CYCLIC SHEAR OF A DISORDERED SOLID

Cyclically driving a disordered system probes its energy landscape





color: motion during the cycle none rearranging yield amplitude

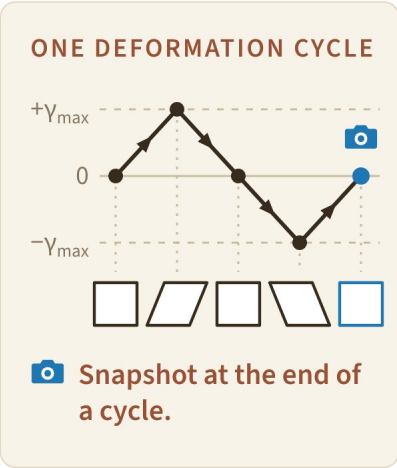


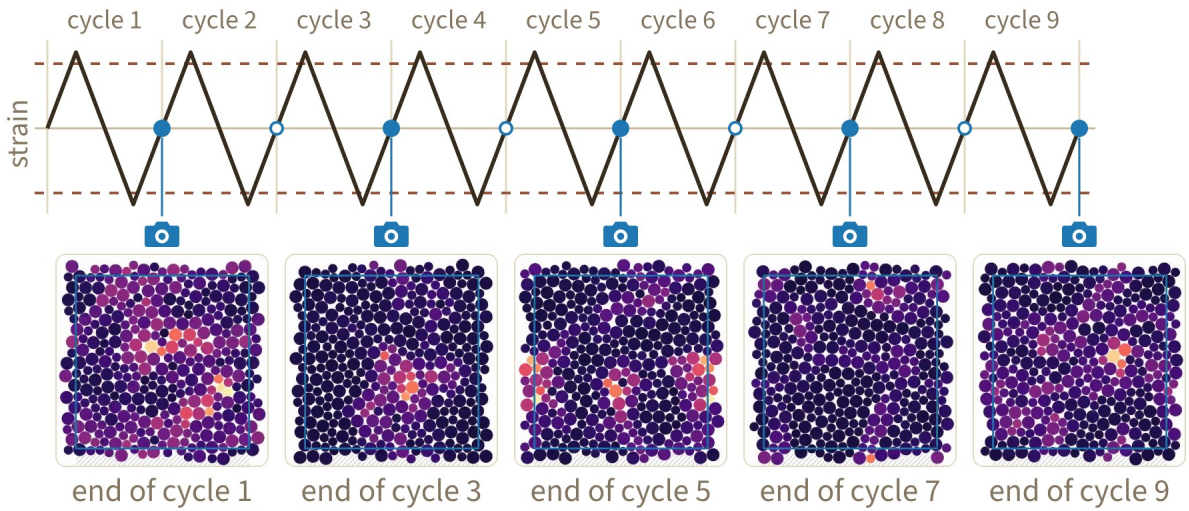
energy landscape, one dot per cycle

SMALL AMPLITUDE **Stuck:** the solid settles into an absorbing state; the cycle-to-cycle change goes to zero.

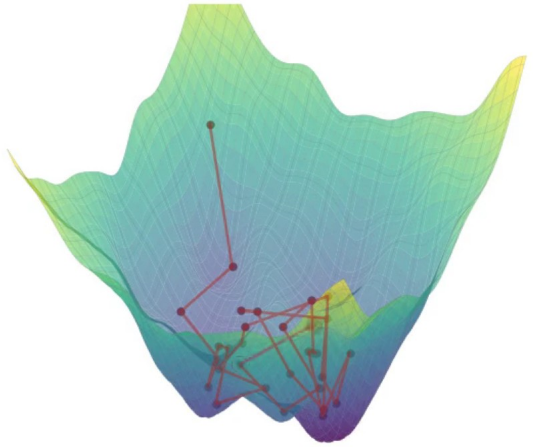
INSPIRATION · CYCLIC SHEAR OF A DISORDERED SOLID

Cyclically driving a disordered system probes its energy landscape





color: motion during the cycle none rearranging yield amplitude

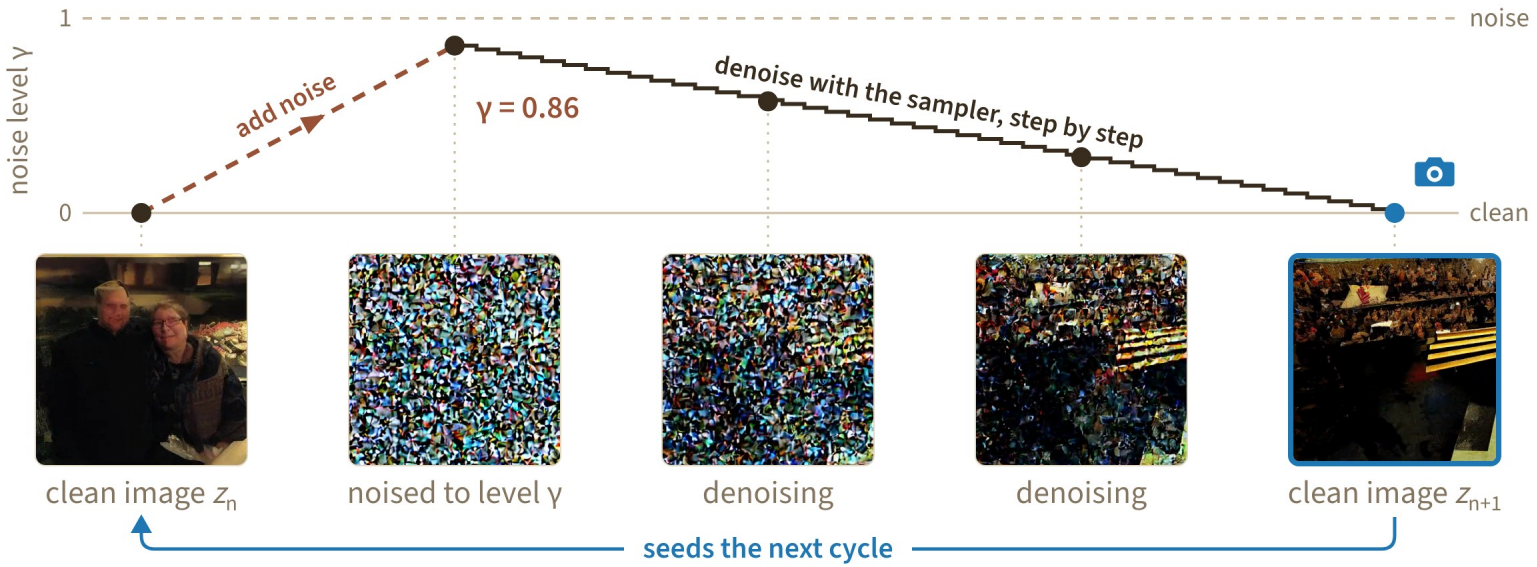


energy landscape, one dot per cycle

LARGE AMPLITUDE **Exploratory:** the solid yields and keeps moving through its energy landscape.

