FROM MODES TO MEMORIES: CHARACTERIZING THE SCALE-SPACE DYNAMICS OF DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models are typically viewed as stochastic processes that transform noise into data. We take a complementary perspective: a diffusion model defines a family of deterministic dynamical systems indexed by noise scale. At each fixed scale σ, we treat the denoiser as a self-map and study its dynamics. For an exact denoiser, fixed points correspond to critical points of the smoothed data density, while attractors correspond to its modes; as σ increases, sample-level modes merge into progressively coarser ones. This suggests a geometric view of memorization: examples that receive excess probability mass due to duplication or overfitting, as well as outliers, should remain distinguishable under stronger smoothing than ordinary examples. We quantify this persistence by the critical scale σ_c, the largest noise scale at which an example is retained by the fixed-scale dynamics. In conditional models, the same construction extends naturally to image–caption pairs. Experiments in controlled settings and on large-scale models show that σ_c tracks memorization arising from duplication, overfitting, and outliers, and identifies both memorized and partially memorized examples in Stable Diffusion. Moreover, σ_c yields interpretable measures of the image spatial distribution and caption dependence of memorization.

1 INTRODUCTION

Diffusion models are typically viewed as stochastic processes that transform noise into data (Ho et al., 2020; Song et al., 2021; Karras et al., 2022; Rombach et al., 2022). They are trained with a local and static criterion, the prediction of a clean signal from a noisy one at each noise level, and are used through a global and dynamic procedure, a sampler that composes these denoisers along a decreasing schedule of noise levels. What a model has learned is then read off the process as a whole, from its samples or from its loss along the schedule.

We take a complementary view and interpret a diffusion model as a family of denoisers D_σ, one for each noise level σ. Fixing σ and iterating the denoiser on its own output, $x_{k+1} = D_\sigma(x_k)$, turns a single trained network into a one-parameter *family of deterministic dynamical systems*, whose fixed points and basins are properties of the network alone (Figure 1, left). For an exact denoiser each of these maps is Gaussian mean shift on the smoothed data density p_σ(Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 2002): its *fixed points* are the critical points of p_σ and its *attractors* are the modes of p_σ. At small σ every training example is its own attractor, at large σ a single attractor remains, and in between the attractors merge into progressively coarser ones (Figures 2–3). Varying σ thus traces a scale space of the attractors of the model.

This scale space gives a geometric view of memorization mechanisms (Carlini et al., 2023; Bonnaire et al., 2025; Ross et al., 2025): an example is retained under its dynamics when it carries excess probability mass, as under duplication or overfitting, or when it is isolated from the remaining data, as outliers. We measure this persistence by the *critical scale* σ_c, the largest noise level at which iterating the fixed-scale denoiser from an example keeps the trajectory close to it (Figure 1). The critical scale requires no labels, applies to conditional models, where it becomes a property of a caption-image pair, and is interpretable. We test it with interventions that cause memorization: on `CIFAR-10` we induce memorization by duplication, by overfitting and by planting outliers, and on Stable Diffusion we use image–caption pairs that are known to be memorized, fully or in part. Our contributions are as follows:

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.

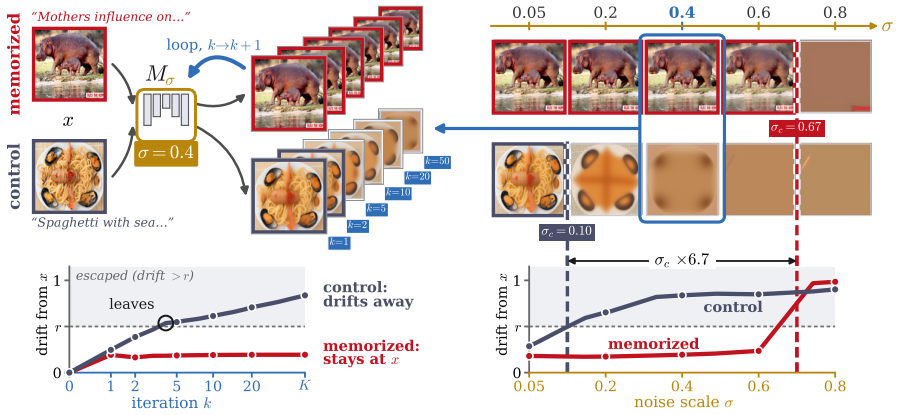

Figure 1: A diffusion model as a family of dynamical systems. Fixing the noise level σ and feeding the denoiser its own output defines a self map M_σ (*left*). A memorized training image is a *fixed point* of this map: its orbit stays at the image, whereas the orbit of a control image drifts away and crosses the escape radius r (*bottom left*). Repeating this retention test at every σ (*right*) yields the critical scale σ_c, the largest noise level at which the image is still returned; it is $6.7\times$ larger for the memorized image, effectively detecting it.

- We show that a diffusion model can be interpreted as a *family of dynamical systems* indexed by the noise level: for an exact denoiser the attractors are the modes of the data density smoothed at scale σ, and varying σ traces a scale space in which these attractors merge.
- In this framework, we define the *critical scale* σ_c as a memorization measure: the scale up to which a sample remains an *attractor* of the dynamics. It extends to conditional models, capturing the caption-image interaction. For the optimal case we prove that σ_c depends on the inter-samples distances and their multiplicity, reflecting memorization induced by data duplication and outliers.
- In controlled experiments we show that σ_c detects memorization induced by duplication, by overfitting and by planting outliers, transferring the mechanism from the theory to trained networks.
- We show that the critical scale is an effective detector on large-scale conditional models such as Stable Diffusion, for both fully and partially memorized examples and that the measures is *interpretable*: being able to characterize whether its retention comes from the caption, from the image, or their interaction and where in the image the memorized content localizes.

2 Diffusion models as a family of dynamical systems

Let $x_0 \in \mathbb{R}^d$ be drawn from a data distribution p, and let $y = x_0 + \sigma\varepsilon$, with $\varepsilon \sim \mathcal{N}(0, I)$, be its observation at noise level $\sigma > 0$. Following Karras et al. (2022), a diffusion model is a family of denoisers $\{D_\sigma : \mathbb{R}^d \to \mathbb{R}^d\}_{\sigma>0}$ trained by minimising $\mathbb{E}\,\|D_\sigma(x_0+\sigma\varepsilon) - x_0\|^2$ over a range of noise levels. Its minimiser is the posterior mean $D_\sigma(y) = \mathbb{E}[x_0 \mid y]$. Noise-prediction models such as DDPM (Ho et al., 2020) and Stable Diffusion (Rombach et al., 2022) reduce to this form: a network ϵ_θ trained to predict ε gives $D_\sigma(y) = y - \sigma\epsilon_\theta(y)$, up to a rescaling of the input (see Appendix F).

We study each denoiser by fixing its noise level and iterating it on its own output, so that each noise level defines a deterministic dynamical system on $\mathbb{R}^d$ (Figure 1, left).

Definition 1 (Denoising dynamics at scale σ)**.** Fix $\sigma > 0$ and let $M_\sigma := D_\sigma$. The *denoising dynamics at scale* σ is the deterministic dynamical system on $\mathbb{R}^d$

$$x_{k+1} = M_\sigma(x_k), \qquad k = 0, 1, 2, \ldots$$

For $x \in \mathbb{R}^d$, the *orbit* of x is the sequence $(M_\sigma^k(x))_{k\geq 0}$, where M_σ^k denotes the k times composition of M_σ. A point $x^* \in \mathbb{R}^d$ is a *fixed point* if $M_\sigma(x^*) = x^*$. Its *basin of attraction* is

$$\mathcal{B}_\sigma(x^*) = \{x \in \mathbb{R}^d : \lim_{k\to\infty} M_\sigma^k(x) = x^*\},$$

and x^* is an *attractor* if $\mathcal{B}_\sigma(x^*)$ contains an open neighbourhood of x^*. Letting σ vary, a diffusion model defines the one-parameter family of dynamical systems $\{M_\sigma\}_{\sigma>0}$.

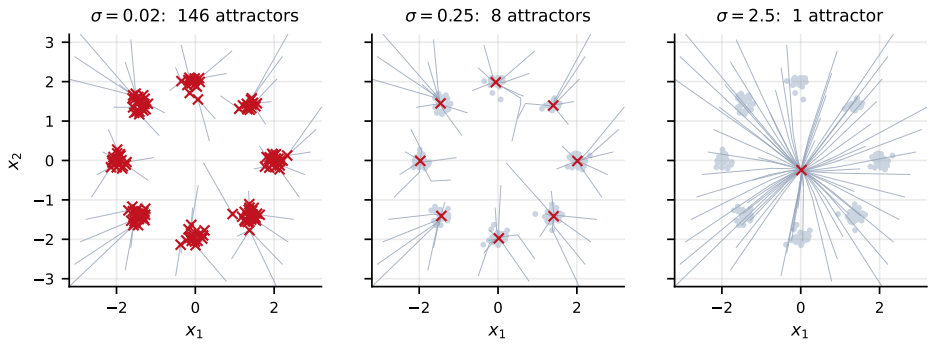

Figure 2: One model, a family of dynamical systems. Orbits of the exact map M_σ (Eq. 1) for samples from an 8-mode mixture of gaussians. *Left:* at small σ each datum carries its own attractor. *Middle:* at intermediate σ the attractors are the 8 population modes. *Right:* at large σ a single attractor remains, at the data mean.

Throughout the paper our goal will be to characterize the dynamics of M_σ as a function of the noise scale σ: see Figures 2 and 3, top row, for a 2D example of the map M at different σ_s and its dynamics. Figure 7 (Appendix A) shows the loop on Stable Diffusion.

When the denoiser is exact, i.e. it predicts the posterior mean $D_\sigma(y) = \mathbb{E}[x_0 \mid y]$, the dynamics of M_σ admit an explicit description:

Proposition 2 (formal: Proposition B.2)**.** *For the exact denoiser, one step of the dynamics is a gradient-ascent step on the smoothed density, $M_\sigma(x) = x + \sigma^2 \nabla \log p_\sigma(x)$. For a training set $p = \frac{1}{N}\sum_i \delta_{x_i}$, M is Gaussian mean shift with bandwidth σ,*

$$M_\sigma(x) = \sum_i w_i(x)\, x_i, \qquad w_i(x) = \frac{\exp\!\big(-\|x - x_i\|^2/2\sigma^2\big)}{\sum_j \exp\!\big(-\|x - x_j\|^2/2\sigma^2\big)}. \tag{1}$$

Its fixed points are the critical points of p_σ, its attractors are the modes of p_σ, and every orbit increases p_σ and converges to a fixed point.

The first identity is Tweedie's formula (Robbins, 1956; Efron, 2011), and the iteration of equation 1 is mean shift (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 2002), the classical mode-seeking clustering algorithm, run at bandwidth σ. The attractors of the denoising dynamics are therefore the modes of the data blurred at scale σ. In the limit $\sigma \to 0$ every training example is an attractor of its own, whereas once σ exceeds half the diameter of the data, a single attractor remains, close to the data mean (Proposition B.7); at intermediate scales the attractors merge progressively, nearby data into cluster modes and cluster modes into coarser ones. Figure 2 shows the three regimes, and Figure 3 traces the merging as a tree, in which the branch of an example terminates where the example ceases to be an attractor of its own.

2.1 Memorization and the critical scale

As σ increases, a training example is absorbed into a coarser attractor, and the model reproduces only its cluster. We regard an example as *memorized* to the extent that it resists this absorption: in Figure 3, non-duplicated examples lose their branch at small σ, whereas the duplicated example retains it until merging with entire groups. To measure the scale of absorption without locating the attractors, we initialize the dynamics at the example and test whether the orbit remains close to it.

Definition 3 (Critical scale)**.** Fix an escape radius $r > 0$ and an iteration budget K. The probe x is *retained at scale* σ if $\|M_\sigma^k(x) - x\| \le r\|x\|$ for every $k \le K$. The critical scale of x is the largest scale at which it is retained,

$$\sigma_c(x) \;=\; \sup\big\{\sigma > 0 :\; x \text{ is retained at scale } \sigma\big\}.$$

The critical scale is thus the largest noise level at which the model, iterated on its own output, still returns x: below σ_c the example lies in a basin of its own, and above it the orbit is drawn towards an attractor possibly shared with other data. Figure 1 illustrates the definition on a memorized and

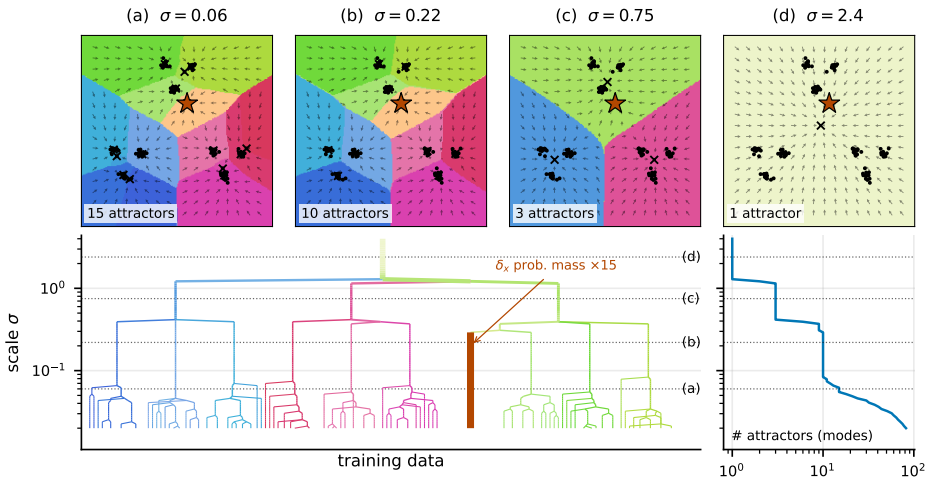

Figure 3: Scale space of attractors. Exact map M_σ on a 2-D training set: three groups of three sub-clusters (black dots), plus one datum repeated 15 times (red star). *Top:* basins of attraction at four scales, coloured by the reached attractor ($\times$); arrows are the drift $M_\sigma(x) - x$. *Bottom left:* merge tree of the attractors reached from the data (width = number of data collected). Single data lose their attractor at small σ; the duplicated datum (red) keeps it until merging with the neighbouring sub-clusters. *Bottom right:* number of attractors. A denoiser trained on the same data reproduces this scale space, basins, merge tree and counts alike (Appendix C, Fig. 8).

a control image on Stable Diffusion model: the memorized image is returned over a range of scales an order of magnitude wider than the control, and σ_c marks the upper end of this range (the full grid of scales and iterations is in Figure 14, Appendix G). Being expressed in units of noise, σ_c is comparable across models and noise schedules.

What the critical scale captures. For the exact denoiser, the critical scale σ_c responds to the probability mass carried by an example and to its distance from the rest of the data.

Theorem 4 (formal: Theorem B.9)**.** *Increasing the probability mass on an example, for instance by duplicating it, reduces the displacement $\|M_\sigma(x) - x\|$ at the example at every noise level and can therefore only increase its critical scale.*

The map is pulled away from the example only by the other data, and additional mass on the example dilutes this pull. The following theorem makes the rate explicit.

Theorem 5 (formal: Thm. B.11)**.** *Consider an example of mass $\theta \leq \frac{1}{2}$ whose only neighbour, carrying the remaining mass, lies at distance d. The example keeps an attractor of its own exactly below a critical scale, which increases with θ. For small θ this scale grows linearly in the distance and only logarithmically in the mass, as $d/\sqrt{2\log(1/\theta)}$.*

Doubling the distance to the neighbour therefore doubles σ_c, whereas doubling the copies raises it by a few percent. The critical scale thus captures both duplicated examples and outliers. At σ_c the attractor disappears in a saddle-node bifurcation, which the escape test detects (Proposition B.12).

Conditional models and the caption gap. A text-to-image model defines one family of dynamics $M_\sigma(\cdot\,; c)$ for each caption c, with $c = \varnothing$ the unconditional model. Memorization is then a property of a caption–image *pair*, and we measure the change that the caption induces in the σ_c of the image.

Definition 6 (Caption gap)**.** With $\sigma_c(x; c)$ the critical scale computed with $M_\sigma(\cdot\,; c)$, the *caption gap* of the pair (x, c) is $\Delta \log \sigma_c(x; c) = \log \sigma_c(x; c) - \log \sigma_c(x; \varnothing)$.

Subtracting the unconditional value removes the retention that the image exhibits on its own. For the exact denoiser, the caption enters the mean shift equation 1 only through the prior weights of the training images, like a duplication count, and Theorem 4 gives the following.

Corollary 7 (formal: Corollary B.14)**.** *For a single iteration, and when the caption leaves the relative weights of the other examples unchanged, the caption gap of (x_i, c) is positive if and only if the caption raises the model's mass on* that *example.*

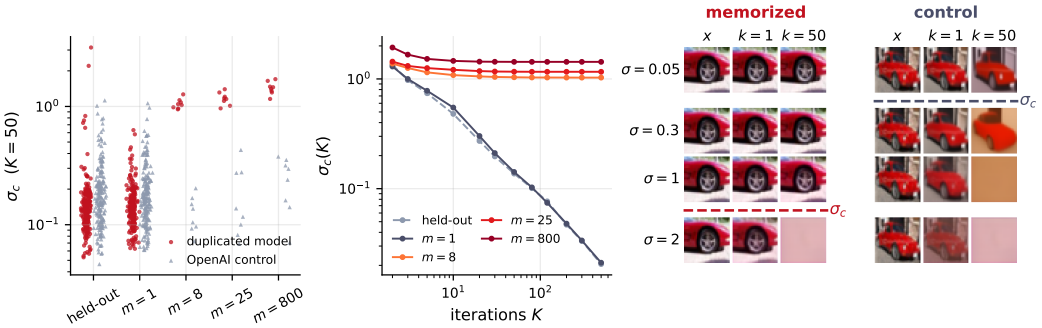

Figure 4: Duplication on full CIFAR-10. *(a)* σ_c of each image, grouped by multiplicity, in the model trained with duplicates (circles) and in the released model trained on the same images without duplicates (triangles). *(b)* σ_c as a function of the iteration budget K (median per group): constant for images with $m \geq 8$, which are fixed points, and decreasing as $1/K$ for held-out and $m = 1$ images, which are transients. *(c)* Orbits $M_\sigma^k(x)$ of a memorized training image ($m = 400$) and of a training image seen once ($m = 1$), with σ fixed along each row and k along the columns; the dashed line marks the σ_c of each image ($K = 50$, $r = 0.5$; 1.62 and 0.14). The memorized image is still reproduced after 50 iterations at $\sigma = 1$, where the other one has already vanished.

From the exact denoiser to trained networks. The results above concern the exact denoiser, of which a trained network is only an approximation, and we do not expect a trained network to reproduce their values. We expect the mechanism to carry over: an example retains a basin of its own when it carries mass or lies far from the rest of the data, and σ_c measures the width of that basin. The experiments of Section 3 verify if memorization from duplicated data and outliers transfers to real data scenarios. On the exact denoiser of a real training set the theory holds quantitatively (Appendix D), and Appendix D.2 examines how closely trained networks approach it.

3 Controlled Experiments

We first evaluate the critical scale in a controlled setting, with DDPM (Ho et al., 2020) models on the `CIFAR-10` dataset (Krizhevsky, 2009). An example can be memorized for different reasons, but these mechanisms are confounded in pretrained models. We therefore isolate them through controlled training interventions and test whether critical scale detects the memorization induced by each. In all experiments σ_c is computed by bisection in $\log \sigma$ on the escape event of Definition 3. Training and evaluation details, and additional results are given in Appendices F and E.

- *Duplication* (§3.1). An example is repeated m times in the training set, so the empirical distribution places mass m/N on it.
- *Rarity* (§3.2). An atypical example is added a few times. It has no training neighbours with which to share a basin, but it also receives almost no probability mass.
- *Overfitting* (Appendix E.2). The training set is reduced at fixed model capacity and training budget, moving the model from generalization to memorization. σ_c tracks this transition and detects memorization at training-set sizes where sampling no longer finds copies.

3.1 Memorization from Data Duplication

Setup. We study memorization induced by duplication, and ask whether σ_c detects it and how it depends on the number of copies. To isolate the effect of duplication from that of membership, we compare two models trained on the same images: one with duplicates and one without. We train the improved-diffusion CIFAR-10 model (Nichol and Dhariwal, 2021) (50M parameters, 500k steps) on the full training set, in which eight images are repeated at each multiplicity $m \in \{2, 8, 25, 50, 100, 200, 400, 800\}$. As control we use the publicly released checkpoint of the same model, trained on the same data without duplicates. We evaluate σ_c on the 64 duplicated images, on 200 training images that appear once ($m = 1$), and on 200 held-out images. Duplication does induce copying: 16 605 of 50 000 generated samples are copies of a duplicated image.

Results. Figure 4a shows that σ_c responds to the intervention rather than to the images. In the model trained with duplicates, the median σ_c is 0.137 on held-out images and 0.142 at $m = 1$, against

1.04–1.43 for $m \geq 8$. In the control model the same images have σ_c between 0.12 and 0.27 at every multiplicity, and the rank correlation with m drops from $\rho = 0.70$ to 0.19. Used as a detector, σ_c separates $m \geq 8$ from $m = 1$ with Area Under the Curve (AUC) 1.000. The duplicated images are fixed points: at $m = 800$ the median σ_c is the same for $K = 50$ and $K = 500$ iterations, while for held-out and $m = 1$ images it decreases by more than two orders of magnitude over the same range (Figure 4b). Above $m = 8$, σ_c increases by only a factor 1.38 while m increases a hundredfold, the saturation expected of a width in which mass enters logarithmically (Appendix E.1).

Takeaway. The critical scale captures memorization from data duplication, which turns transient states in the image orbits into stable fixed points. Additional copies increase σ_c only slightly.

3.2 Memorization of rare samples

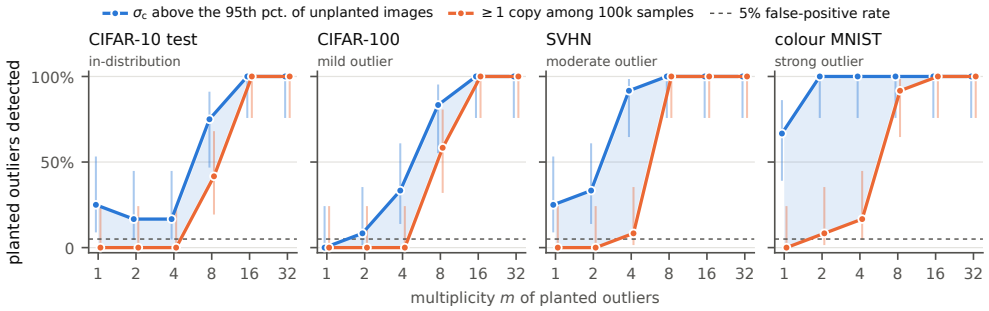

Figure 5: Rarity: detection thresholds of σ_c and of sampling. Fraction of added images detected as a function of their multiplicity m, for sources of increasing atypicality. *Orange*: the image appears at least once among 10^5 samples; the threshold is $m \approx 8$ for every source. *Blue*: σ_c exceeds the 95th percentile of the images of the same source not added to the model; the threshold decreases to $m = 1$ as the images become more atypical. The shaded region is memorization detected by σ_c but not by sampling.

Setup. We study memorization of outliers, which Theorem 5 predicts to be retained with few copies because of their distance to the other samples. We train two models, A and B, on the same subset of $N = 10^4$ `CIFAR-10` images, to which we add images from 4 datasets of increasing atypicality: `CIFAR-10` test images, `CIFAR-100` classes with no `CIFAR-10` counterpart, `SVHN` digits, and colourised `MNIST` digits. The added images statistics are matched to the `CIFAR-10`. From each source, 36 images are added to A, 36 other images to B, and 100 images to neither, with six images at each multiplicity $m \in \{1, 2, 4, 8, 16, 32\}$. Each image is thus memorized in one model and not in the other, and the images added to neither model serve as a reference distribution. We draw 10^5 samples from each model and count near-copies of the added images.

Results. Sampling and σ_c have different detection thresholds (Figure 5). Sampling detects added images only from $m \approx 8$, for every source; below this multiplicity no image is generated. We flag an image with σ_c when its value exceeds the 95th percentile non-added images from the same source, which gives a false-positive rate of 5% without using labels. The resulting threshold decreases with atypicality: $m = 8$ for `CIFAR-10` and `CIFAR-100`, $m = 4$ for `SVHN` and $m = 1$ for colour `MNIST`. The paired design confirms that the effect is due to training on the image and not to the image itself. Among the added images that were never sampled in the 10^5 drawns, σ_c flags 88.2% (`MNIST`), 48.6% (`SVHN`), 25.6% (`CIFAR-10`), and 22.0% (`CIFAR-100`). Additional analysis on how σ_c depends on distance and on multiplicity in these models is in Appendix E.3.

Takeaway. The critical scale can detect memorized samples that are unlikely to be generated: σ_c depends also on distance, generation only on mass. An isolated example (outlier) can be memorized, since it has room for its own basin, yet not be sampled, since it carries little probability mass.

4 Detecting Memorization in Stable Diffusion

We evaluate Stable Diffusion v1.4 (Rombach et al., 2022) on memorized `LAION` training examples identified by Webster (2023): *matching verbatim* (MV) examples, whose generations reproduce the paired training image, and *template verbatim* (TV) examples, whose generations preserve only part of its content or structure. After removing duplicates and mismatched image–caption pairs

Table 1: Memorization-detection on Stable Diffusion v1.4. AUC for distinguishing MV and TV examples from controls, caption-swapped pairs, and MV from TV. Our global scores use $r = 0.25$, and local $|\Delta \log \sigma_c|$ uses $r = 0.1$. For both variants of Wen et al. (2024), we report its best-performing configuration (Appendix F).

Score	MV vs. control	MV vs. swap	TV vs. control	TV vs. swap	MV vs. TV
$\|\epsilon_c - \epsilon_\varnothing\|$ Wen et al. (2024)	**0.999**	0.606	**0.999**	0.184	0.902
$\|\epsilon_c(x_0) - \epsilon_\varnothing(x_0)\|$	**0.999**	0.887	**0.996**	0.696	0.868
σ_c, $K = 16$	0.917	0.904	0.695	0.803	0.909
$\Delta \log \sigma_c$, $K = 16$	0.8569	**0.908**	0.824	**0.939**	0.820
$\|\Delta \log \sigma_c\|$, $K = 16$	0.984	0.860	0.879	0.354	0.905
local $\|\Delta \log \sigma_c\|$, $K = 16$	**0.997**	0.865	**0.993**	0.458	**0.935**

(Appendix F), we retain 70 MV and 24 TV examples. As controls we use 94 training images from `LAION-MI`(Dubiński et al., 2024), drawn from LAION Aesthetics v2 5+(Schuhmann et al., 2022), so that all images are training members and differ only in their degree of memorization. Each memorized image is evaluated with its own caption, a *swap* caption (captions randomly permuted across memorized images) and the empty caption; each control with its own and the empty caption.

Scores. From these runs we compute the critical scale σ_c under the own caption and the caption gap $\Delta \log \sigma_c$ of Definition 6, which measures how much the caption increases the retention of that image (Figures 21–25 in Appendix G.5). Both are global and do not say *where* in the image memorization occurs. We therefore also apply the retention test separately at each latent location $u \in U$, which gives a local gap $\Delta \log \sigma_c^u$ and the local score $\frac{1}{|U|} \sum_{u \in U} |\Delta \log \sigma_c^u|$ (see Appendix F). To separate the benefit of spatial resolution from that of discarding the sign, we also report the absolute global gap $|\Delta \log \sigma_c|$. We first evaluate critical-scale scores for detecting memorized examples (§4.1), then examine whether local gaps identify memorized regions (§4.2), and finally use mismatched captions to identify images memorized in the unconditional branch (§4.3).

4.1 Critical scale detects memorization

Setup. We report AUC for separating: (i) MV and TV examples from controls, (ii) caption-swapped pairs, and (iii) MV from TV. We compare our scores with the prompt-level memorization score $\|\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \varnothing)\|$ of Wen et al. (2024). Since this measure depends only on the caption, we also consider an image-dependent variant: x_t is obtained by adding noise to the candidate image x_0. This provides a stronger baseline for testing if a particular image-caption pair is memorized.

Results. Table 1 reports the detection results (see Appendix G.1 for an ablation on K and r). All scores separate memorized examples from control, with local $|\Delta \log \sigma_c|$ performing best among our variants. It achieves AUCs of 0.997 for MV and 0.993 for TV, nearly matching the prompt-level baseline (0.999). Its largest gain occurs for TV, where it raises the AUC from 0.879 to 0.993, consistent with partial memorization being localized. It also best separates MV from TV (0.935). For caption swaps, the signed gap performs best because a mismatched caption can reduce conditional retention below unconditional retention, reversing the gap's sign (Figure 17, Appendix G.2). Taking the absolute value discards this directional information. This advantage is most pronounced for TV, where $\Delta \log \sigma_c$ achieves 0.939 AUC versus 0.696 for the image-dependent baseline.

Takeaway. Critical scale is an effective global measure of memorization at scale, and computing it per spatial location improves detection. Comparing unconditional retention with retention under a mismatched caption further reveals whether an image was memorized with a particular caption.

4.2 Local critical scale localizes memorized regions

Setup. We qualitatively assess if the local gap $|\Delta \log \sigma_c^u|$ identifies memorized regions in TV examples. As a reference, we generate five samples from each training caption: consistently reproduced regions indicate memorized content, whereas variable regions indicate non-memorized content.

Results. Figure 6 (left) shows that large values of $|\Delta \log \sigma_c^u|$ concentrate in regions reproduced consistently across generations. By contrast, regions that vary across samples generally have gaps close to zero. Thus, the spatial variation of the critical-scale gap aligns with the apparent boundary

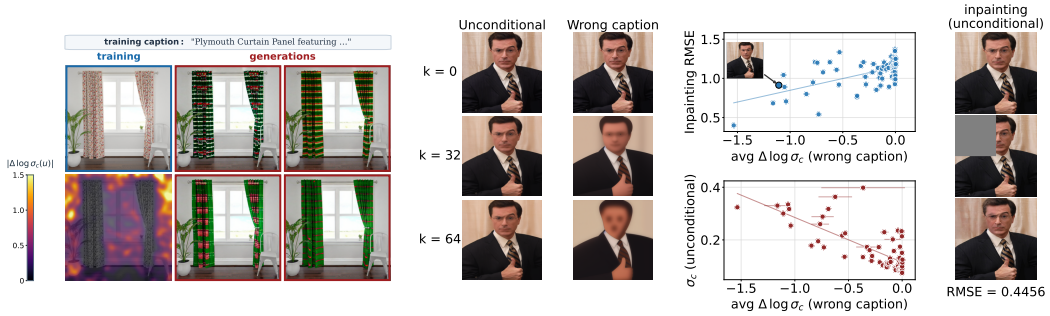

Figure 6: Critical-scale gaps reveal spatial and image-driven memorization. *Left:* A TV example: the training image, generations from its training caption, and the local gap $|\Delta \log \sigma_c^u|$. The gap is large where all generations copy the training image (the room) and near zero where they vary (the curtain pattern). *Right:* A wrong caption exposes images already retained by the unconditional model. Center: wrong-caption gap $\Delta \log \sigma_c$ against unconditional inpainting RMSE (top) and unconditional σ_c (bottom); bars include std deviation over 5 captions. Side: a negative-gap image is retained by unconditional dynamics but lost when conditioning on a wrong caption; unconditional inpainting restores the masked region.

between memorized and non-memorized content. The trajectories in Figures 22 and 23 provide complementary evidence: in partially memorized images, non-memorized content disappears at smaller noise scales, whereas memorized regions persist longer. Additional qualitative examples are provided in Figure 18 in Appendix G.3.

Takeaway. The local critical-scale gap provides an *interpretable* spatial measure of where memorized content appears within an image.

4.3 Critical-Scale Gap Identifies Image-Driven Memorization

Setup. We aim to find images that the unconditional branch has memorized by itself, and not only through their captions. A strongly negative $\Delta \log \sigma_c$ indicates stronger retention under unconditional than conditional dynamics. However, since unconditional memorization is rare, searching the full training set for such examples is computationally prohibitive. We therefore use the MV examples under five unrelated captions drawn from LAION-MI, and compute $\Delta \log \sigma_c$ for each image-caption pair. This allows us to rank images by evidence of unconditional retention. Since finding an image through unconditional sampling is not feasible, we evaluate this ranking through inpainting. We mask each image quadrant in turn, reconstruct it from the remaining three, and report the mean latent RMSE to the ground-truth region. For each image, we generate 10 samples per quadrant. Low reconstruction error provides evidence that the image is retained independently of its caption.

Results. Figure 6 (right) shows an example with a strongly negative gap. Its unconditional dynamics retain the image more strongly than the mismatched-caption dynamics, and unconditional inpainting accurately reconstructs the missing content. This provides qualitative evidence that the image is memorized by the unconditional branch. This reconstruction is consistent with the memorized image acting as an attractor and the masked image lying within its basin of attraction (see Figure 7 in Appendix A). Across images, strongly negative gaps are associated with lower inpainting RMSE (top center) and larger unconditional σ_c (bottom center), supporting their use as evidence of unconditional retention. Gaps near zero are less informative because they can arise when either both dynamics retain the image or neither does. Figure 19 in Appendix G.4 illustrates both near-zero regimes, while Figure 20 provides additional examples with strongly negative gaps.

Takeaway. A strongly negative critical-scale gap provides evidence that an image is retained independently of its caption, identifying cases in which the image itself contributes to memorization and can be reconstructed by the unconditional branch.

5 Related Work

Dynamical and multiscale views of diffusion models. Mean shift turns gradients of kernel density estimates into mode-seeking dynamics (Fukunaga and Hostetler, 1975; Cheng, 1995; Comani-

ciu and Meer, 2002); with Gaussian kernels, it equals posterior-mean denoising of the empirical distribution, whose smoothed-density critical points are fixed points and modes are attractors. The link between denoising and the score, established for denoising autoencoders (Alain and Bengio, 2014), underlies score-based diffusion via reverse-time SDEs/ODEs (Song et al., 2021). These dynamics reveal how structure emerges across noise scales: global high-variance features precede low-variance details (Wang and Vastola, 2023), a speciation transition marks the emergence of broad structure (Biroli et al., 2024), and high- and low-level features evolve differently, reflecting the data hierarchy (Sclocchi et al., 2025). These works follow generation trajectories: instead, we fix the noise scale and iterate the denoiser, obtaining a family of dynamical systems whose attractors and persistence across scales describe the learned structure.

Memorization and generalization in diffusion models. The empirical optimum of denoising score matching reproduces the training data: Gu et al. (2025) study when learned models approach it, and Baptista et al. (2025) how explicit and implicit regularization in the reverse dynamics prevent it. Kadkhodaie et al. (2024) attribute generalization to geometry-adaptive inductive biases, Shah et al. (2025) tie memorization mainly to low-noise denoising, and Bonnaire et al. (2025) show that good generation precedes memorization during training. Closest to us, Fumero et al. (2026a) iterate an autoencoder's map, whose latent fixed points and attractors reflect memorization and generalization. A diffusion model instead induce a family of dynamical systems, so attractor persistence can be tracked along a scale axis. Pham et al. (2025) view diffusion models as associative memories, and Jain et al. (2025) argue that classifier-free guidance steers reverse trajectories into the basins of memorized images. Both study the reverse generation trajectory, while our fixed-scale dynamics measure how strongly a given image-caption pair is retained across noise scales.

Detecting memorization in diffusion models. Early work detected replication by matching generations to training data (Somepalli et al., 2023) or extracting training examples (Carlini et al., 2023; Webster, 2023). Later methods avoid exhaustive generation. Wen et al. (2024) detect memorized prompts from the conditional–unconditional noise-prediction gap. Jeon et al. (2025) use probability-landscape sharpness and Ross et al. (2025) the local geometry of model generations to detect memorization. Jiang et al. (2025) address image-level detection without the paired prompt and Chen et al. (2025) localize memorized regions. Our score instead depends jointly on a candidate image–caption pair, testing whether that pair is memorized rather than whether the caption alone elicits memorization. Comparing conditional and empty-caption critical scales further detects memorization in the unconditional branch. Membership inference (Carlini et al., 2023; Hu and Pang, 2023; Zhai et al., 2024; Duan et al., 2023; Matsumoto et al., 2023) instead asks whether an example was used in training, not whether it is retained strongly enough to be reproduced.