METHOD · ONE CYCLE

Cyclic denoising: add noise to a level γ, denoise back, repeat



ONE DENOISING CYCLE

$0 \rightarrow \gamma \rightarrow 0$

Add noise up to the level γ, then run the sampler back to a clean image.

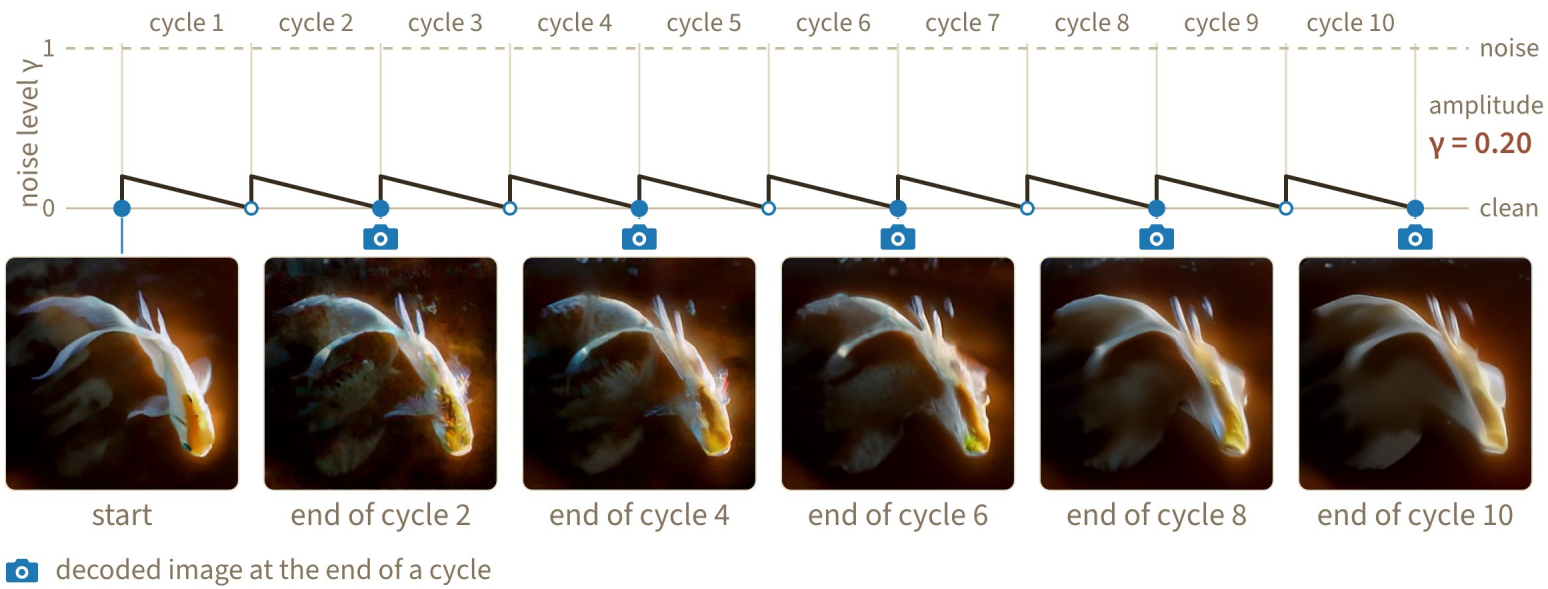
Snapshot at the end of a cycle.

Cycling amplitude $\gamma \in [0, 1]$: the fraction of the inference trajectory traversed per cycle; it sets how much noise is added before denoising.

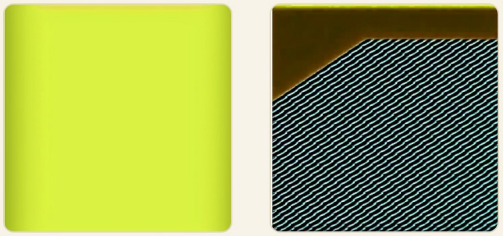
Stable Diffusion v1.4, unconditional; DDPM sampler with fresh noise at every step, 50 inference steps.