6 Conclusions

In this paper we proposed to interpret a diffusion model as a *family of dynamical systems* indexed by the noise level, whose attractors form a scale space of the data. Memorization then becomes the persistence of an example as an attractor, measured by the critical scale σ_c. On Stable Diffusion, σ_c detects fully and partially memorized image–caption pairs and offers two forms of interpretation: comparing conditional and unconditional critical scales isolates the caption-image interaction in memorization, while local variants of σ_c can identify which regions within an image are memorized. Further analysis is needed to quantify the precision and generality of this localization, which in principle can be extended also to the caption parts. Our theory holds for the exact denoiser. While the mechanism transfers to trained networks (see Section 3), closing this gap formally–for instance via analytic models of the inductive biases of trained diffusion models Kamb and Ganguli (2024)–is a natural direction for future work. Beyond memorization, the orbits and basins of the denoising maps may reveal broader structure in what a model has learned Sclocchi et al. (2025), and locating where individual examples are retained in the weights Hintersdorf et al. (2024) suggests a route towards targeted unlearning.

AI USE STATEMENT

We used generative AI tools as coding assistants when implementing and modifying the codebase, and for language editing. They were not used to produce research results on their own or to replace the authors' scientific judgment. The authors designed the methodology and experiments, and reviewed and validated all AI-assisted code and text.

ETHICS STATEMENT

This work is primarily theoretical and methodological, focusing on the dynamics of denoising maps and on memorization in diffusion models. All datasets and models are publicly available (CIFAR-10, MNIST, SVHN, LAION-5B (Schuhmann et al., 2022) and Stable Diffusion (Rombach et al., 2022), and the memorized prompts we study were identified in prior work (Webster, 2023). No human subjects are involved and no new data is collected. Our detectors probe which training images a model has memorized. In principle, such insights could be misused to recover training data from pretrained models. However, our experiments score images that are already public, we release no new private or copyrighted content, and the techniques should be applied responsibly, for instance to audit and reduce memorization. We hope our findings contribute to a better understanding of how diffusion models generalize and memorize.

REPRODUCIBILITY STATEMENT

All formal statements and proofs are reported in Appendix B. We implement all experiments in PyTorch (Paszke et al., 2019) and the HuggingFace `diffusers` library (von Platen et al., 2022). We use only publicly released checkpoints, and the CIFAR-10 models we train follow the public recipe of Nichol and Dhariwal (2021). Models, datasets, training recipes, noise schedules, estimator hyperparameters and the filtering of the memorization set are listed in Appendix F. We will open-source our codebase for all experiments upon acceptance.

REFERENCES

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning Research*, 15(110):3743–3773, 2014.

Ricardo Baptista, Agnimitra Dasgupta, Nikola B Kovachki, Assad Oberai, and Andrew M Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

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

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *Advances in Neural Information Processing Systems*, 2025.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Chen Chen, Daochang Liu, Mubarak Shah, and Chang Xu. Exploring local memorization in diffusion models via bright ending attention. In *International Conference on Learning Representations*, volume 2025, pages 29827–29841, 2025.

Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995. doi: 10.1109/34.400568.

Dorin Comaniciu and Peter Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. doi: 10.1109/34.1000236.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8717–8730. PMLR, 2023.

Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemyslaw Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4860–4869, 2024.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975. doi: 10.1109/TIT.1975.1055330.

Marco Fumero, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Navigating the latent space dynamics of neural models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 24765–24793, 2026a.

Marco Fumero, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Navigating the latent space dynamics of neural models. In *International Conference on Learning Representations*, 2026b.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding nemo: Localizing neurons responsible for memorization in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hailong Hu and Jun Pang. Loss and likelihood based membership inference of diffusion models. In *Information Security – 26th International Conference, ISC 2023*, volume 14411 of *Lecture Notes in Computer Science*, pages 121–141. Springer, 2023.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12871–12879. IEEE, 2025.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding and mitigating memorization in generative models via sharpness of probability landscapes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 27091–27112. PMLR, 2025.

Yue Jiang, Haokun Lin, Yang Bai, Bo Peng, Zhili Liu, Yueming Lyu, Yong Yang, Xing Zheng, and Jing Dong. Image-level memorization detection via inversion-based inference perturbation. In *International Conference on Learning Representations*, 2025.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations (ICLR)*, 2024.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops*. IEEE, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PmLR, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: Emergence of diffusion models from associative memory. *arXiv preprint arXiv:2505.21777*, 2025.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 157–163. University of California Press, 1956.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Brendan Leigh Ross, Hamidreza Kamkari, Tongzi Wu, Rasa Hosseinzadeh, Zhaoyan Liu, George Stein, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. A geometric framework for understanding memorization in generative models. In *International Conference on Learning Representations*, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025.

Kulin Shah, Alkis Kalavasis, Adam Klivans, and Giannis Daras. Does generation require memorization? creative diffusion models using ambient diffusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 54143–54166. PMLR, 2025.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

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 (ICLR)*, 2021.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Binxu Wang and John J Vastola. Diffusion models generate images like painters: an analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

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

Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, and Yang Liu. Membership inference on text-to-image diffusion models via conditional likelihood discrepancy. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

A Denoising dynamics in Stable Diffusion

Figure 7 shows Definition 1 on Stable Diffusion v1.4. The image is encoded once into the latent space, where the map is iterated; the decoder is used only to display the iterates.

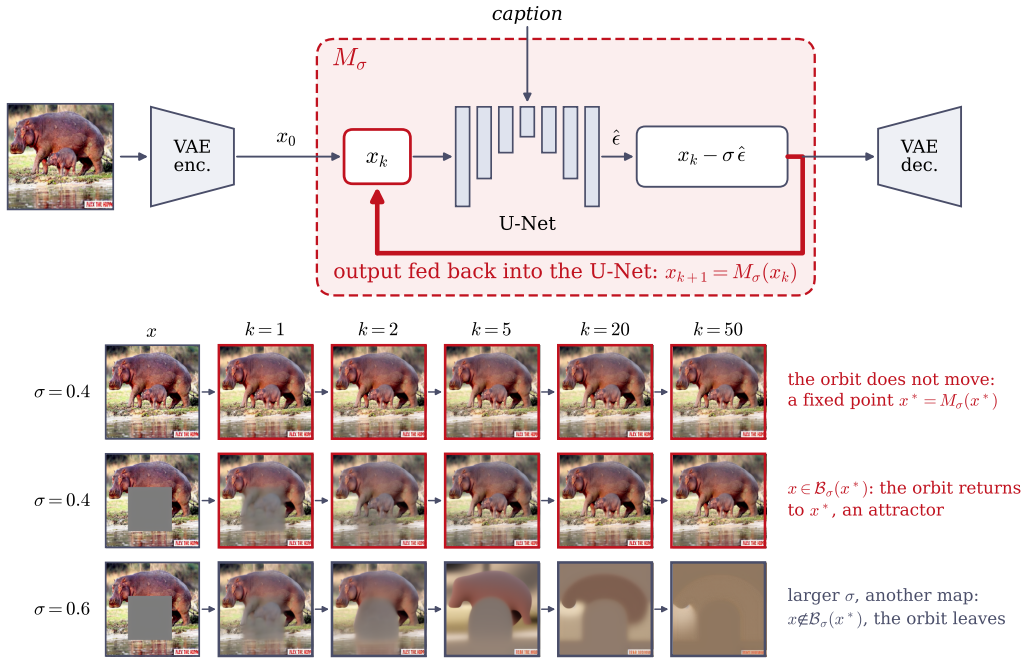

Figure 7: Definition 1 on Stable Diffusion. *Top:* the map M_σ. The image is encoded once into the latent x_0; the U-Net is evaluated with the caption at noise level σ, and its output $x_k - \sigma\hat{\epsilon}$ is fed back as the next input. *Bottom:* each row is an orbit $(M_\sigma^k(x))_{k\geq 0}$, decoded for display, of a memorized image under its own caption ($w = 1$). *First row:* started at the image, the orbit does not move; the image is a fixed point x^*. *Second row:* started with part of the image replaced by a grey square, the orbit restores it and converges to the same x^*. The occluded image lies in the basin $\mathcal{B}_\sigma(x^*)$, so x^* is an attractor. *Third row:* the same start at a larger σ, which defines a different map in the family $\{M_\sigma\}_{\sigma>0}$. There the start is no longer in the basin, and the orbit leaves.

B Proofs

Setting and notation. Throughout, p is a probability measure supported in a compact convex set $S \subset \mathbb{R}^d$ of diameter $R < \infty$. This covers image data in $[-1,1]^d$ and, taking S to be its convex hull, any finite training set. Write $\varphi_\sigma(u) = (2\pi\sigma^2)^{-d/2} e^{-\|u\|^2/2\sigma^2}$. The smoothed density $p_\sigma(x) = \int \varphi_\sigma(x - x_0)\, p(dx_0)$ is C^∞ and strictly positive, and derivatives may be taken under the integral because S is bounded. The posterior over clean signals given x is $\pi_x(dx_0) = \varphi_\sigma(x - x_0)\, p(dx_0)/p_\sigma(x)$. $\mathbb{E}[\cdot \mid x]$ and $\mathrm{Cov}[\cdot \mid x]$ denote its mean and covariance, so $D_\sigma(x) = \mathbb{E}[x_0 \mid x]$. A fixed point x^* of M_σ is an *attractor* if every eigenvalue of $\nabla M_\sigma(x^*)$ has modulus less than 1. A *mode* is a local maximum of p_σ, and it is *nondegenerate* if $\nabla^2 \log p_\sigma$ is negative definite there. Two elementary facts are used throughout.

Lemma B.1. *For every $x \in \mathbb{R}^d$:* (F1) $D_\sigma(x) \in S$; (F2) $0 \preceq \mathrm{Cov}[x_0 \mid x] \preceq \frac{R^2}{4} I$.

Proof. (F1): $D_\sigma(x)$ is the mean of a probability measure on the convex set S. (F2): for a unit vector v, $v^\top x_0$ lies in an interval of length at most R, and a random variable confined to an interval of length ℓ has variance at most $\ell^2/4$. □

B.1 THE EXACT MAP (SECTION 2)

Intuition. The exact denoiser averages the training data, weighting each clean signal by how plausible it is as the source of x. Moving to that average moves x uphill on the smoothed density: the step is the score (Tweedie). How sharply the output reacts to the input is the spread of the plausible sources, i.e. the posterior covariance. So a point is held in place when the posterior at it is narrower than the noise, which is exactly a mode of p_σ.

Proposition B.2 (Exact denoisers are mean shift; formal version of Proposition 2)**.** *Let $D_\sigma(x) = \mathbb{E}[x_0 \mid x]$ be the exact denoiser. Then M_σ is Gaussian mean shift with bandwidth σ on the smoothed density p_σ:*

(i) $M_\sigma(x) = x + \sigma^2 \nabla \log p_\sigma(x)$, and for $p = \frac{1}{N}\sum_i \delta_{x_i}$ it is given by equation 1.
(ii) The fixed points of M_σ are the critical points of p_σ. A fixed point x^* is an attractor iff $\mathrm{Cov}[x_0 \mid x^*] \prec \sigma^2 I$, i.e. iff it is a nondegenerate mode of p_σ.
(iii) Every step increases p_σ. The map has no cycles, and every orbit converges to a fixed point when the fixed points are isolated.

Proposition B.2 combines four statements, which we prove in turn: Tweedie's identities (Proposition B.3), the fixed points and their stability (Proposition B.4), the ascent property (Lemma B.5) and its consequence for orbits (Corollary B.6).

Proposition B.3 (Tweedie)**.** *$D_\sigma(x) = x + \sigma^2 \nabla \log p_\sigma(x)$ and $\nabla D_\sigma(x) = \sigma^{-2}\mathrm{Cov}[x_0 \mid x]$.*

Proof of Proposition B.3. This is Tweedie's formula (Robbins, 1956; Efron, 2011). Since $\nabla_x \varphi_\sigma(x - x_0) = \sigma^{-2}(x_0 - x)\varphi_\sigma(x - x_0)$, integrating against p gives $\nabla p_\sigma(x) = \sigma^{-2} p_\sigma(x)(D_\sigma(x) - x)$, which is the first identity. For the second, note that $\nabla_x \log[\varphi_\sigma(x - x_0)/p_\sigma(x)] = \sigma^{-2}(x_0 - D_\sigma(x))$. Differentiating $D_\sigma(x) = \int x_0 \, \pi_x(dx_0)$ under the integral therefore gives $\nabla D_\sigma(x) = \sigma^{-2}\mathbb{E}[x_0(x_0 - D_\sigma(x))^\top \mid x] = \sigma^{-2}\mathrm{Cov}[x_0 \mid x]$. □

Combining the two identities of Proposition B.3,

$$\nabla^2 \log p_\sigma(x) = \sigma^{-2}\big(\nabla D_\sigma(x) - I\big) = \sigma^{-4}\mathrm{Cov}[x_0 \mid x] - \sigma^{-2} I. \tag{2}$$

For a finite $p = \sum_i \mu_i \delta_{x_i}$, where $\mu_i \propto m_i$ and m_i is the multiplicity of x_i, the posterior is discrete and its mean is

$$M_\sigma(x) = \sum_i w_i(x)\, x_i, \qquad w_i(x) = \frac{m_i \, e^{-\|x - x_i\|^2/2\sigma^2}}{\sum_j m_j \, e^{-\|x - x_j\|^2/2\sigma^2}}. \tag{3}$$

This is equation 1 when all $m_i = 1$. The weighted form is the one used for $\sigma_{\mathrm{c}}^{\mathrm{emp}}$.

Proposition B.4 (Fixed points and stability)**.** *x^* is a fixed point of M_σ iff $\nabla p_\sigma(x^*) = 0$. The Jacobian ∇M_σ is symmetric positive semidefinite everywhere, so the linearised map neither rotates nor oscillates. A fixed point is an attractor iff $\mathrm{Cov}[x_0 \mid x^*] \prec \sigma^2 I$, equivalently iff x^* is a nondegenerate mode of p_σ. Saddles of p_σ carry the basin boundaries.*

Proof of Proposition B.4. Fixed points. By Proposition B.3, $M_\sigma(x) - x = \sigma^2 \nabla \log p_\sigma(x)$. Since $p_\sigma > 0$, $M_\sigma(x^*) = x^*$ iff $\nabla p_\sigma(x^*) = 0$.

Jacobian. $\nabla M_\sigma = \sigma^{-2}\mathrm{Cov}[x_0 \mid x]$ is symmetric positive semidefinite. Its eigenvalues are therefore real, so the linearised map has no rotation, and non-negative, so it has no sign-reversing (oscillating) direction.

Stability. Because the eigenvalues are non-negative, they all lie in $[0, 1)$ iff $\mathrm{Cov}[x_0 \mid x^*] \prec \sigma^2 I$. By equation 2 this holds iff $\nabla^2 \log p_\sigma(x^*) \prec 0$. At a critical point, that is a nondegenerate local maximum.

Saddles. If $\nabla^2 \log p_\sigma(x^*)$ has a positive eigenvalue, equation 2 gives $\nabla M_\sigma(x^*)$ an eigenvalue greater than 1, so x^* repels along that direction. A saddle attracts along some directions and repels

along others. By Corollary B.6 below, a point in no attractor's basin converges to such a non-maximal critical point. Generically this is a saddle, and the sets converging to saddles separate the basins. □

The linearisation is local. The global behaviour comes from the classical ascent property of Gaussian mean shift (Comaniciu and Meer, 2002): each step climbs p_σ by at least its own squared length.

Lemma B.5 (Ascent). *For every x, $\log p_\sigma(M_\sigma(x)) - \log p_\sigma(x) \geq \|M_\sigma(x) - x\|^2/2\sigma^2$.*

Proof. For any y, $p_\sigma(y)/p_\sigma(x) = \mathbb{E}\big[\exp\{(\|x-x_0\|^2 - \|y-x_0\|^2)/2\sigma^2\} \mid x\big]$. Jensen's inequality bounds its logarithm below by the posterior mean of the exponent. That mean equals $\big(\|x\|^2 - \|y\|^2 - 2(x-y)^\top D_\sigma(x)\big)/2\sigma^2$, which at $y = D_\sigma(x)$ is $\|x - D_\sigma(x)\|^2/2\sigma^2$. □

Corollary B.6 (Orbits of the exact map). *Let $x_{k+1} = M_\sigma(x_k)$. Then (i) M_σ has no periodic orbit other than fixed points; (ii) every limit point of (x_k) is a fixed point, and if the fixed points are isolated, (x_k) converges; (iii) every attractor is a local maximum of p_σ.*

Proof. (i) By Lemma B.5, p_σ strictly increases along any step that moves, so an orbit cannot return to its starting point. (ii) Summing Lemma B.5 over k and using $p_\sigma \leq (2\pi\sigma^2)^{-d/2}$ gives $\sum_k \|x_{k+1} - x_k\|^2 < \infty$, so the steps tend to 0. By (F1) the orbit is bounded. If $x_{k_j} \to z$, then also $x_{k_j+1} \to z$, and continuity of M_σ gives $M_\sigma(z) = z$. The limit set of a bounded sequence whose steps tend to 0 is connected, and a connected set of isolated points is a single point. (iii) If every orbit started in a neighbourhood U of x^* converges to x^*, then $p_\sigma(x) \leq p_\sigma(x^*)$ for all $x \in U$, because p_σ increases along these orbits. □

Proof of Proposition B.2. (i) is the first identity of Proposition B.3, and equation 1 is equation 3 with all $m_i = 1$. (ii) is Proposition B.4. (iii) is Lemma B.5 together with Corollary B.6(i)–(ii). □

Both ends of the scale axis. At large σ every posterior is nearly the whole data distribution, so the map barely depends on x and collapses everything to one point. At small σ the posterior near a datum is almost entirely that datum, so the map is nearly constant on a small ball around it, which then holds one attractor. Both statements are contraction arguments.

Proposition B.7. *(a) If $\sigma > R/2$, M_σ has exactly one fixed point, and every orbit converges to it. As $\sigma \to \infty$ this fixed point tends to $\mathbb{E}_p[x_0]$. (b) Let $p = \sum_i \mu_i \delta_{x_i}$ with pairwise distances at least δ, let $B_i = \{x : \|x - x_i\| \leq \delta/4\}$, and let $\epsilon_i = \frac{1-\mu_i}{\mu_i} e^{-\delta^2/4\sigma^2}$. If $\epsilon_i R \leq \delta/4$ and $\epsilon_i R^2 < \sigma^2$, which holds for all small enough σ, then B_i contains exactly one fixed point. It is an attractor, and it lies within $\epsilon_i R$ of x_i.*

Proof. (a) By (F2), $\|\nabla M_\sigma(x)\| \leq R^2/4\sigma^2 < 1$ everywhere, and by (F1) M_σ maps S into itself. By Banach's fixed-point theorem, M_σ has a unique fixed point x^*_σ in S. Every orbit enters S after one step, so every orbit converges to it, and by (F1) there are no fixed points outside S. As $\sigma \to \infty$, $\pi_x \to p$ uniformly in $x \in S$, so $x^*_\sigma = \mathbb{E}[x_0 \mid x^*_\sigma] \to \mathbb{E}_p[x_0]$.