METHOD · REPEATED CYCLES

Cyclic denoising: add noise to a level γ, denoise back, repeat



TRIVIAL ABSORBING STATES



fixed point
γ = 0.2

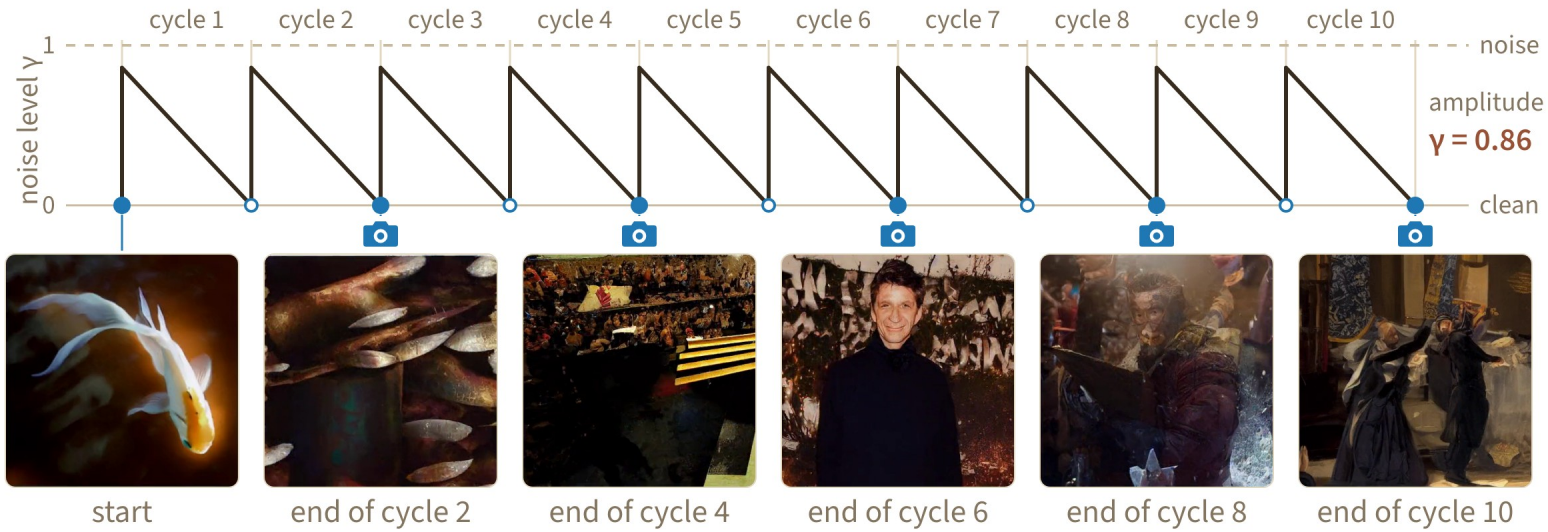
limit cycle
γ = 0.1

after 10,000 cycles

SMALL AMPLITUDE Each cycle changes only fine details, so the image drifts slowly into a **trivial absorbing state**.

METHOD · REPEATED CYCLES

Cyclic denoising: add noise to a level γ, denoise back, repeat



STROBOSCOPIC TRAJECTORY

$z_0 \to z_1 \to z_2 \to \cdots$

The clean state at the end of each cycle.

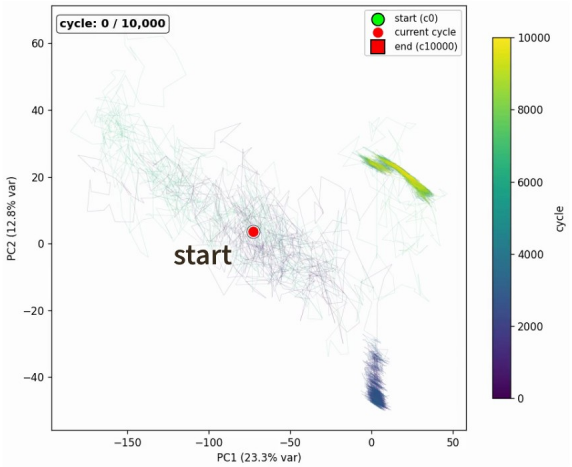
ORDER PARAMETER

$\cos(z_n, z_{n-1})$

1 if the image returns, **near 0 if it is replaced**.

LARGE AMPLITUDE The dynamics defines a **stroboscopic trajectory** on the generative landscape. The **order parameter** tracks its cycle-to-cycle change.

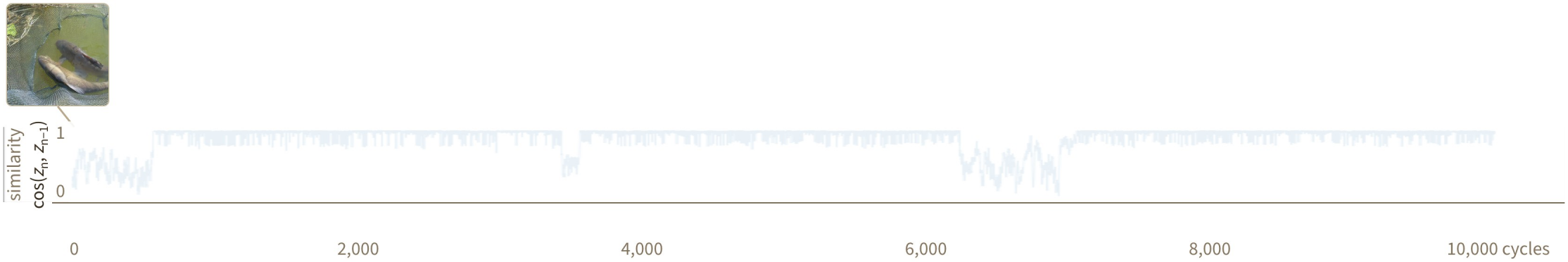
The trajectory gets trapped in memorized training images



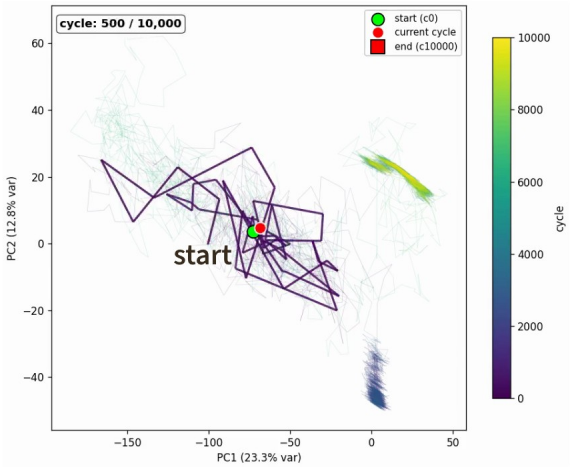
latent trajectory, 2D projection (PCA)