(b) Let $x \in B_i$ and $j \neq i$. Then $\|x - x_j\| \geq 3\delta/4$ and $\|x - x_i\| \leq \delta/4$, so $\|x - x_j\|^2 - \|x - x_i\|^2 \geq \delta^2/2$ and

$$\frac{w_j(x)}{w_i(x)} \leq \frac{\mu_j}{\mu_i}\, e^{-\delta^2/4\sigma^2}, \qquad \text{hence} \qquad \sum_{j \neq i} w_j(x) \leq \frac{1-\mu_i}{\mu_i}\, e^{-\delta^2/4\sigma^2} = \epsilon_i. \tag{4}$$

Two bounds follow, using $\mathrm{Cov}[x_0 \mid x] \preceq \mathbb{E}[(x_0 - x_i)(x_0 - x_i)^\top \mid x]$ for the second:

$$\|M_\sigma(x) - x_i\| = \Big\| \sum_{j \neq i} w_j(x)(x_j - x_i) \Big\| \leq \epsilon_i R \leq \delta/4, \tag{5}$$

$$\|\nabla M_\sigma(x)\| \leq \sigma^{-2} \Big\| \sum_{j \neq i} w_j(x)(x_j - x_i)(x_j - x_i)^\top \Big\| \leq \frac{\epsilon_i R^2}{\sigma^2} < 1. \tag{6}$$

The first gives $M_\sigma(B_i) \subset B_i$, so M_σ is a contraction of the convex set B_i into itself, and Banach's theorem gives a unique fixed point there. Its Jacobian has norm below 1, so it is an attractor, and by the first bound it lies within $\epsilon_i R$ of x_i. □

B.2 The critical scale and mass monotonicity (Sections 2.1–2.1)

Proposition B.8 (Critical scale at $K = 1$)**.** *For the exact denoiser, the critical scale with a single iteration is*

$$\sigma_{\mathrm{c}}^{(1)}(x) = \sup \big\{ \sigma > 0 : \ \sigma^2 \, \|\nabla \log p_\sigma(x)\| \leq r\|x\| \big\}. \tag{7}$$

Proof of Proposition B.8. At $K = 1$ the probe is retained at σ iff $\|M_\sigma(x) - x\| \leq r\|x\|$, and by Proposition B.3 $\|M_\sigma(x) - x\| = \sigma^2 \|\nabla \log p_\sigma(x)\|$. Take the supremum over the retained scales. □

Two direct consequences of Definition 3 are worth recording. σ_{c} is nonincreasing in K, since retention for $K + 1$ steps implies retention for K. The retention set need not be an interval, which is why we verify it empirically (Appendix C).

Theorem B.9 (Mass monotonicity; formal version of Theorem 4)**.** *Let $p_\theta = (1 - \theta)q + \theta\delta_{x^*}$ put mass θ on x^*, with q any distribution of bounded support. Then $\|D_\sigma(x^*) - x^*\|$ is strictly decreasing in θ, and $\sigma_{\mathrm{c}}^{(1)}(x^*)$ is nondecreasing in θ.*

Duplicating a datum m times among N others corresponds to $\theta = m/(N + m)$.

Intuition for Theorem B.9. At the probe, the posterior splits its vote between the probe's own atom and the rest of the data. Only the rest pulls the probe away, and it pulls with the same force whatever the probe's mass. More mass on the probe only shrinks the share of the vote the rest receives. The proof makes this exact.

Proof of Theorem B.9. Let D^q and π_x^q be the exact denoiser and the posterior of q, and set $c = \varphi_\sigma(0)$ and $B = \int \varphi_\sigma(x^* - x_0)\, q(dx_0) > 0$. At x^* the posterior of p_θ is a mixture of the atom and the posterior of q:

$$\pi_{x^*} = \lambda_\theta \, \delta_{x^*} + (1 - \lambda_\theta) \, \pi_{x^*}^q, \qquad \lambda_\theta = \frac{\theta c}{\theta c + (1 - \theta) B}. \tag{8}$$

Taking means,

$$D_\sigma^\theta(x^*) - x^* = (1 - \lambda_\theta)\big(D^q(x^*) - x^*\big). \tag{9}$$

The vector $D^q(x^*) - x^*$ does not depend on θ, and $1 - \lambda_\theta$ is strictly decreasing in θ. Hence the displacement

$$f_\theta(\sigma) := \|D_\sigma^\theta(x^*) - x^*\| = (1 - \lambda_\theta) \, \|D^q(x^*) - x^*\| \tag{10}$$

is strictly decreasing in θ at every σ with $D^q(x^*) \neq x^*$, and zero at every other σ. For $\theta < \theta'$ this gives $f_{\theta'} \leq f_\theta$, so the $K = 1$ retention sets are nested,

$$\{\sigma : \ f_\theta(\sigma) \leq r\|x^*\|\} \ \subseteq \ \{\sigma : \ f_{\theta'}(\sigma) \leq r\|x^*\|\}, \tag{11}$$

and their suprema, $\sigma_{\mathrm{c}}^{(1)}(x^*)$ by Proposition B.8, are nondecreasing in θ. □

Lemma B.10 (Strict mass monotonicity). *In Theorem B.9, suppose $x^* \neq 0$ and $\bar{\sigma} = \sigma_c^{(1)}(x^*)$ under θ is finite. Then $\sigma_c^{(1)}(x^*)$ under any $\theta' > \theta$ is strictly larger than $\bar{\sigma}$.*

Proof. Let $\tau = r\|x^*\| > 0$. The retention set is closed because f_θ is continuous, and every $\sigma > \bar{\sigma}$ escapes, so

$$f_\theta(\bar{\sigma}) \leq \tau \quad \text{and} \quad f_\theta(\sigma) > \tau \ \ (\sigma > \bar{\sigma}) \qquad \Longrightarrow \qquad f_\theta(\bar{\sigma}) = \tau > 0. \tag{12}$$

In particular $D^q(x^*) \neq x^*$ at $\bar{\sigma}$, and equation 9 gives

$$f_{\theta'}(\bar{\sigma}) = \frac{1-\lambda_{\theta'}}{1-\lambda_\theta} f_\theta(\bar{\sigma}) < \tau. \tag{13}$$

By continuity $f_{\theta'} < \tau$ on an interval above $\bar{\sigma}$. □

B.3 Two atoms (Theorem 5)

Theorem B.11 (Two atoms; formal version of Theorem 5). *Let $p = \theta\delta_{+a} + (1-\theta)\delta_{-a}$ with $a = d/2$ and $\theta \leq \frac{1}{2}$: the probe carries mass θ and its only neighbour sits at distance d. The probe carries its own attractor iff $\sigma < \sigma_c(\theta)$, where σ_c is strictly increasing on $(0, \frac{1}{2}]$, $\sigma_c(\frac{1}{2}) = d/2$, and*

$$\sigma_c(\theta) = \frac{d}{\sqrt{2\log(1/\theta)}}\,(1 + o(1)) \qquad \text{as } \theta \to 0.$$

Intuition. With two atoms the map lives on the line through them. There it is a sigmoid of the position, and the sigmoid's slope is the separation over the noise, $\beta = a^2/\sigma^2$, while its offset is the log-mass ratio h. A steep sigmoid crosses the diagonal three times: one attractor at each atom and a saddle between them. As σ grows the sigmoid flattens, and the crossing near the lighter atom disappears by merging with the saddle. The lighter atom survives while the slope beats the offset, which is why separation (through β) enters linearly and mass (through h) only logarithmically.

Proof of Theorem B.11. Place the atoms at $\pm ae$ for a unit vector e, with the probe at $+ae$ carrying mass $\theta \leq \frac{1}{2}$.

Step 1 (reduction to a line). By (F1) all fixed points lie on the segment $[-ae, ae]$. Write $x = aue + x_\perp$ with $x_\perp \perp e$. Then

$$\|x \mp ae\|^2 = \|x_\perp\|^2 + a^2(u \mp 1)^2, \qquad \|x + ae\|^2 - \|x - ae\|^2 = 4a^2u, \tag{14}$$

so the factor depending on $x_\perp$ cancels from the weights of equation 3, and

$$\frac{w_+(x)}{w_-(x)} = \frac{\theta}{1-\theta}\, e^{2a^2u/\sigma^2} = e^{2(\beta u + h)}, \qquad M_\sigma(x) = a\,(w_+ - w_-)\,e = a\tanh(\beta u + h)\,e, \tag{15}$$

with

$$\beta = \frac{a^2}{\sigma^2}, \qquad h = \frac{1}{2}\log\frac{\theta}{1-\theta} \leq 0. \tag{16}$$

M_σ depends on x only through u, and ∇M_σ vanishes orthogonally to e. Fixed points and their stability are therefore those of the scalar map $g(u) = \tanh(\beta u + h)$ on $[-1, 1]$.

Step 2 (fixed points). For $|u| < 1$,

$$u = g(u) \iff H(u) = h, \qquad H(u) = \operatorname{arctanh} u - \beta u, \qquad H'(u) = \frac{1}{1-u^2} - \beta, \tag{17}$$

and $H \to \mp\infty$ as $u \to \mp 1$. If $\beta \leq 1$, H is increasing and there is exactly one fixed point. If $\beta > 1$, let $u_\beta = \sqrt{1 - 1/\beta}$ and

$$h^*(\beta) := \beta u_\beta - \operatorname{arctanh} u_\beta > 0. \tag{18}$$

Then H increases on $(-1, -u_\beta)$ up to $h^*(\beta)$, decreases on $(-u_\beta, u_\beta)$ down to $-h^*(\beta)$, and increases on $(u_\beta, 1)$.

Step 3 (stability). At a fixed point $\tanh(\beta u + h) = u$, so

$$g'(u) = \beta(1 - u^2) \geq 0, \qquad g'(u) < 1 \iff H'(u) > 0. \tag{19}$$

Fixed points on the increasing branches of H are attractors, and a fixed point on the decreasing branch is repelling: it is the saddle between the two basins.

Step 4 (the probe's attractor). On $(u_\beta, 1)$, H takes every value in $(-h^*, \infty)$ exactly once. Since $h \leq 0$, the probe carries its own attractor iff

$$\beta > 1 \quad \text{and} \quad h^*(\beta) > |h|. \tag{20}$$

Otherwise the unique fixed point has $u \leq 0$, on the neighbour's side or at the midpoint.

Step 5 (the critical scale). Using $1/(1 - u_\beta^2) = \beta$,

$$\frac{dh^*}{d\beta} = u_\beta + \beta u'_\beta - \frac{u'_\beta}{1 - u_\beta^2} = u_\beta > 0, \qquad h^*(1) = 0, \qquad h^*(\infty) = \infty. \tag{21}$$

So $\sigma \mapsto h^*(a^2/\sigma^2)$ decreases strictly from ∞ at $\sigma \to 0$ to 0 at $\sigma = a$, and equation 20 holds iff $\sigma < \sigma_c(\theta)$, where $\sigma_c(\theta)$ is the unique solution of

$$h^*\big(a^2/\sigma_c(\theta)^2\big) = |h(\theta)| = \frac{1}{2}\log\frac{1-\theta}{\theta}. \tag{22}$$

The right-hand side decreases strictly on $(0, \frac{1}{2}]$, so $\sigma_c(\theta)$ increases strictly, and $h(\frac{1}{2}) = 0$ gives $\sigma_c(\frac{1}{2}) = a = d/2$.

Step 6 (asymptotics). As $\beta \to \infty$,

$$u_\beta = 1 - \frac{1}{2\beta} + O(\beta^{-2}), \qquad \operatorname{arctanh} u_\beta = \frac{1}{2}\log\frac{1+u_\beta}{1-u_\beta} = \frac{1}{2}\log(4\beta) + O(\beta^{-1}), \tag{23}$$

$$h^*(\beta) = \beta - \tfrac{1}{2}\log(4\beta) - \tfrac{1}{2} + O(\beta^{-1}). \tag{24}$$

As $\theta \to 0$, $|h| = \frac{1}{2}\log(1/\theta) + O(\theta) \to \infty$, so $\beta_c \to \infty$ and equation 24 gives

$$\beta_c = \tfrac{1}{2}\log(1/\theta)\,(1 + o(1)), \qquad \sigma_c = \frac{a}{\sqrt{\beta_c}} = \frac{d}{\sqrt{2\log(1/\theta)}}\,(1 + o(1)). \tag{25}$$

□

Remarks. (i) For $\theta > \frac{1}{2}$ the atoms swap roles, and bistability ends at the same scale: $\sigma_c(\theta) = \sigma_c(1 - \theta)$. (ii) The $\log\beta$ term in equation 24 makes the exact σ_c smaller than the leading-order formula: solving Step 5 numerically gives 0.74, 0.83 and 0.87 of the formula's value at $\theta = 10^{-2}, 10^{-4}, 10^{-6}$. So the formula of Theorem 5 gives how σ_c scales (linearly in d, logarithmically in θ), not its value. (iii) At σ_c the probe's attractor and the saddle merge at u_{β_c} and both disappear: a saddle-node. The attractor is still at distance $a(1 - u_{\beta_c})$ from the probe when this happens.

Theorem B.11 locates the scale at which an attractor ceases to exist, while Definition 3 thresholds a displacement. The next proposition says the two agree whenever the attractor is still within the escape radius when it vanishes.

Proposition B.12 (Escape test in the two-atom model)**.** *In the setting of Theorem B.11, let $K = \infty$ and let the escape radius $\tau = r\|x\|$ satisfy $\tau < a$. Then the probe is retained exactly for $\sigma \in (0, \sigma_{\text{esc}}]$, with $\sigma_{\text{esc}} \leq \sigma_c(\theta)$. Equality holds iff $a(1 - u_{\beta_c}) \leq \tau$, where $\beta_c = a^2/\sigma_c(\theta)^2$.*

Proof. g is increasing and $g(1) < 1$, so the orbit started at $u = 1$ decreases monotonically to the largest fixed point $u^*(\sigma) \leq 1$. Its largest displacement from the probe is its limit, so

$$\sup_k \|M_\sigma^k(x) - x\| = a\big(1 - u^*(\sigma)\big), \qquad \text{the probe is retained} \iff a\big(1 - u^*(\sigma)\big) \leq \tau. \tag{26}$$

For $\sigma \leq \sigma_c$, u^* is the probe's attractor, the root of $H(u) = h$ on the rising branch. As σ grows, H increases pointwise on $u > 0$, so this root moves continuously left, down to u_{β_c} at σ_c. For $\sigma > \sigma_c$, $u^* \leq 0$ and the displacement is at least $a > \tau$. So the retention set is

$$\big\{\sigma \leq \sigma_c : \; a\big(1 - u^*(\sigma)\big) \leq \tau\big\}, \tag{27}$$

an interval, which reaches σ_c iff the displacement at σ_c is at most τ. □

Since $1 - u_{\beta_c} \approx 1/(2\beta_c) \approx 1/\log(1/\theta)$, the condition holds for any fixed radius once θ is small. The escape test then returns the saddle-node scale. This is a statement about the exact map, not about a trained network.

B.4 From two atoms to a training set

Intuition. Remark (ii) and equation 24 say that, at leading order, the lighter atom keeps its attractor while it wins the posterior vote at its own location, $\beta > |h|$. For a whole training set the same criterion reads: the probe's m copies must outweigh the kernel mass of all other rows at the probe,

$$S(\sigma) := \sum_j m_j \, e^{-\|x - x_j\|^2/2\sigma^2} < m, \tag{28}$$

the sum running over the other rows. Each term of S increases with σ, so the balance holds on an interval $(0, \sigma_{\mathrm{bal}})$. This criterion is a heuristic. It is leading-order, it looks only at the probe's own location, and it has no escape threshold. What can be proved is how the closed forms of Section 2.1 relate to it.

Proposition B.13. *Let $n > m$ be the number of other rows, d_{nn} the distance to the nearest of them, and σ_1 the solution of $S(\sigma_1) = 1$. Set $d_{\mathrm{kern}} = \sigma_1\sqrt{2\log n}$. Then*

$$\sigma_{\mathrm{bal}} \;\geq\; \frac{d_{\mathrm{nn}}}{\sqrt{2\log(n/m)}} \qquad \textit{and} \qquad d_{\mathrm{kern}} \;\geq\; d_{\mathrm{nn}}, \tag{29}$$

with equality in both iff all other rows are at distance d_{nn}. Moreover, a single atom of n rows at distance d_{kern} has the same kernel mass as the actual rows at σ_1.

Proof. Every other row is at distance at least d_{nn}, so

$$S(\sigma) \leq n\, e^{-d_{\mathrm{nn}}^2/2\sigma^2}, \tag{30}$$

with equality iff all of them are at distance d_{nn}. The three claims follow from this bound:

$$n\, e^{-d_{\mathrm{nn}}^2/2\sigma^2} < m \iff \sigma < \frac{d_{\mathrm{nn}}}{\sqrt{2\log(n/m)}} \qquad \Longrightarrow \quad \sigma_{\mathrm{bal}} \geq \frac{d_{\mathrm{nn}}}{\sqrt{2\log(n/m)}}, \tag{31}$$

$$1 = S(\sigma_1) \leq n\, e^{-d_{\mathrm{nn}}^2/2\sigma_1^2} \iff d_{\mathrm{nn}} \leq \sigma_1\sqrt{2\log n} = d_{\mathrm{kern}}, \tag{32}$$

$$n\, e^{-d_{\mathrm{kern}}^2/2\sigma_1^2} = 1 = S(\sigma_1). \qquad \square$$

So equation 40, with $n \approx N$, is a lower bound on the balance scale: lumping every row at the nearest distance overcounts the competition. Lumping them at d_{kern} instead gives $d_{\mathrm{kern}}/\sqrt{2\log(n/m)}$, which is exact at $m = 1$ and approximate otherwise. In words, d_{kern} is the distance at which one atom carrying all other rows would weigh as much as the actual neighbours do against a single copy. Because the balance itself is only a heuristic, the ceiling σ_c^{emp} is not taken from either formula. It is computed by running equation 3 through the network's own escape test (Appendix D).