○ **start** cycle 0, an ImageNet photograph



The trajectory gets trapped in memorized training images

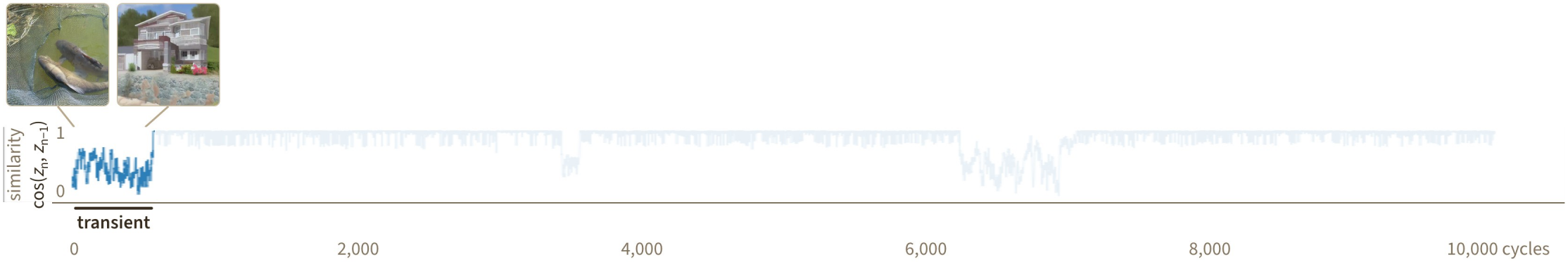


latent trajectory, 2D projection (PCA)

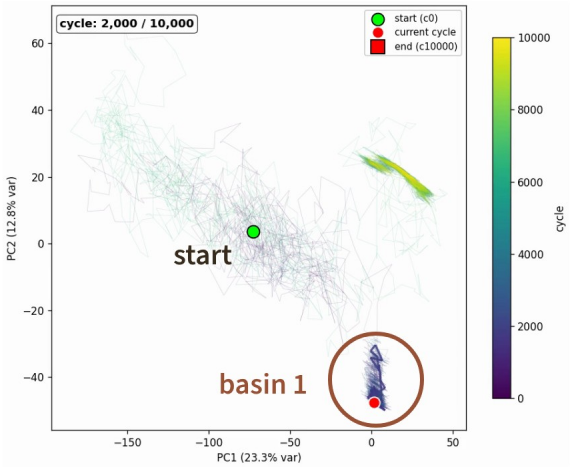


start cycle 0, an ImageNet photograph

transient about 550 cycles



The trajectory gets trapped in memorized training images



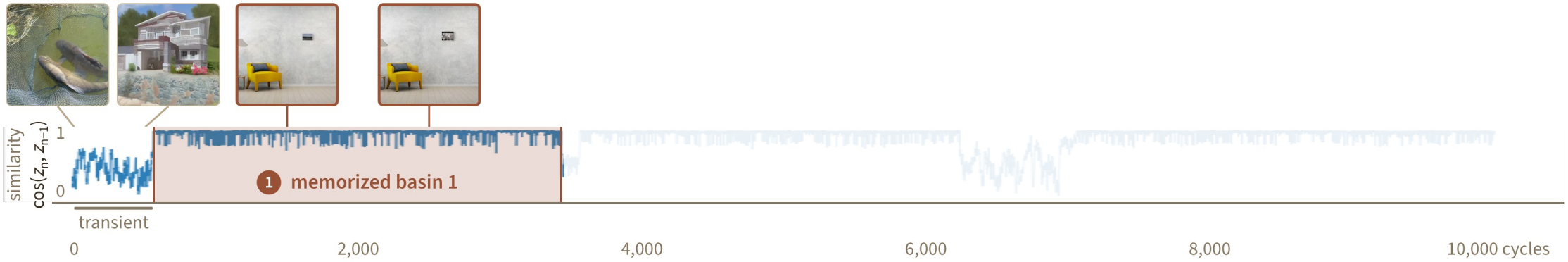
latent trajectory, 2D projection (PCA)



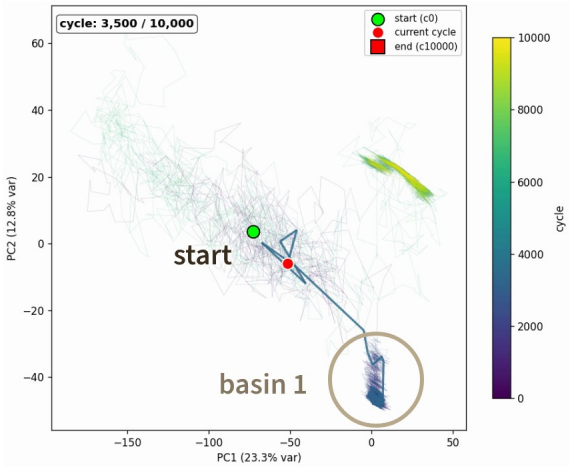
start cycle 0, an ImageNet photograph

transient about 550 cycles

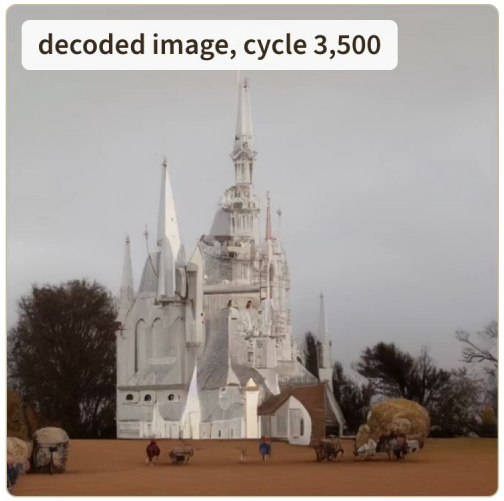
1 **memorized basin 1** about 2,900 cycles



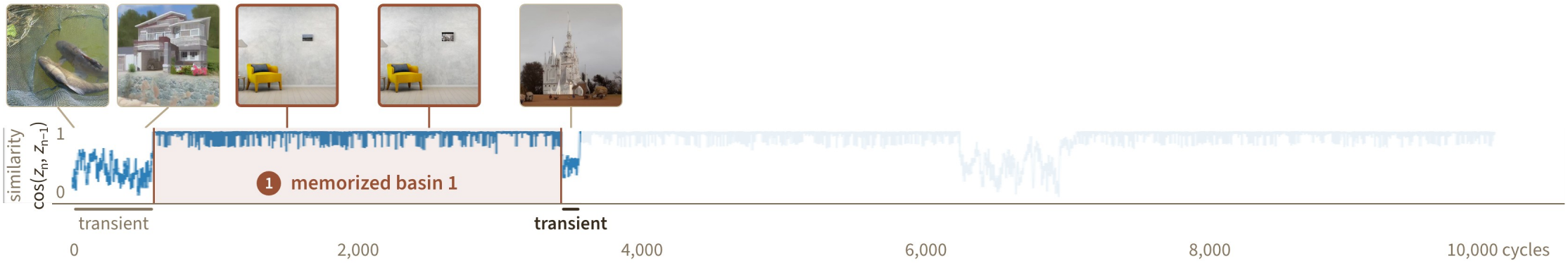
The trajectory gets trapped in memorized training images