B.5 Conditional models and guidance (Section 2.1)

No argument above uses that p is a marginal distribution, so every result holds for $p(\cdot \mid c)$. For $p(x_0 \mid c) = \sum_i w_i(c)\delta_{x_i}$, equation 3 with $\mu_i = w_i(c)$ gives the caption-reweighted mean shift

$$M_\sigma(x; c) = \sum_i \tilde{w}_i(x, c)\, x_i, \qquad \tilde{w}_i(x, c) \propto w_i(c) \exp\!\big(-\|x - x_i\|^2/2\sigma^2\big). \tag{33}$$

The Gaussian kernel, which carries σ, is unchanged, and the caption enters only through the prior weights $w_i(c)$.

Corollary B.14 (Caption gap; formal version of Corollary 7)**.** *At $K = 1$, and when the caption changes the weight of x_i but not the relative weights of the other examples, $\Delta \log \sigma_c(x_i; c) > 0$ iff $w_i(c) > w_i(\varnothing)$.*

Intuition for Corollary B.14. A caption acts on the empirical mean shift only through the prior weights, exactly as a duplication count does. If it raises the weight of x_i and leaves the rest of the data as it was, it is a duplication of x_i, and Theorem B.9 applies.

Proof of Corollary B.14. By assumption, $w_j(c)/w_j(\varnothing)$ is the same for every $j \neq i$, so the normalized remainder is the same under both conditions:

$$p(\cdot \mid c') = w_i(c')\,\delta_{x_i} + \big(1 - w_i(c')\big)\,q_i, \quad c' \in \{c, \varnothing\}, \qquad q_i = \frac{\sum_{j\neq i} w_j(\varnothing)\,\delta_{x_j}}{1 - w_i(\varnothing)}. \tag{34}$$

These are the family p_θ of Theorem B.9 at $\theta = w_i(c)$ and at $\theta = w_i(\varnothing)$. By Lemma B.10, applied in whichever direction the weight moves,

$$\sigma_{\mathrm{c}}^{(1)}(x_i; c) > \sigma_{\mathrm{c}}^{(1)}(x_i; \varnothing) \iff w_i(c) > w_i(\varnothing), \tag{35}$$

when both are finite. □

Without the assumption, equation 9 still describes the step, but the caption also changes the pull of the remainder q_i.

Guidance. With the guided noise prediction $\epsilon_w = \epsilon_\varnothing + w(\epsilon_c - \epsilon_\varnothing)$, the map becomes

$$M_\sigma^{(w)}(x) = x + \sigma^2 \nabla \log \Big[p_\sigma(x)^{1-w} p_\sigma(x \mid c)^w\Big]. \tag{36}$$

It climbs the tilted density $p_\sigma(x)\big(p_\sigma(x \mid c)/p_\sigma(x)\big)^w$, whose critical points for $w > 1$ are not those of $p_\sigma(\cdot \mid c)$. Only $M_\sigma^{(1)}$ describes the fixed points of the conditional model.

Lemma B.15 (Guided map)**.** *The guided map is equation 36, and its Jacobian is*

$$\nabla M_\sigma^{(w)}(x) = \frac{(1-w)\mathrm{Cov}[x_0 \mid x] + w\mathrm{Cov}[x_0 \mid x, c]}{\sigma^2}. \tag{37}$$

For $0 \leq w \leq 1$ it is positive semidefinite; for $w > 1$ it is a difference of positive semidefinite matrices and may be indefinite.

Proof of Lemma B.15. Both denoisers are affine in the noise prediction, $D_\sigma(x) = x - \sigma\epsilon_\varnothing$ and $D_\sigma(x; c) = x - \sigma\epsilon_c$, so guidance with $\epsilon_w = (1-w)\epsilon_\varnothing + w\epsilon_c$ gives

$$M_\sigma^{(w)} = (1-w)\,D_\sigma(\cdot) + w\,D_\sigma(\cdot\,; c). \tag{38}$$

Applying Proposition B.3 to each term gives the map and its Jacobian. For $0 \leq w \leq 1$ the Jacobian is a convex combination of positive semidefinite matrices; for $w > 1$ the coefficient $1 - w$ is negative. □

The consequence for the dynamics: for $0 \leq w \leq 1$ everything in this appendix carries over to the tilted density $p_\sigma^{1-w} p_\sigma(\cdot \mid c)^w$. For $w > 1$ a negative eigenvalue lets the map overshoot and oscillate, and a maximum of the tilted density is an attractor only if every eigenvalue of its Jacobian stays above -1.

C Remarks on the critical scale

Computing the critical scale. We compute σ_{c} by bisection in $\log\sigma$ on the escape event of Definition 3. The definition is most natural when the set of retained scales is an interval, i.e. when the escape event is monotone in σ; in that case bisection returns σ_{c} exactly. We do not assume monotonicity but verify it on the trained models, and observe no violation on either model (on Stable Diffusion, none for 16 images over the noise grid). For $K > 1$ the escape event depends on the whole orbit, and σ_{c} has no closed form analogous to Proposition B.8.

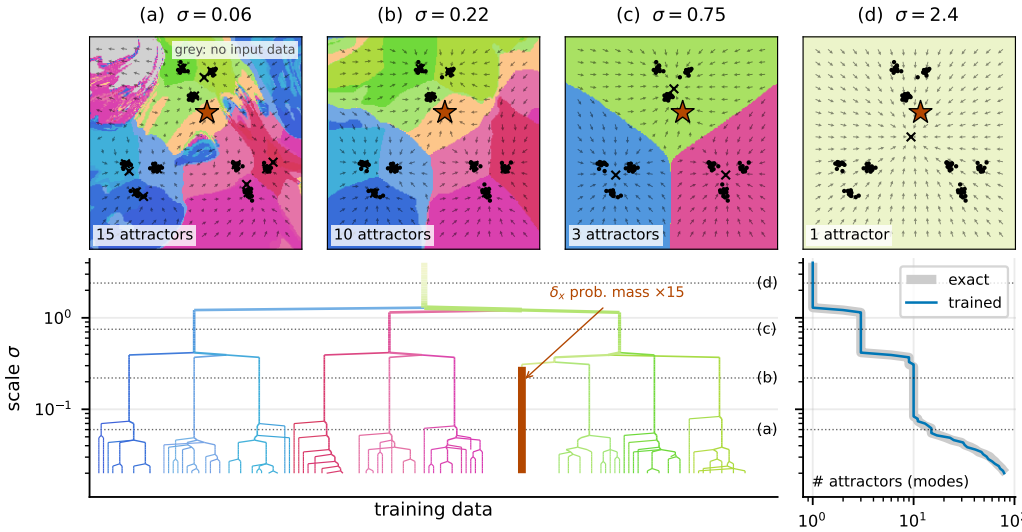

Figure 8: A trained denoiser reproduces the scale space of Fig. 3. Same training set and same panels as Fig. 3, with the exact map M_σ replaced by $M_\sigma(x) = D_\sigma(x)$ for an EDM-preconditioned MLP trained on the 141 points. *Top:* basins at the four scales of Fig. 3, coloured by the attractor ($\times$) reached from the data; grey marks the grid points that reach none of them, in the corners of the domain where the network saw no training data. *Bottom left:* merge tree of the trained map. *Bottom right:* its attractor count against that of the exact map. The two maps agree: the counts are equal at 80% of the 90 scales, the partitions of the training set they induce have adjusted Rand index 0.99 on average (0.76 at worst, at the smallest scales where single data split), and the duplicated datum keeps its attractor over the same range of scales in both.

Related iterations. The map of Definition 1 feeds the clean estimate back to the denoiser without adding noise, so that it is deterministic and noise cannot create spurious endpoints. Two related iterations have different fixed points and are not equivalent to M_σ. In the renoising iteration $x_{k+1} = D_\sigma(x_k + \sigma\varepsilon_k)$ the endpoints depend on the noise realisation, and the round trip $x_{k+1} = \mathrm{Sample}_{\sigma\to 0}(x_k + \sigma\varepsilon_k)$ is a generative operator.

D THE EXACT DENOISER AND THE RETENTION COEFFICIENT

The theorems of Section 2.1 describe the exact denoiser of a fixed distribution. The exact denoiser of a training set, M_σ^{emp}, is the multiplicity-weighted mean shift equation 3 over the number of samples: the map of a model that has stored its training set exactly. It costs one softmax over the training set per step, so it can be iterated from the same images and scored with the same escape test as the network (settings in Appendix F). We write

$$\sigma_{\mathrm{c}}^{\mathrm{emp}}(x) \;=\; \sigma_{\mathrm{c}}(x) \text{ computed with } M_\sigma^{\mathrm{emp}} \text{ in place of the network, at the same } K \text{ and } r. \tag{39}$$

Section D.1 checks the theory on this map; Section D.2 compares trained networks with it.

D.1 THE THEORY ON THE EXACT DENOISER

A bifurcation, not a threshold. For this map a training image is an exact fixed point at every scale below the transition, so its critical scale is the saddle-node of Theorem 5: on the duplication training set $\sigma_{\mathrm{c}}^{\mathrm{emp}}$ is unchanged between escape radii $r = 0.1$ and $r = 0.5$ and between $K = 20$ and $K = 500$ iterations (largest group-median difference 1%).

Distance and multiplicity. With the nearest training image as the only neighbour, Theorem 5 gives, for an example of multiplicity m among N rows,

$$\sigma_{\mathrm{c}}^{\mathrm{emp}}(x) \;\approx\; \frac{d_{\mathrm{nn}}(x)}{\sqrt{2\log(N/m)}}, \qquad d_{\mathrm{nn}}(x) = \min_{x_i \neq x} \|x - x_i\|, \tag{40}$$

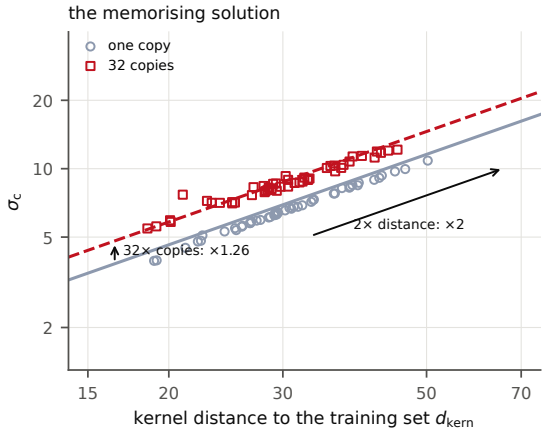

Figure 9: The exact denoiser follows the law. σ_c^{emp} of the added images of the rarity experiment against their kernel distance d_{kern} to the training set (Eq. 28), for images added once (grey) and 32 times (red), under the escape test with $K = 1$, $r = 0.2$. The lines are the law $\sigma_c = d_{kern}/\sqrt{2\log(N/m)}$ of Theorem 5, drawn without fitting. Doubling the distance doubles the critical scale, whereas 32 copies increase it by a factor 1.26.

refined by the balance equation 28 of Appendix B.4. The critical scale should thus grow linearly with distance and only logarithmically with multiplicity. The exact denoiser follows this law without fitting (Figure 9): doubling the kernel distance doubles σ_c^{emp}, whereas 32 copies increase it by a factor 1.26. The log–log slope in distance is 1.03 $[1.00, 1.05]$ for one copy and 0.88 $[0.79, 0.93]$ for 32 copies, against the predicted 1, and a joint fit over all 288 images of the rarity experiment gives 1.95 per doubling of distance against 1.06 per doubling of multiplicity, where the law predicts 2 and 1.05.

Which distance the theory means. Solving the balance equation 28 image by image predicts σ_c^{emp} at rank correlation 0.99–1.00 at every multiplicity of the rarity experiment, with an error of 7–10%, and on the duplication training set at $\rho = 0.99$ with a constant factor 0.83. The nearest-neighbour formula equation 40 reaches $\rho = 0.87$–0.95 with an error of 43%, and lies below the computed value, as Proposition B.13 requires. The gaussian kernel distance d_{kern}, is the notion of distance the theory uses.

D.2 Trained networks against the exact denoiser

The retention coefficient. A network should not hold an image longer than the memorizing solution does, which suggests reporting

$$\kappa(x) = \frac{\sigma_c(x)}{\sigma_c^{emp}(x)}, \tag{41}$$

the fraction of that solution's retention the network reaches at x: near 0 where the model generalizes through x, approaching 1 where it has stored x. Both terms are measured with the same test and the same threshold $r\|x\|$. A held-out image has no atom in the training set and leaves the exact map at the first step at every σ, so κ is defined for training images only.

The bound holds. We use the 1047 training images of Section 3: 264 from the duplication experiment, 495 from the overfitting experiment (five subsets of 100, excluding five images that escape at the lower end of the interval), and 288 from the rarity experiment. No image has $\kappa > 1$. The largest values are 0.78 (duplication), 0.82 (overfitting) and 0.65 (rarity). The memorizing solution therefore bounds every trained model we measure, and σ_c can be read as a position between a model that generalizes through the image and one that has stored it.

What moves, and what does not. Across the three interventions the ceiling σ_c^{emp} stays within a factor 4.2 per image and between 4.9 and 10.3 in group median, whereas σ_c varies by a factor 335 (Table 2). Each intervention raises κ, from 0.01–0.06 for images that are not memorized to 0.2–0.6 for images that are. The training set fixes the scale on which retention is measured, and training

Table 2: The three interventions on a common scale. σ_c^{emp} is the critical scale of the exact empirical denoiser of the training set of each model under the same escape test, equation 39; $\kappa = \sigma_c/\sigma_c^{emp}$ is the fraction of this solution that the network implements. Group medians, training images only: the ceiling varies little, and the effect is carried by κ.

intervention	group	σ_c^{emp}	σ_c	κ
duplication	$m = 1$	4.99	0.142	0.028
	$m = 8$	5.60	1.036	0.194
	$m = 800$	9.91	1.431	0.145
	$m = 800$, control model	9.91	0.269	0.024
overfitting	$N = 5 \cdot 10^4$	4.93	0.056	0.012
	$N = 2000$	6.11	0.243	0.043
	$N = 100$	7.84	4.674	0.598
rarity	model not trained on the image	7.68	0.451	0.058
	CIFAR-10 added, $m = 1$	7.28	0.447	0.076
	colour MNIST added, $m = 1$	6.85	1.459	0.182
	colour MNIST added, $m = 32$	10.30	4.024	0.432

The rarity rows use $K = 1$, $r = 0.2$ and the others the settings of Appendix F, so κ is comparable within a block but not across blocks.

decides how far along it the model goes; this is why the overfitting sweep moves σ_c by two orders of magnitude while the exact denoiser of its subsets barely moves.

The two maps release an image for different reasons. On the exact denoiser a training image is a fixed point until the saddle-node, so its critical scale does not depend on the iteration budget. On a network it does, and the dependence itself separates the groups: at $m = 800$ the median σ_c is 1.431 at both $K = 50$ and $K = 500$, whereas for $m = 1$ and held-out images it falls by factors of 165 and 193 over the same range. A memorized image is retained because it is a fixed point of the network's map; an ordinary image is retained only transiently, and a longer iteration releases it. Consequently κ compares two escape scales measured the same way, not two maps: at a fixed budget it is a conservative reading of how much of the memorizing solution the network implements, and for images without a basin it is an upper bound rather than an estimate.

E CIFAR-10: ADDITIONAL RESULTS

This appendix reports additional results for the experiments of Section 3. Training and evaluation details are given in Appendix F.

E.1 MEMORIZATION FROM DATA DUPLICATION (SECTION3.1)

Copying and sample quality. With the copy criterion of Carlini et al. (2023), 16 605 of 50 000 samples are copies. The FID increases from 3.90 of the control model to 4.70 for non duplicated samples (when this are counted in the FID computation FID goes to 26.46). Therefore aside for the planted memorized images the retained model still perform well on quality of generations.

Membership. σ_c separates training images that appear once from held-out (test set) images with AUC 0.528, i.e. it does not detect membership in this case: this is due to the fact that not accounting for the memorized content the models have similar FID and fit well the CIFAR data distribution therefore there is few difference between training and test distribution. This is consistent with a model that copies neither group, and agrees with the overfitting experiment at $N = 5 \cdot 10^4$. Figure 10 illustrates this on one image per group.

Dependence on multiplicity. The median σ_c relative to its value at $m = 8$ is:

m	8	25	50	100	200	400	800
network	1.00	1.12	1.18	1.34	1.24	1.26	1.38
exact empirical denoiser	1.00	1.11	1.08	1.23	1.19	1.29	1.77

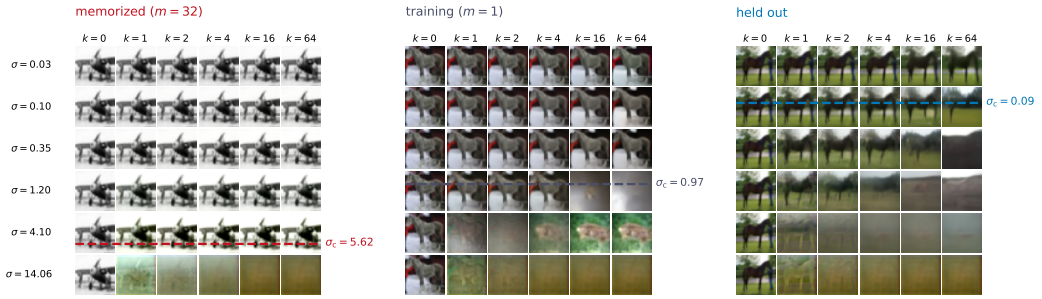

Figure 10: Orbits across scales, one image per group. Orbits of the model trained with duplicates, with σ fixed along each row and the iteration index k along each column, for a memorized image ($m = 800$), a training image with $m = 1$ and a held-out image, each chosen at the median σ_c of its group. The dashed line marks the σ_c of the image ($K = 64$, $r = 0.5$). The memorized image is unchanged after 64 iterations at $\sigma = 1.43$, whereas the other two converge to a uniform colour at $\sigma = 0.24$, with nearly identical critical scales (0.122 and 0.118).

A hundredfold increase in the number of copies changes the critical scale by less than a factor 1.4, and the memorizing solution of the same training set saturates in the same way: mass enters the width of a basin only logarithmically, so once an image has a basin, further copies cannot widen it much. Copies still change what the sampler does, as the copy rate shows. Accordingly κ does not order the duplicated images by multiplicity (0.028 at $m = 1$, 0.19 at $m = 8$, 0.15 at $m = 800$; $\rho = -0.01$), but it separates duplicated from non-duplicated images, and in the control model it lies between 0.014 and 0.037 at every m.