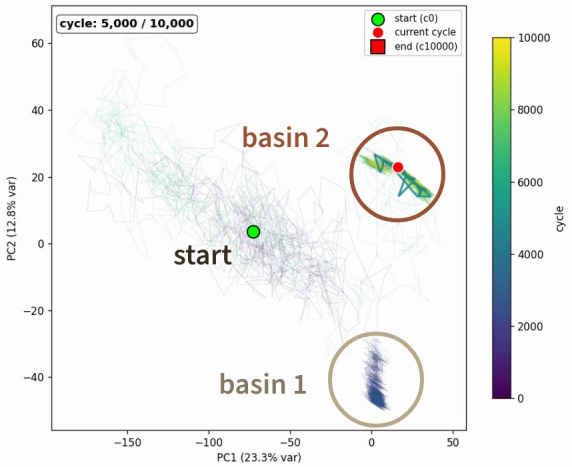
latent trajectory, 2D projection (PCA)



- start cycle 0, an ImageNet photograph
- transient about 550 cycles
- **1 memorized basin 1** about 2,900 cycles
- **transient** about 125 cycles



The trajectory gets trapped in memorized training images



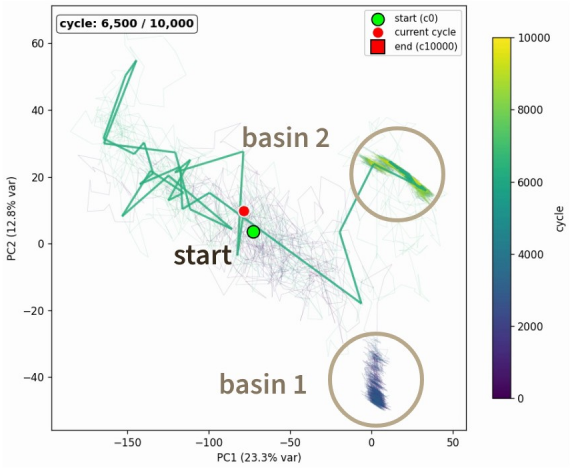
latent trajectory, 2D projection (PCA)



- start cycle 0, an ImageNet photograph
- transient about 550 cycles
- **1 memorized basin 1** about 2,900 cycles
- transient about 125 cycles
- **2 memorized basin 2** about 2,700 cycles



The trajectory gets trapped in memorized training images



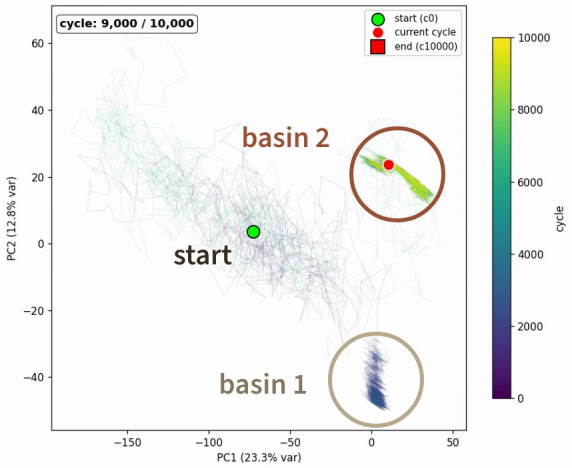
latent trajectory, 2D projection (PCA)



- start cycle 0, an ImageNet photograph
- transient about 550 cycles
- **1 memorized basin 1** about 2,900 cycles
- transient about 125 cycles
- **2 memorized basin 2** about 2,700 cycles
- **transient** about 800 cycles



The trajectory gets trapped in memorized training images



latent trajectory, 2D projection (PCA)

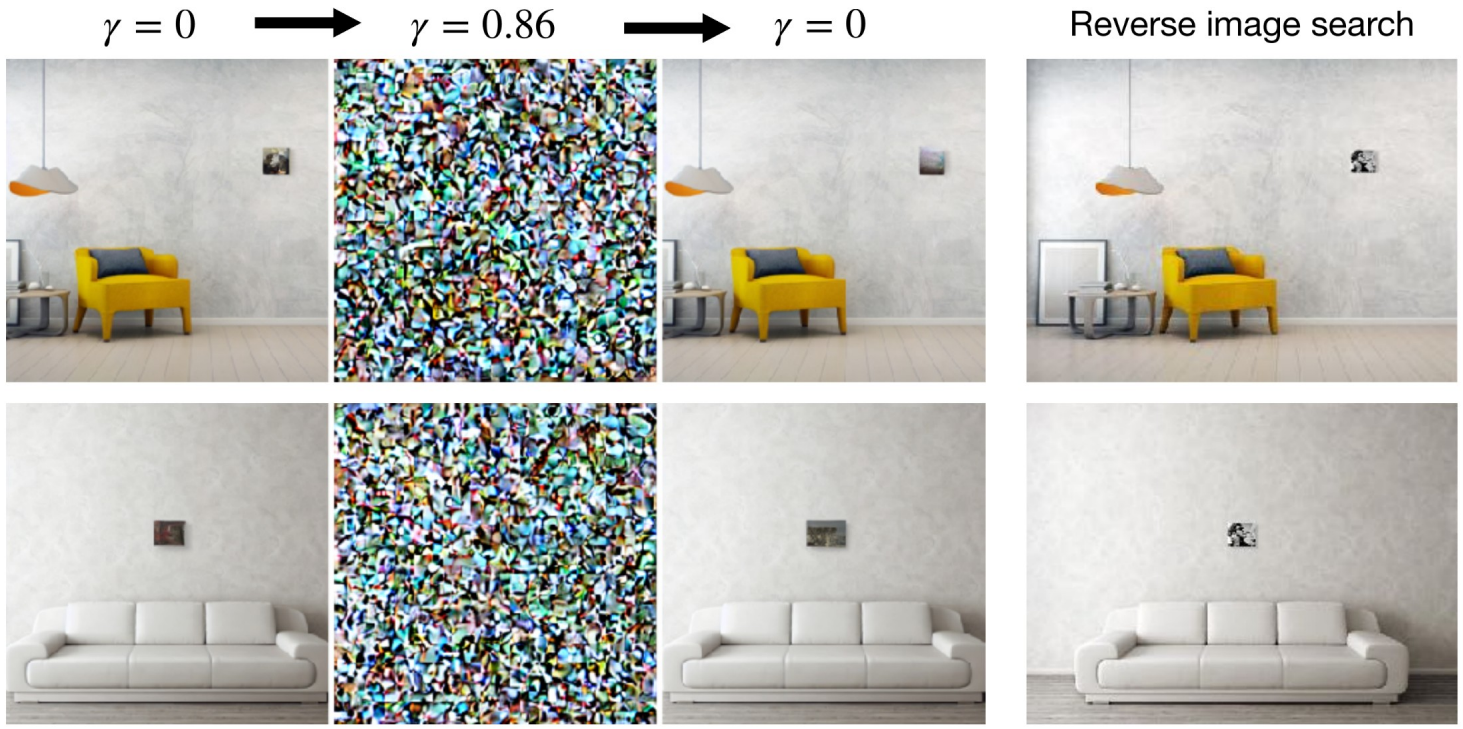


- start cycle 0, an ImageNet photograph
- transient about 550 cycles
- **1 memorized basin 1** about 2,900 cycles
- transient about 125 cycles
- **2 memorized basin 2** about 2,700 cycles
- transient about 800 cycles
- **3 memorized basin 2, revisited** the last 2,900 cycles



EXTRACTION · ONE CYCLE FROM EACH BASIN, AND THE CLOSEST WEB IMAGE

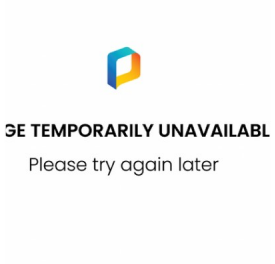
The basins are memorized training templates



Stock templates duplicated across the web, found with **no prompt** and **no access to the training set**.

EXTRACTION · THE AMPLITUDE ACTS AS A STABILITY FILTER

Different amplitudes expose different memories

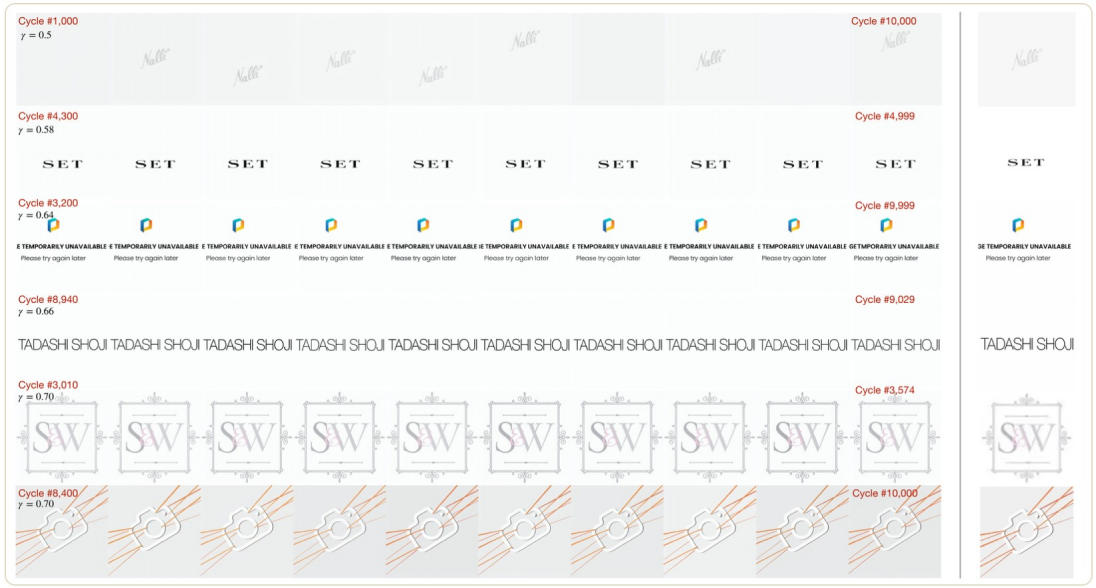
	brand logo	placeholder	brand logo	backpack	golf bag	tote bag
RECOVERED BY CYCLING						
CLOSEST WEB IMAGE						
	γ = 0.58	γ = 0.64	γ = 0.70	γ = 0.72	γ = 0.84	γ = 0.86

Moderate amplitude: logos, placeholders → **High amplitude:** product photographs