E.2 From memorization to generalization: memorization from overfitting

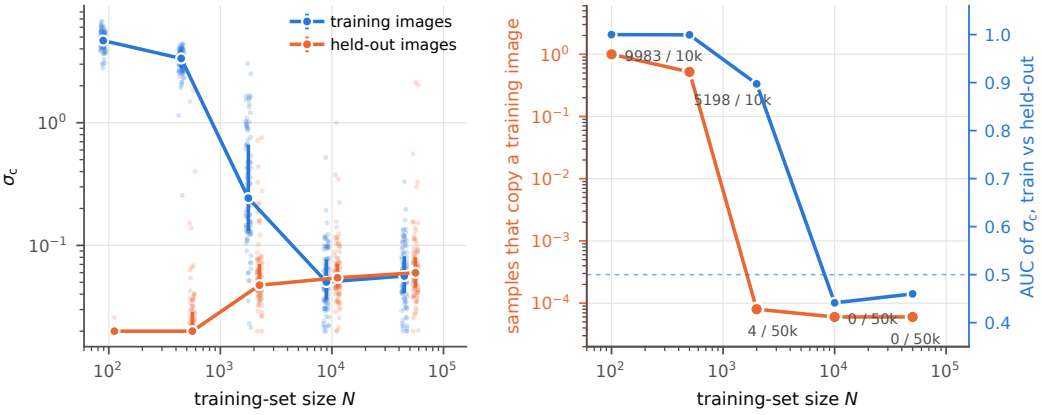

Figure 11: Overfitting: the same images across training-set sizes. *(a)* σ_c of the 100 training images shared by all models (blue) and of 100 held-out images (orange): medians with interquartile range, individual images in the background. Images that escape at the lower end of the bisection interval are drawn at that end, so the held-out medians for $N \leq 500$ are upper bounds. *(b)* Fraction of generated samples that are copies (orange, left axis; when no copy is found the point is drawn at the 95% upper bound) and AUC of σ_c between training and held-out images (blue, right axis). At $N = 2000$ sampling no longer detects memorization while σ_c still does.

Setup. We study memorization induced by training on a small dataset, and compare σ_c with memorization measured by sampling. We train five models with identical architecture, optimiser and training budget on nested subsets of CIFAR-10 of size $N \in \{100, 500, 2000, 10^4, 5 \cdot 10^4\}$ inducing a transition from a memorization (overfitting) to generalization phase (Fumero et al. (2026b); Kadkhodaie et al. (2024); Pham et al. (2025)). Since only N varies, a smaller N means more passes over each image. The first 100 images of the permutation belong to every subset, so σ_c is evaluated on the same 100 training images in every model, together with 100 held-out test images. Indepen-

dently of σ_c, we measure memorization as the fraction of generated samples that are near-copies of a training image. We also evaluate the public 50M improved-diffusion model on the same images.

Results. The median σ_c of the training images decreases by two orders of magnitude, from 4.67 at $N = 100$ to 0.056 at $N = 5 \cdot 10^4$, where it coincides with that of the held-out images (0.060; Figure 11a). The decrease is not a power law: σ_c is nearly constant up to $N = 500$, collapses between $N = 500$ and $N = 2000$, and is flat afterwards. At both ends σ_c agrees with sampling. For $N \leq 500$, 52–100% of the samples are copies and σ_c separates training from held-out images with AUC 1.00; for $N \geq 10^4$ and for the public model, no sample is a copy and σ_c is at the held-out level. The two measures differ at $N = 2000$: only 4 of $5 \cdot 10^4$ samples are copies, yet σ_c separates training from held-out images with AUC 0.897 (Figure 11b).

Every image here appears once, so the training set alone changes little across the sweep: what changes by two orders of magnitude is how closely the trained model reproduces the memorizing solution at these images (Appendix D.2).

Summary. Retention is lost later than generation: a model can keep its training images as attractors after its sampler has stopped producing them. Since σ_c measures retention, it detects memorization in a regime in which the copy rate is already zero.

Comparison to full OpenAi checkpoint. The public model behaves like our models with $N \geq 10^4$ (median 0.068 against 0.072, AUC 0.45). This is consistent with sampling: an exhaustive search finds no extractable training image, and the number of copies remains at the held-out level up to $3 \cdot 10^5$ samples. The images that sampling does recover from this model belong to clusters of near-duplicates in CIFAR-10, for which σ_c is at the held-out level.

Exact denoiser. Every image has $m = 1$, so equation 20 depends on N only through $\log N$ and predicts $\sigma_c(100)/\sigma_c(5 \cdot 10^4) = 1.53$, against a measured ratio of 83. The exact empirical denoiser of each subset agrees with the formula: its critical scale decreases from 7.84 to 4.93, a factor of 1.6. Held-out images, which have no atom, leave the exact map immediately (91 to 100 of 100 at every N), so their retention by the network is due to generalization.

E.3 Memorization of rare samples (Section 3.2)

Detection by sampling. No added image with $m \leq 4$ is copied more than 0.25 times per 10^5 samples for any source. At $m = 8$ the number of copies per image is between 1.3 and 13.6, and for $m \geq 16$ every added image is generated. For colour MNIST, σ_c flags 8 of 12 images with a single copy and all 12 with two.

Paired comparison. The median ratio between σ_c in the model trained on the image and σ_c in the other model increases from 1.02 (CIFAR-10, $m = 1$) through 2.5 (SVHN, $m = 4$) and 3.0 (MNIST, $m = 1$) to between 3.4 and 8 for $m \geq 8$. For the two CIFAR sources with $m \leq 4$, the unpaired AUC against images of the same source is between 0.36 and 0.66: these images cannot be detected in a single model, but they are detected by the paired comparison. Figure 12 shows three added images that are never sampled, one per source, chosen at the median of their group.

Isolation. Within a source and multiplicity, the increase of σ_c caused by adding an image is correlated with its distance to the nearest training image (Spearman $\rho = 0.33$, $p = 5 \cdot 10^{-5}$, $n = 144$ for $m \leq 4$). An isolated image does not share its basin at low noise with any training image, so a single copy suffices to create it; a typical image is absorbed into the solution that the model generalizes from its neighbours unless m is large. The probability that the sampler reaches the basin from pure noise, by contrast, grows with m and not with isolation, which explains why the two detection thresholds differ.

Distance and multiplicity in the trained models. The design spans five doublings of multiplicity and less than one doubling of distance, since the sources were matched in norm and total variation, so the two factors are compared per doubling. A joint fit over the 288 added images gives a factor 4.25 $[3.4, 5.3]$ per doubling of kernel distance against 1.49 $[1.44, 1.54]$ per doubling of multiplicity: in the trained models, as in the memorizing solution (Appendix D.1), distance is the *stronger* factor. The network responds to multiplicity more than the exact denoiser does, which is expected, since each copy also adds training steps on the image. Pixel distances are correlated with $\|x\|$ (see *Norm* below),

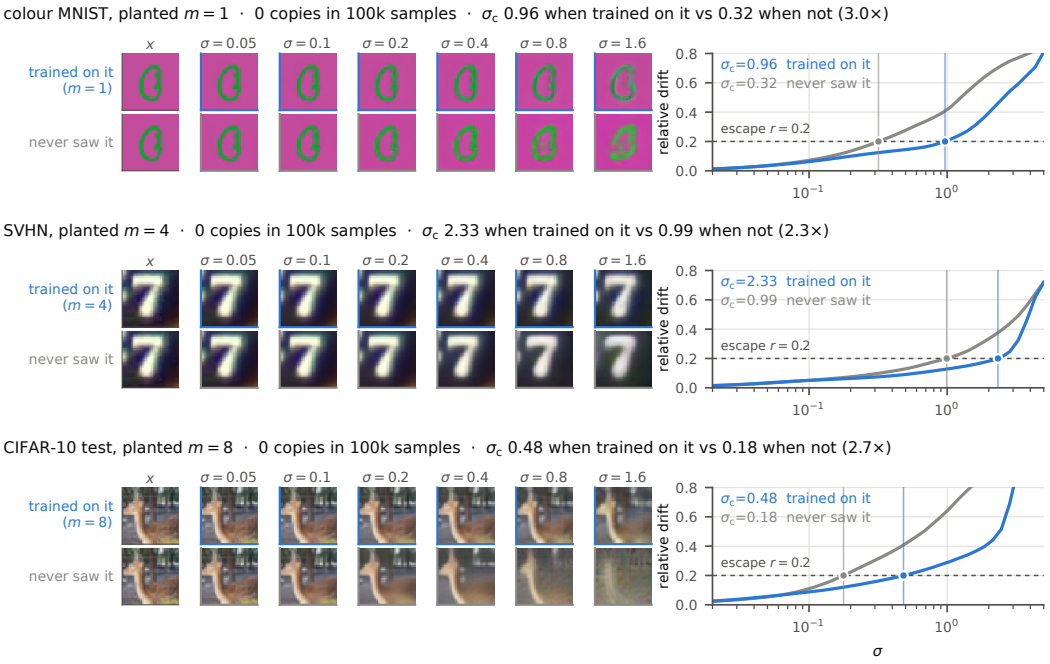

Figure 12: Paired comparison on three images that are never sampled. One added image per source, chosen at the median of its group. Each row shows $M_\sigma(x)$, one denoising step applied to the clean image at noise level σ, in the model trained on the image (top, blue) and in the model not trained on it (bottom, grey), together with the relative displacement $\|M_\sigma(x) - x\|/\|x\|$ whose crossing of r defines σ_c. The model not trained on the image moves it towards what it has learned (blur, generic texture) at a much smaller σ.

Agreement with the exact denoiser. Within a source and multiplicity the network orders the images as the exact map does ($\rho = 0.78$). The ordering of the sources is not an effect of the norm: at equal $\|x\|$, κ is larger than for CIFAR-100 by a factor 3.1 for colour MNIST and 2.2 for SVHN ($t = 13.7$ and 9.8), and by a factor 1.13 for CIFAR-10 ($t = 1.5$).

Escape radius. The radius $r = 0.02$ selected in advance on the public model is not suitable for this noise schedule: the escape occurs at $\sigma \approx 0.025$, below the informative range, and all AUCs except that of colour MNIST are close to 0.5. The results of Section 3.2 use $r = 0.2$, selected from a sweep over (K, r). The separation is stable for $r \in [0.1, 0.5]$ and $K \in [1, 20]$. The paired design determines the direction of the effect independently of this choice, since the same image is evaluated with the same r in two models.

F Experimental details

Noise-prediction models as denoisers. DDPM and Stable Diffusion corrupt the data as $x_t = \sqrt{\bar\alpha_t}\, x_0 + \sqrt{1-\bar\alpha_t}\,\varepsilon$ at discrete steps t and train $\epsilon_\theta(x_t, t)$ to predict ε. Dividing by $\sqrt{\bar\alpha_t}$ gives $y = x_0 + \sigma\varepsilon$ with $\sigma(t) = \sqrt{(1-\bar\alpha_t)/\bar\alpha_t}$, and the denoiser is $D_\sigma(y) = y - \sigma\,\epsilon_\theta(\sqrt{\bar\alpha_t}\, y, t)$.

CIFAR-10 duplication (Section3.1). The model with duplicates is trained with the improved-diffusion CIFAR-10 configuration (Nichol and Dhariwal, 2021), using the model and diffusion definitions of the original implementation: 50M parameters, $T = 4000$ steps with a cosine schedule, learned variances, dropout 0.3, 128 channels at three resolutions, learning rate 10^{-4}, batch size 128, EMA 0.9999, no horizontal flips, and 500k steps. The control is the released checkpoint `cifar10_uncond_50M_500K.pt`. Eight images are duplicated at each $m \in \{2, 8, 25, 50, 100, 200, 400, 800\}$ in addition to the 50 000 training images, giving 62 616 training examples. The multiplicities are chosen so that $m = 800$ corresponds to a per-image mass of 1.28%. Images in the lowest 5% of nearest-neighbour distance within CIFAR-10, i.e. the near-

duplicates already present in the dataset, are excluded from both the duplicated images and the $m = 1$ images. σ_c is computed on the 64 duplicated images, 200 training images and 200 held-out images by bisection on $[0.02, 12]$ with $r = 0.5$. Copies are counted with the criterion of Carlini et al. (2023) on 50 000 DDIM samples with 100 steps, and FID is computed against the training set with 50 000 samples per model. The two models are not exactly matched: the held-out loss of the model with duplicates is 1–4.5% higher than that of the released checkpoint at every t, and its gap between training and held-out loss is about 5%, against 0.8% for the released model, although it sees each non-duplicated image 20% less often. The released checkpoint should therefore be regarded as a reference trained without duplicates rather than as an exactly matched control.

CIFAR-10 overfitting (SectionE.2). Each model is a 25.8M-parameter DDPM trained for 15k steps at batch size 128, so the number of passes over the data decreases from 19 200 at $N = 100$ to 38 at $N = 5 \cdot 10^4$. The subsets are nested prefixes of a single permutation, so the subset with $N = 100$ is contained in all others. σ_c is computed by bisection on $[0.02, 20]$ with 12 steps, $K = 120$ and $r = 0.5$, on the first 100 images of the permutation and on 100 CIFAR-10 test images. For the copy rate we draw DDIM samples with 100 steps, 10^4 per model for $N \leq 500$ and $5 \cdot 10^4$ otherwise. A sample is a copy when the ratio of its distances to the nearest and second-nearest training images is below $1/3$; the same criterion applied to a held-out set of the same size finds no copies at any N. The public model is the improved-diffusion checkpoint `cifar10_uncond_50M_500K`, evaluated on the same images with the same settings.

CIFAR-10 rarity (Section3.2). The base set consists of $N = 10^4$ images from the same permutation, and the models use the same architecture and 15k training steps. The added images bring each training set to 11 512 examples. Since σ_c on CIFAR-10 depends strongly on $\|x\|$ and on typicality, unmatched images would obtain high values for trivial reasons; the added images are therefore matched to the base set in RMS norm and total variation, and images with a near-duplicate in the CIFAR-10 training set are excluded. Colour MNIST cannot be matched exactly (RMS 0.56 against 0.49, total variation 0.038 against 0.056) and should be regarded as the extreme of the atypicality range. We draw 10^5 DDIM samples per model ($\eta = 1$, 100 steps). A sample is a copy when its nearest image among the CIFAR-10 training set and all images of the four sources is at RMS distance below 0.07 on $[0, 1]$, and below 0.45 times the mean distance to the next 50 images; no image that was not added to a model was copied in $2 \cdot 10^5$ samples. σ_c is computed with $K = 1$ and $r = 0.2$ by bisection on the range of the noise schedule up to $\sigma = 20$, with 14 steps.

Filtering the memorization dataset. For the Stable Diffusion experiments, we use the matching-verbatim (MV) and template-verbatim (TV) examples released by Webster (2023). The collection contains many duplicate or near-duplicate entries, particularly among the TV examples, where the same memorized template appears with different visual patterns. To prevent these repeated templates from being overrepresented in our evaluation, we manually deduplicate the collection, retaining one MV copy per image and one TV example per template. We additionally remove a small number of mismatched image–caption pairs. The released collection associates each memorized image with its training caption, but does not guarantee that sampling from that caption reproduces that particular image. In several cases, the stored caption instead reproduces a different memorized image. Because our method evaluates memorization at the level of image–caption pairs, we exclude these mismatches. Figure 13 shows representative examples removed by both filtering steps.

Comparison to Wen et al. (2024). For the results reported in Table 1, we report the best-performing configuration for both variants of Wen et al. (2024): $n = 32$ samples with first and first 10 steps for $\|\epsilon_c - \epsilon_\varnothing\|$ and $\|\epsilon_c(x_0) - \epsilon_\varnothing(x_0)\|$, respectively.

Local $|\Delta \log \sigma_c|$. In order to retain the image local information, we apply the retention test of Definition 3 separately to each 4-dimensional patch x_u of the latent space, $||M_\sigma^k(x_u) - x_u|| \leq r||x_u||$. We first apply the denoising map for K iterations over a logarithmic grid of 24 noise levels. Then, we estimate the local σ_c^u as the geometric midpoint between the noise level that first escapes the radius r and the previous value in the grid, and define local $|\Delta \log \sigma_c|$ by averaging over locations. Unless otherwise stated, we use $K = 16$ and $r = 0.1$ for the local gap experiments.



(a) Duplicate examples (b) Mismatched image–caption pairs

Figure 13: Examples removed when filtering the memorization dataset. (a) Each row shows a group of duplicate examples, of which we retain only one. **(b)** The first image in each row is the stored training image, followed by three generations from its training caption. We remove cases in which the caption does not reproduce the stored image.

G Stable Diffusion: additional experiments

Orbits across the (σ, k) grid. Figure 14 expands Figure 1: for the same memorized and control pairs, every row is an orbit at one noise scale, so both the orbit (a row) and the end of the orbit across scales (the last column) can be read off one grid. More qualitative examples are in Appendix G.5.

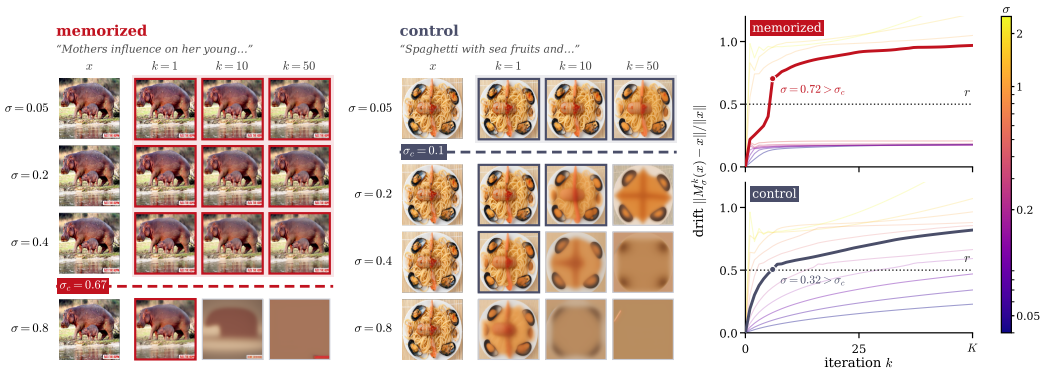

Figure 14: The full (σ, k) grid behind Figure 1: the critical scale on Stable Diffusion v1.4, for the same memorized caption–image pair (left) and control pair (middle), each run with its own caption at guidance scale 1. Each row iterates M_σ from the image x at one noise scale σ; a framed iterate $M_\sigma^k(x)$ is still within $r\|x\|$ of x. At small σ the model keeps returning x, which therefore sits in a basin of its own; at large σ the orbit drifts away towards content shared with other data. The critical scale σ_c (dashed line) separates the two regimes: the memorized image withstands far more noise than the control before it escapes. Right: the drift $\|M_\sigma^k(x) - x\|/\|x\|$ along the orbit, one curve per scale coloured by σ; the thick curve is a scale above σ_c and the dot marks where it crosses r. Here $K = 50$ and $r = 0.5$.