EXTRACTION · ATTRACTORS RECOVERED IN STABLE DIFFUSION V1.4 WITH NO PROMPT

Recovered attractors at moderate and high amplitudes

MODERATE AMPLITUDE γ = 0.50 to 0.70



Brand logos, placeholders and web-crawl artifacts

HIGH AMPLITUDE γ = 0.72 to 0.88

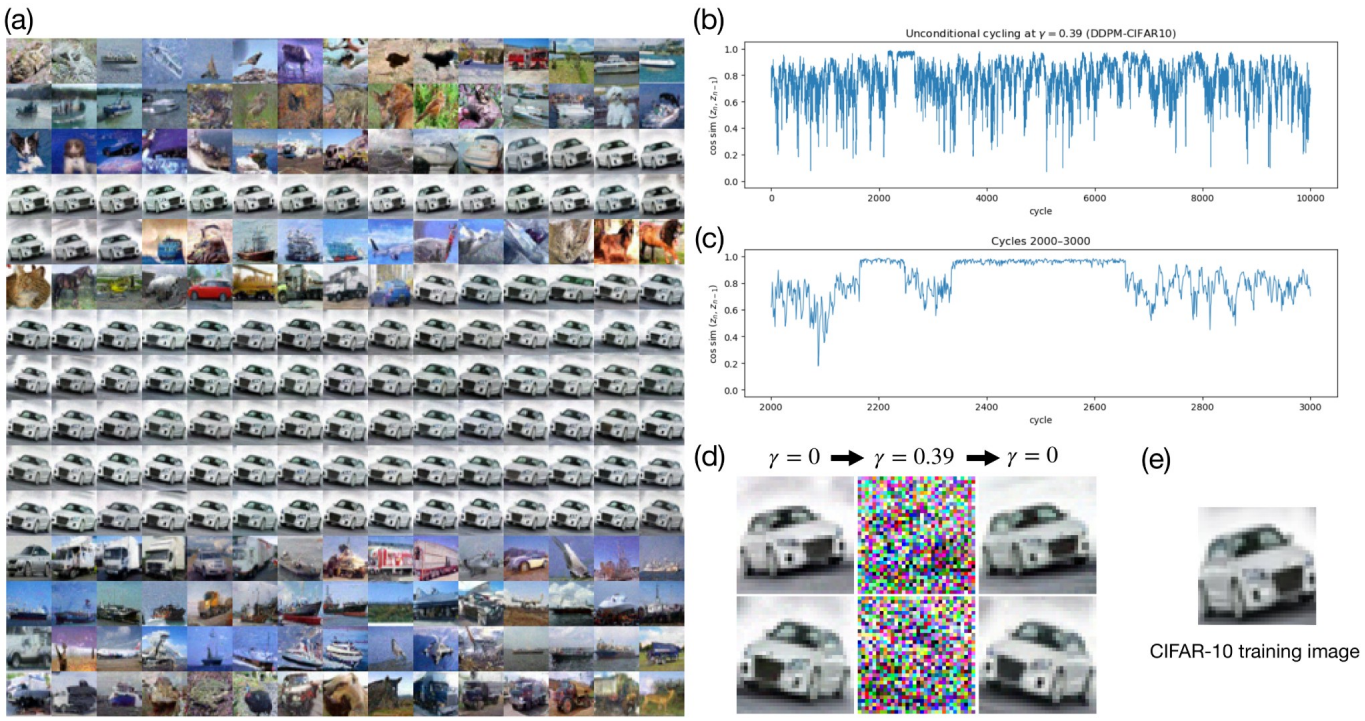


Product photographs and room templates

Each row: ten snapshots of one trajectory while it stays in one attractor. Last column: the closest web image.

EXTRACTION · CIFAR-10 DDPM, AMPLITUDE 0.39

The same dynamics in a pixel-space model

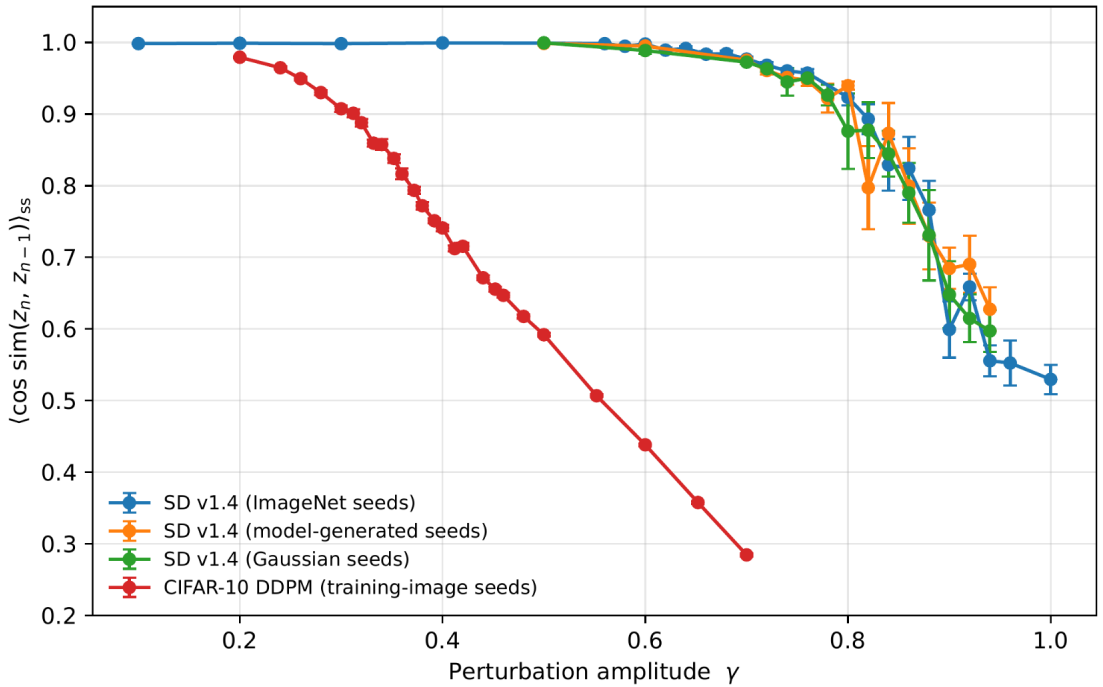


The car is a training image independently flagged as memorized, recovered here **from the model dynamics alone**.