G.1 Ablation

Figures 15 and 16 assess sensitivity to the number of iterations K and escape radius r. Performance is stable around our primary choices: $K = 16$, with $r = 0.25$ for global scores and $r = 0.1$ for the local gap.

For $r = 0.25$, the global scores remain broadly stable through $K = 64$ and deteriorate at larger values. With $r = 0.1$, their performance is lower and begins to decline earlier; this sensitivity is particularly pronounced for σ_c on TV examples. By contrast, the local gap performs similarly for

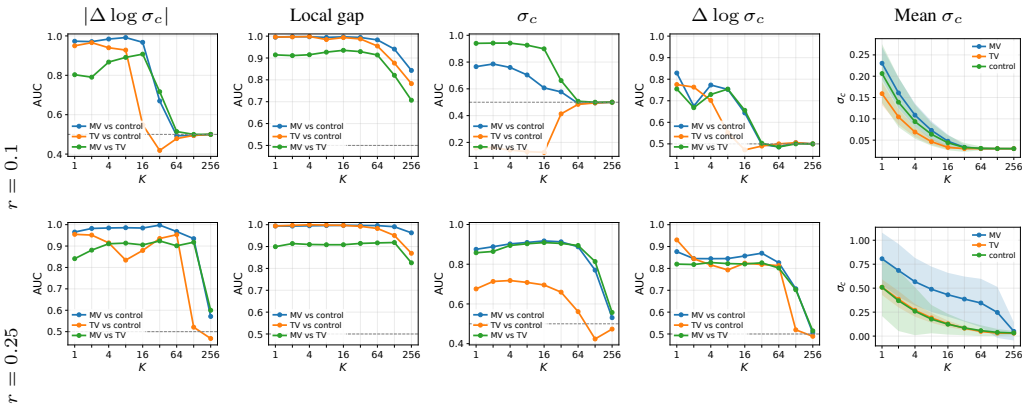

Figure 15: Ablation on the number of iterations K. Detection performance as a function of K for escape radii $r = 0.1$ and $r = 0.25$. The first four columns report AUC for the different critical-scale scores; the final column reports the mean critical scale (solid line) $\pm$ one standard deviation (shaded region). Performance is stable around the primary choice $K = 16, r = 0.25$.

both radii across most values of K, with only a modest decline beyond $K = 128$. Noticeably, mean σ_c decreases with K, reducing separation between memorized and control examples and causing the drop in detection performance at large K.

The radius ablation shows a similar pattern: global scores are more sensitive to r, especially on TV examples, whereas the local gap maintains strong performance as r varies by an order of magnitude.

The main results and the K-ablation estimate σ_c by log-space bisection. To evaluate many radii efficiently, the radius ablation instead uses a logarithmic grid of 24 noise levels and takes the geometric midpoint between the last retained level and the first escaped level. Overall, these results show that the reported performance is not specific to the selected hyperparameters and that the local gap is particularly robust.

G.2 Further experimental results

Figure 17 shows the score distributions for the main detection tasks using the best critical-scale metric in each setting. For separating memorized examples from controls, we use the local $|\Delta \log \sigma_c|$ with $K = 16$ and $r = 0.1$. For distinguishing MV and TV pairs from their caption-swapped counterparts, we use the signed $\Delta \log \sigma_c$ with $K = 16$ and $r = 0.25$.

The caption-swap distributions illustrate why the sign of the gap matters. Most swapped pairs have a negative gap, indicating that an unrelated caption weakens retention relative to the unconditional branch; this behavior is especially frequent for TV examples. Taking the absolute value folds these negative scores onto the positive axis, obscuring the separation between original and swapped pairs.

G.3 Detecting memorized regions: additional qualitative examples

Figure 18 presents four additional TV examples. As in Figure 6 (left), large local gaps align with content reproduced consistently across generations, whereas regions that vary tend to have gaps near zero. These examples further show that the local critical-scale gap captures the spatial extent of memorized content across different images.

G.4 Further analysis of mismatched caption gap

Ambiguity near zero. As shown in Figure 6 (right), the gap between mismatched-caption and unconditional critical scales is most informative when strongly negative. Values near zero, however, can represent two different regimes. Both branches may retain the image, indicating unconditional memorization and therefore yielding successful inpainting, or neither branch may retain it, indicating the absence of unconditional memorization. Because the gap measures relative retention, it cannot

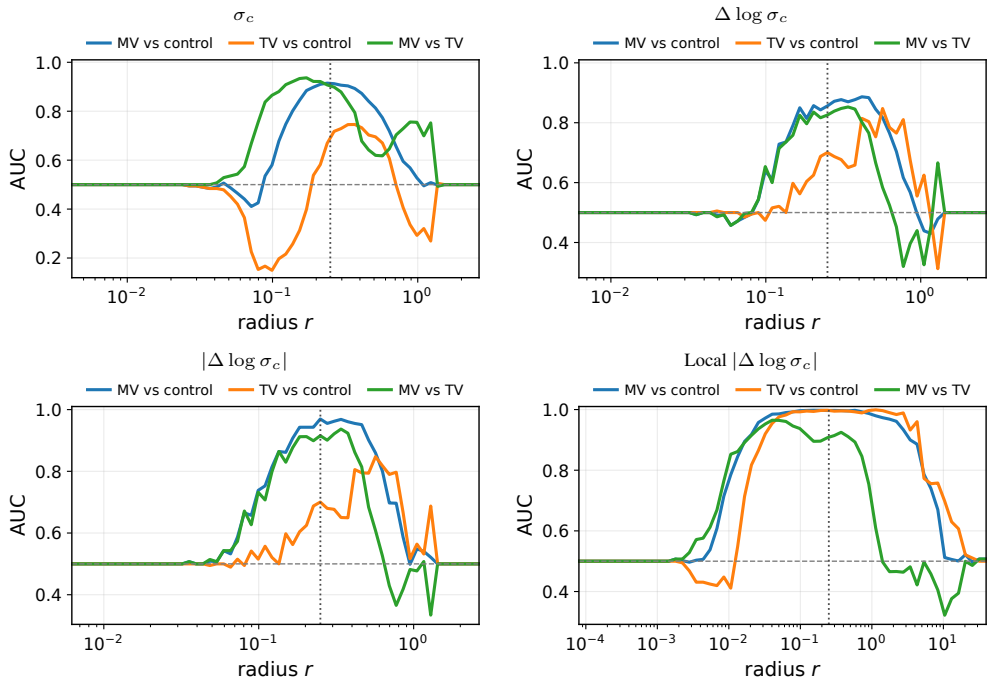

Figure 16: Ablation on the escape radius r. Detection performance (AUC) as a function of r using $K = 16$. For this ablation, σ_c is approximated over a logarithmic grid of 24 noise levels rather than estimated by log-space bisection. Performance is stable around the primary choice $K = 16, r = 0.25$ (global measures), $r = 0.1$ (local $|\Delta \log \sigma_c|$).

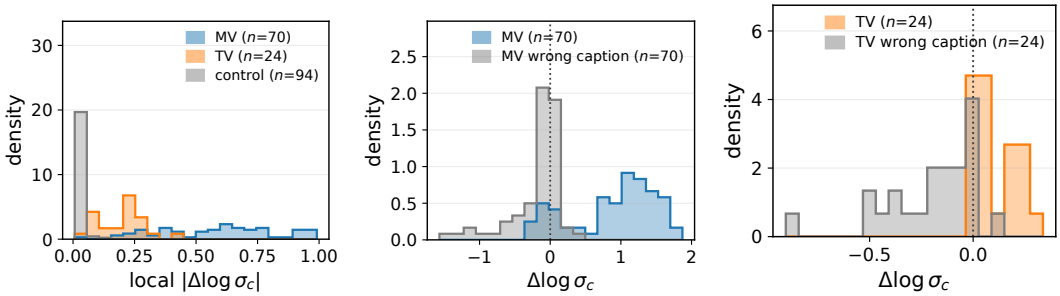

Figure 17: Critical-scale score distributions across evaluation groups. Histograms of the scores used to separate the different groups. *Left:* Local $|\Delta \log \sigma_c|$ for MV, TV, and control examples ($K = 16$, $r = 0.1$). *Center-Right:* Signed $\Delta \log \sigma_c$ for MV and TV pairs with their original and swapped captions ($K = 16$, $r = 0.25$). Swapped pairs frequently have negative gaps.

distinguish between these cases. Figure 19 (top) illustrates the first regime and Figure 19 (bottom) the second: despite having similar gaps, the examples have markedly different critical scales and inpainting performance.

Further examples with negative gap. Figure 20 provides two further examples with negative gaps. Both show evidence of unconditional memorization, but the example with the more negative gap yields more faithful reconstructions, consistent with stronger relative retention by the unconditional branch.

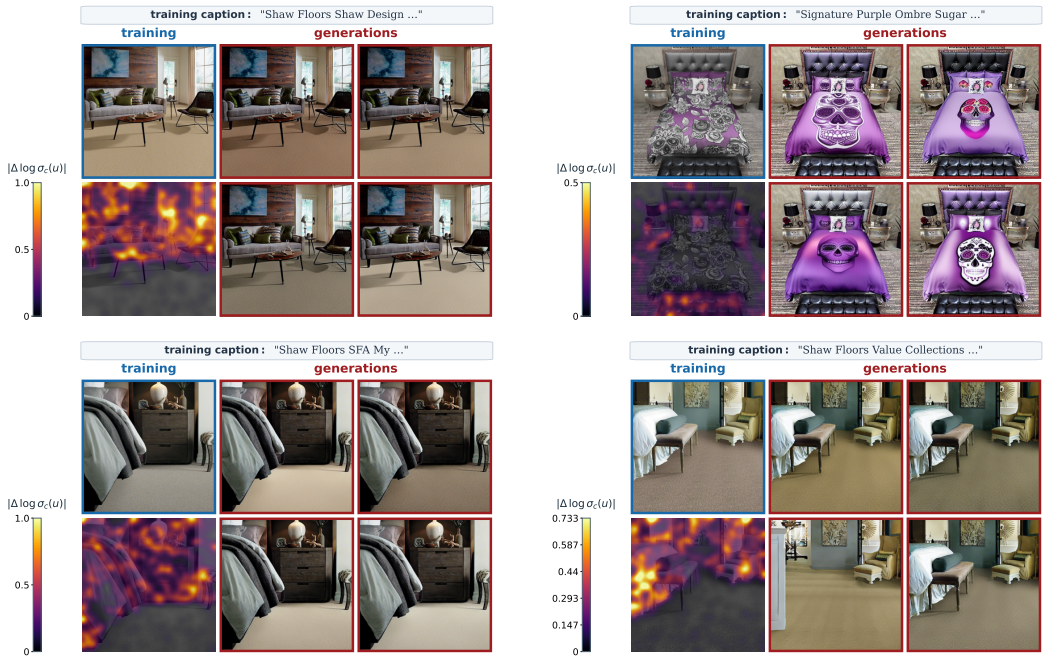

Figure 18: Additional examples of localized memorization. Each panel shows a TV training image, generations from its caption, and the corresponding local $|\Delta \log \sigma_c^u|$ map. Large gaps align with regions reproduced consistently across generations, whereas variable regions generally have gaps near zero.

G.5 Further qualitative examples

Figures 21–23 provide additional comparisons of fixed-scale trajectories and critical scales for matching-verbatim (MV), template-verbatim (TV), and control training images. Overall, retention is strongest for MV examples, followed by TV examples and then randomly selected controls: MV images remain close to their initial states for more iterations and at larger noise scales than TV images, which in turn persist longer than controls.

The TV trajectories also reveal spatial differences in retention. In the bedroom example of Figures 23 and 25, the surrounding room remains stable after the bedsheet pattern has been smoothed, consistent with the memorized and variable regions identified in Figure 18. Similarly, in Figure 22, the non-memorized curtain pattern disappears first, whereas the memorized plant and chair persist longer. These examples provide qualitative evidence that fixed-scale dynamics retain memorized regions longer than non-memorized ones, further motivating a local measure for partially memorized images.

Figure 24 illustrates how the caption gap distinguishes memorized image–caption pairs from controls. For the MV example, conditioning on the training caption retains the image to larger noise scales than the unconditional branch, producing a large positive $\Delta \log \sigma_c$. For the control, the two branches escape at similar scales, and the gap is close to zero.

Figure 25 further illustrates how retention varies across memorization types and image regions. In the upper comparison, the MV example has the largest σ_c, followed by the TV example and then the control. The lower comparison shows two TV examples: the image with memorized content occupying a larger region has the larger σ_c. Within both TV images, non-memorized content—the floor in the first and the bedsheet pattern in the second—disappears at smaller noise scales than the memorized regions.

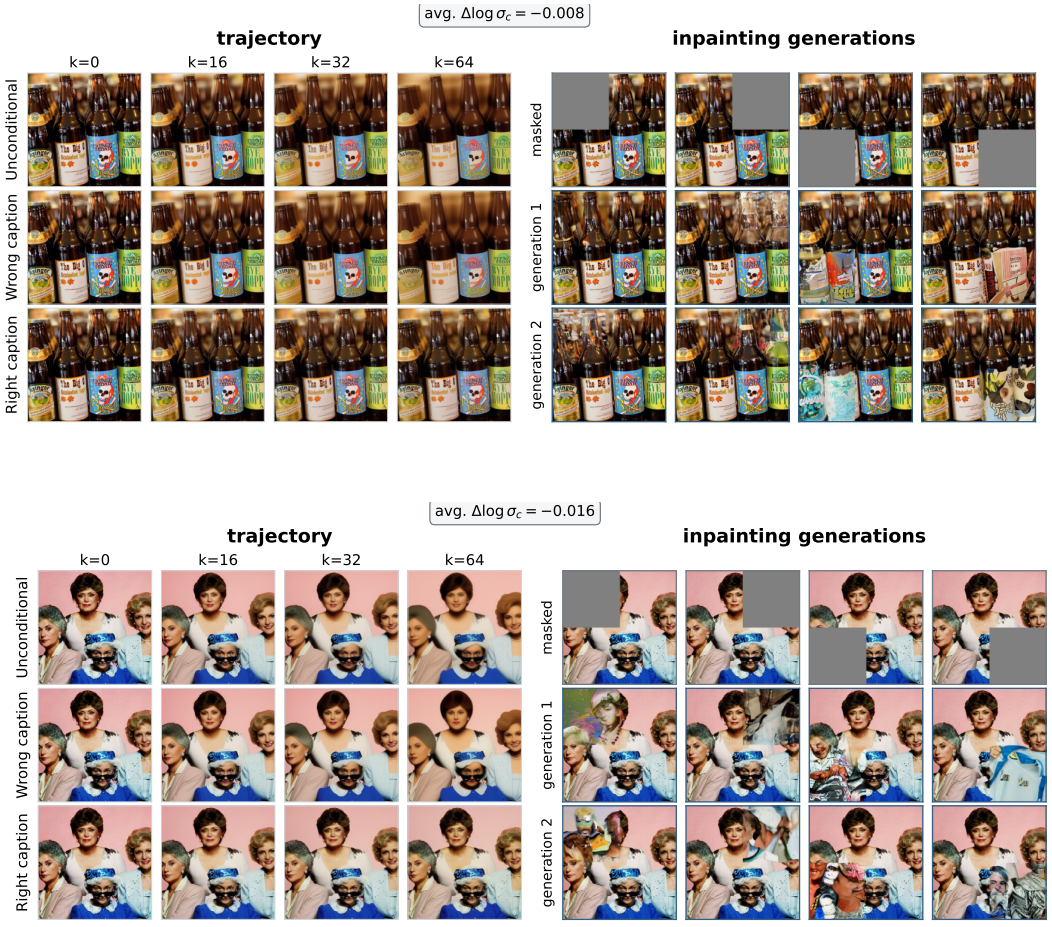

Figure 19: Near-zero gaps are ambiguous. The two examples have similar mismatched-caption gaps but different absolute retention. *Top:* Both branches retain the image, and unconditional reconstruction succeeds. *Bottom:* Neither branch retains the image, and unconditional reconstruction fails.

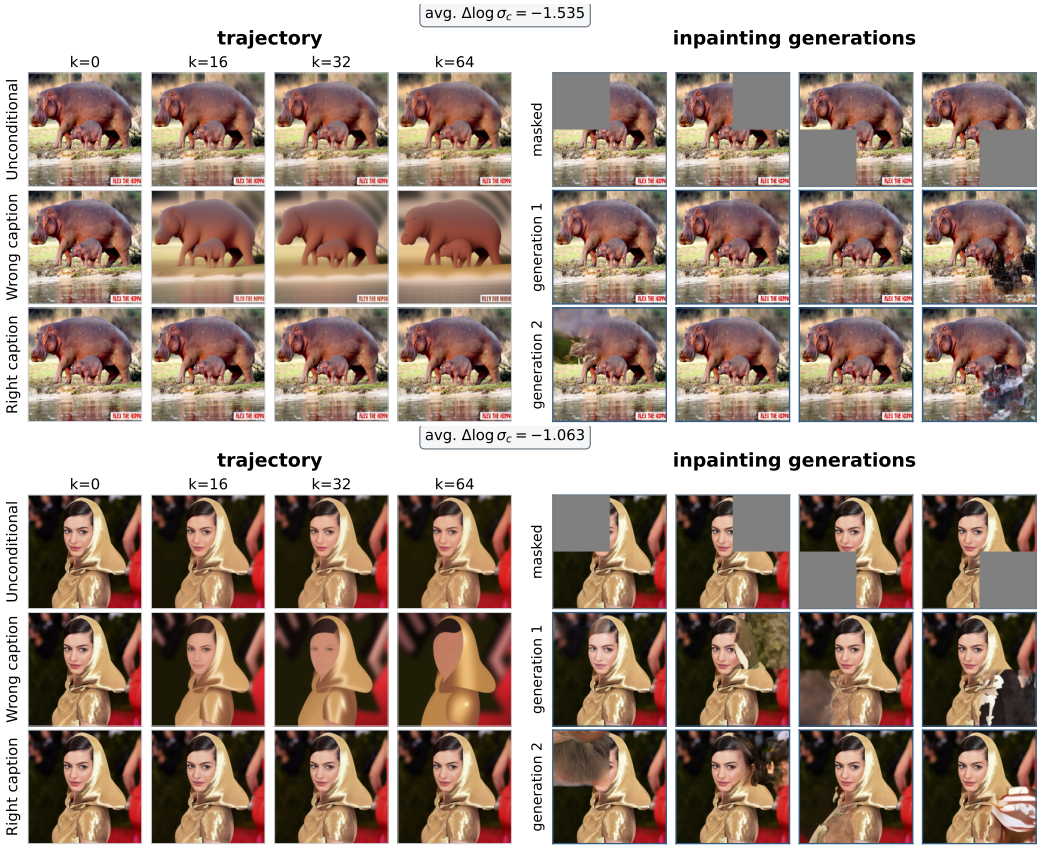

Figure 20: Additional examples with negative mismatched-caption gaps. Both examples exhibit unconditional retention and successful inpainting. The more negative gap in the top example corresponds to more faithful reconstructions than the less negative gap in the bottom example.

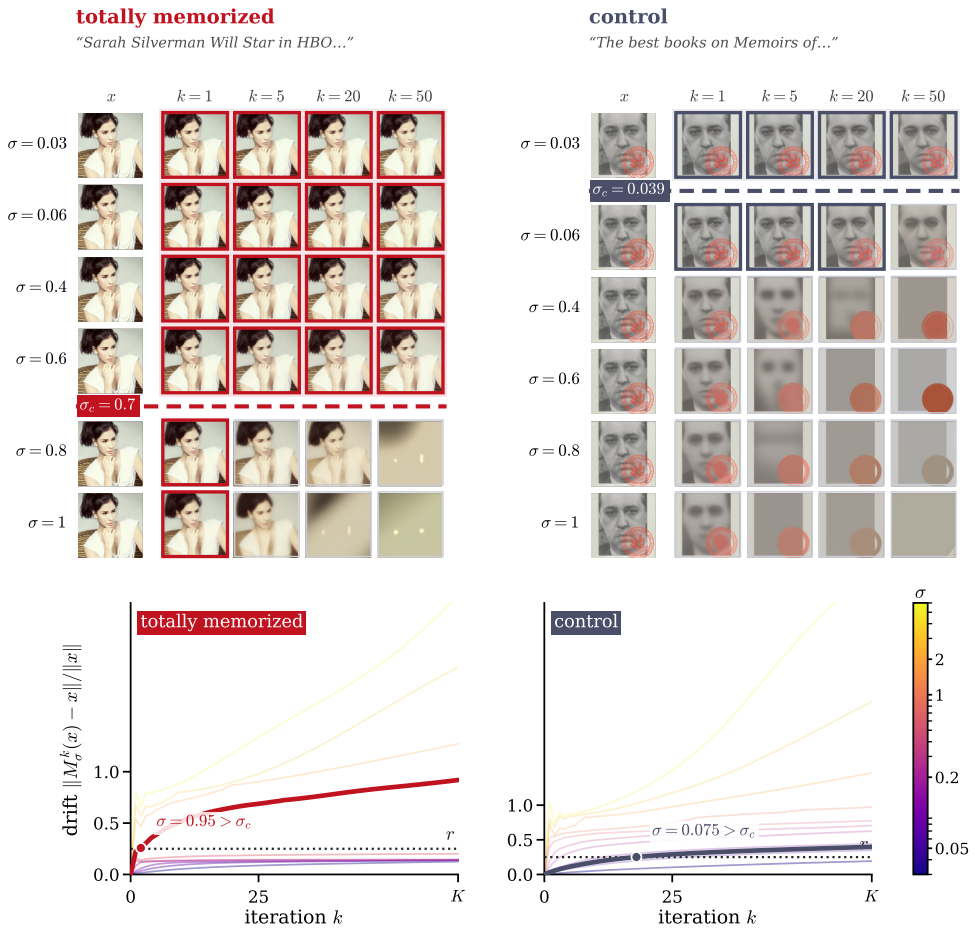

Figure 21: Trajectory and critical scale MV vs control. *Top:* critical scale on for a totally memorized caption–image pair (*left*) and a control pair (*right*), each run with its own caption at guidance scale 1. Each row iterates M_σ from the image x at one noise scale σ; a framed iterate $M_\sigma^k(x)$ is still within $r\|x\|$ of x. At small σ the model keeps returning x, which therefore sits in a basin of its own; at large σ the orbit drifts away towards content shared with other data. The critical scale σ_c (dashed colored line) separates the two regimes: the memorized image withstands far more noise than the control before it escapes. *Bottom:* the drift $\|M_\sigma^k(x) - x\|/\|x\|$ along the orbit, one curve per scale coloured by σ; the thick curve is a scale above σ_c and the dot marks where it crosses r.

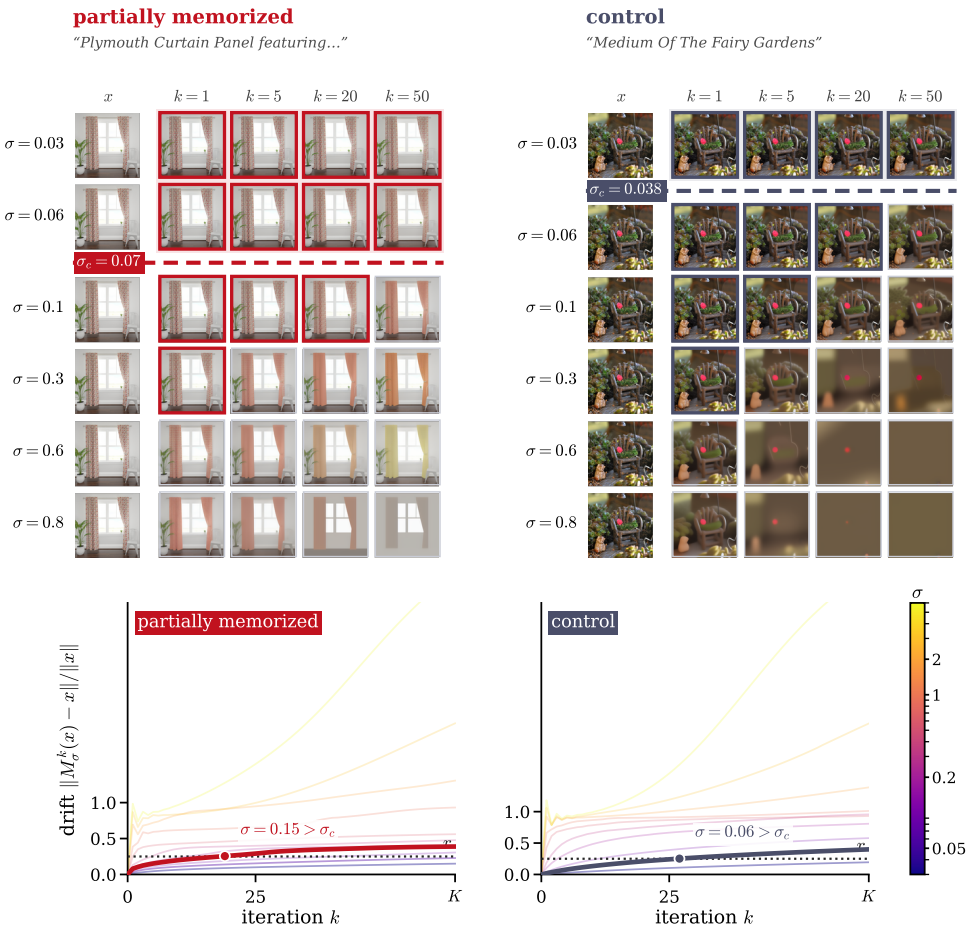

Figure 22: Trajectory and critical scale TV vs control. *Top:* critical scale on for a partially memorized caption–image pair (*left*) and a control pair (*right*), each run with its own caption at guidance scale 1. Each row iterates M_σ from the image x at one noise scale σ; a framed iterate $M_\sigma^k(x)$ is still within $r\|x\|$ of x. At small σ the model keeps returning x, which therefore sits in a basin of its own; at large σ the orbit drifts away towards content shared with other data. The critical scale σ_c (dashed colored line) separates the two regimes: the memorized image withstands far more noise than the control before it escapes. *Bottom:* the drift $\|M_\sigma^k(x) - x\|/\|x\|$ along the orbit, one curve per scale coloured by σ; the thick curve is a scale above σ_c and the dot marks where it crosses r.

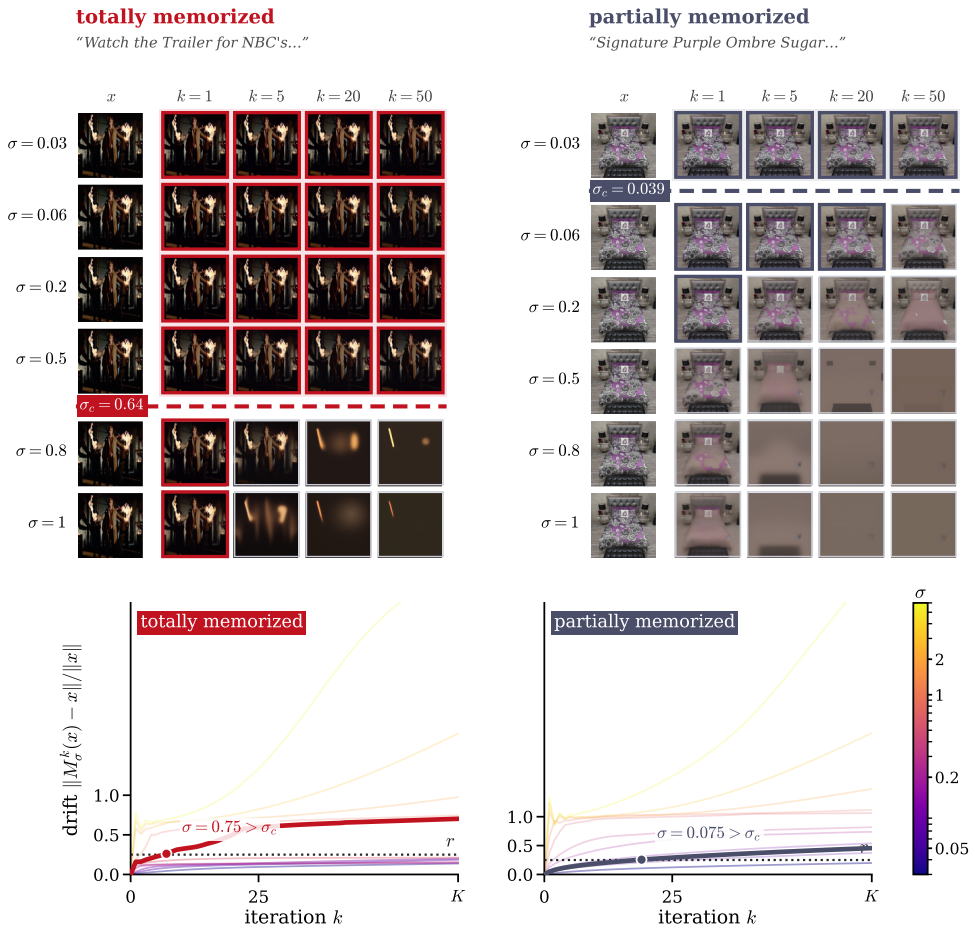

Figure 23: Trajectory and critical scale MV vs TV. *Top:* critical scale on for a totally memorized caption–image pair (*left*) and a partially memorized pair (*right*), each run with its own caption at guidance scale 1. Each row iterates M_σ from the image x at one noise scale σ; a framed iterate $M_\sigma^k(x)$ is still within $r\|x\|$ of x. At small σ the model keeps returning x, which therefore sits in a basin of its own; at large σ the orbit drifts away towards content shared with other data. The critical scale σ_c (dashed colored line) separates the two regimes: the memorized image withstands far more noise than the control before it escapes. *Bottom:* the drift $\|M_\sigma^k(x) - x\|/\|x\|$ along the orbit, one curve per scale coloured by σ; the thick curve is a scale above σ_c and the dot marks where it crosses r.

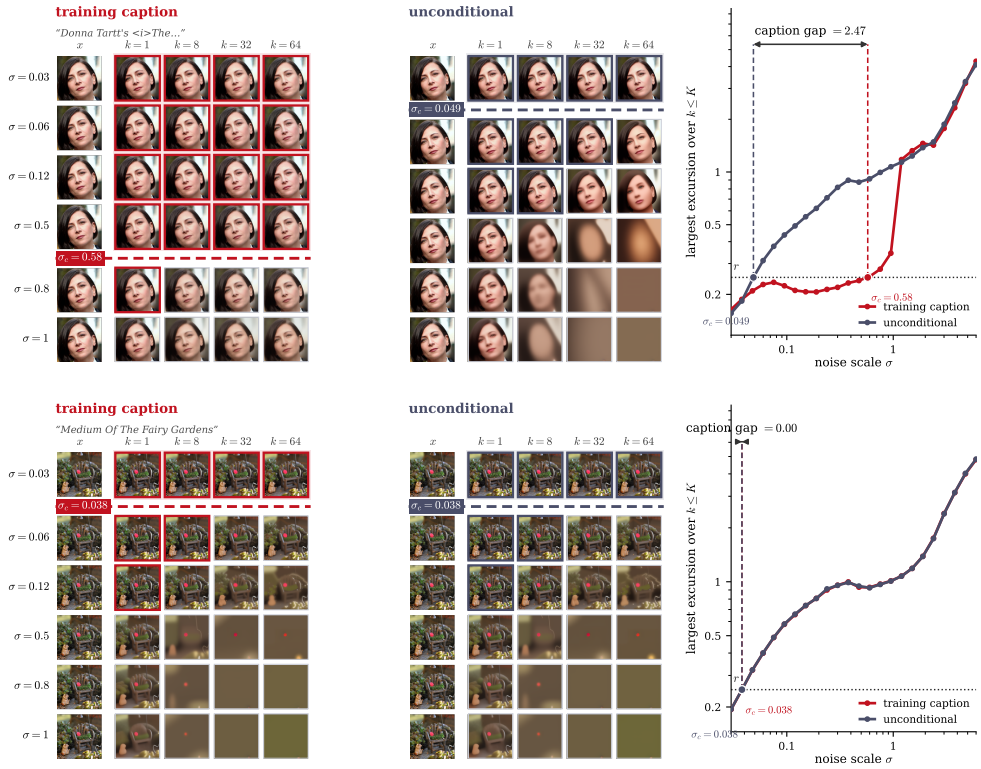

Figure 24: $\Delta \log \sigma_c$ for MV and control images. The top row shows an MV image–caption pair and the bottom row a control. The left and center panels show fixed-scale trajectories across noise levels under training-caption and unconditional conditioning, respectively. The right panels plot $\max_{k \le K} ||M_\sigma^k(x) - x||/||x||$ for both branches. Their crossings with the dotted threshold r define the conditional and unconditional critical scales, whose log difference is $\Delta \log \sigma_c$. The MV pair has a large gap because its caption substantially increases retention, whereas the control has nearly equal critical scales and a gap close to zero.

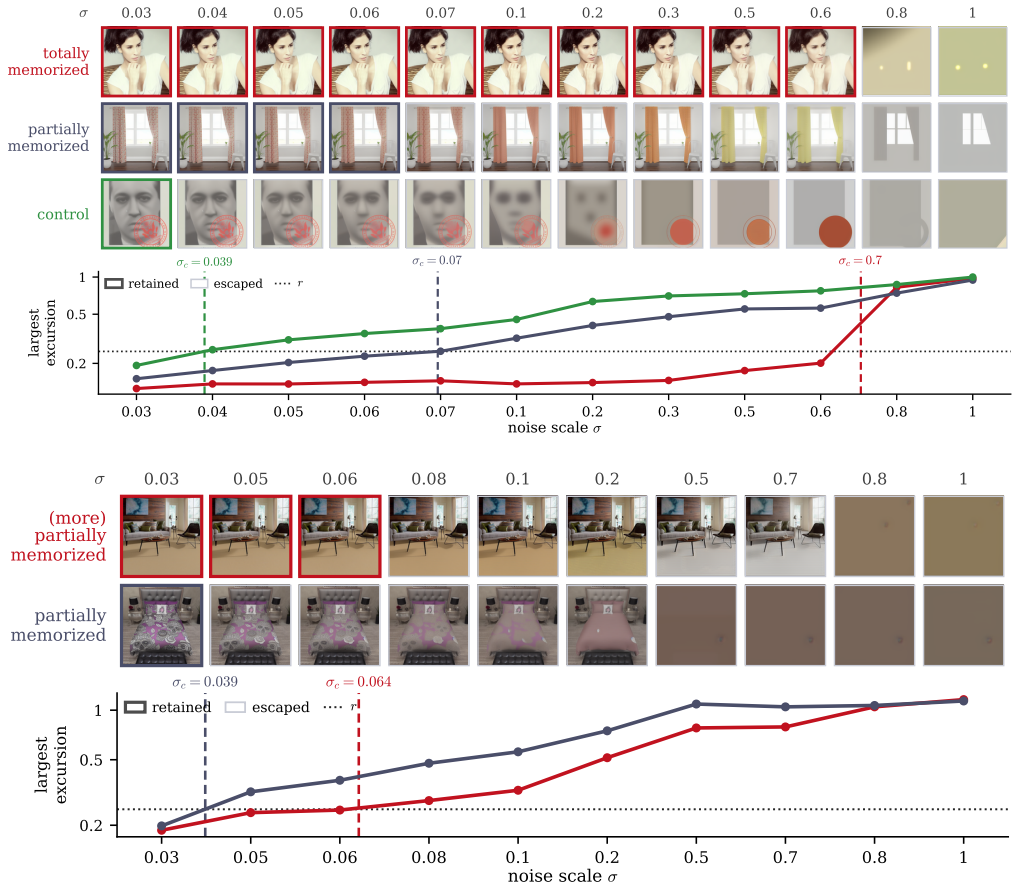

Figure 25: Fixed-scale trajectories across noise levels. Each comparison shows the iterate $M_\sigma^K(x)$ across noise scales (top) and the maximum normalized displacement $\max_{k \le K} |M_\sigma^k(x) - x|/|x|$ as a function of σ (bottom). The dotted line marks the escape radius r; its crossing determines σ_c. The upper panel compares an MV, a TV, and a control image, ordered from longest to shortest retention. The lower panel compares two TV images with different amounts of memorized content. The image with more memorized content is retained longer.