Model: huggingface.co/google/ddpm-cifar10-32, an unconditional pixel-space DDPM

STABILITY · STEADY-STATE SIMILARITY AGAINST AMPLITUDE

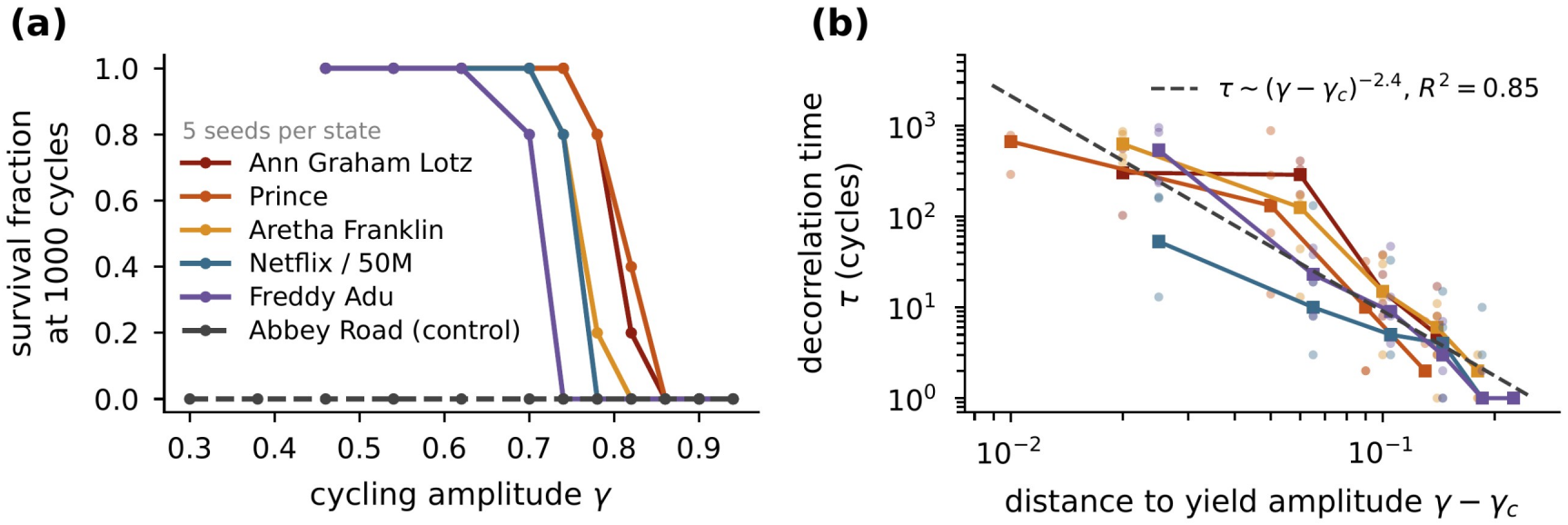
Like the sheared solid, the diffusion model also yields



Increasing the amplitude takes the model from a **stuck** phase to an **exploratory** phase.

The CIFAR-10 model (red) yields at a lower γ, in part because its noise schedule differs; γ is comparable only within a model.

Stability under cyclic denoising as a memorization diagnostic



(a) **Known memorized images**, cycled without prompts, yield at their own yield amplitudes (0.72 to 0.81).

(b) Signatures of **critical slowing down**: the decorrelation time shows a power-law divergence.

STABILITY · 98 MEMORIZED IMAGES AGAINST 105 POST-TRAINING CONTROLS

Stability under cyclic denoising as a memorization diagnostic

MEMORIZED SET

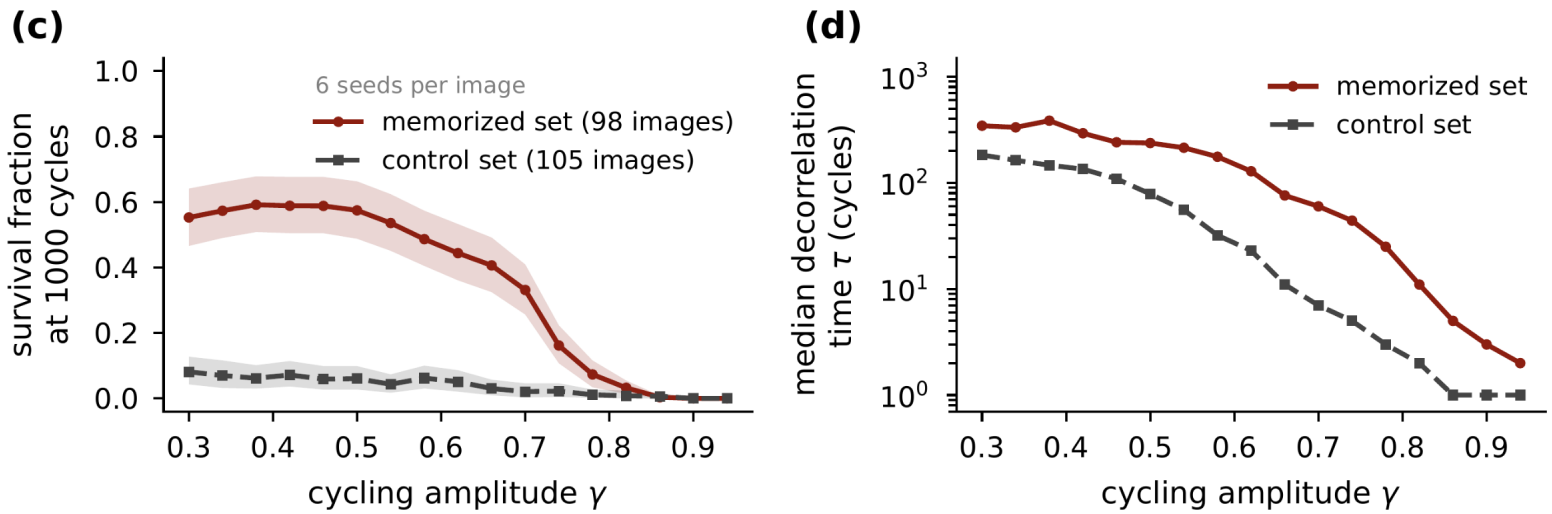
62 of 98

resolve a yield amplitude; median 0.69

CONTROL SET

9 of 105

resolve a yield amplitude; median 0.47



A memorization test for candidate images: each image is encoded and cycled with no prompt, and its survival is measured.

17 amplitudes, six seeds, 1,000 cycles each.

SUMMARY

A model's own dynamics reveal what it has memorized

- **Cyclic denoising** emerges as both a probe of the generative landscape and a physics-inspired extraction attack for diffusion models.
- **The deepest attractors are ultrastable:** regenerated from near-total corruption and sustained through thousands of noising-denoising cycles. **Many correspond to memorized training images.**
- **An extraction attack from the dynamics alone:** no prompts, no captions, no training data.
- **Stability under cycling is a memorization diagnostic:** a prompt-free test for candidate images.
- **Standard sampling may substantially underestimate memorization risk**, because some training examples are hidden in deep basins that are rarely visited unless the model is driven.

Thank you

Funded by the Simons Foundation, grant MPS-T-MPS-00839534.
We also thank NYU High Performance Computing and Prof. David J. Heeger for discussions.

NEW YORK UNIVERSITY

SIMONS FOUNDATION

Movies and project page
rishabh-tifr.github.io/cyclic-denoising

