ANCHORED OR DRIFTING: WHAT RECURSIVE SELF-GENERATION REVEALS ABOUT TRAINING DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Large generative models are known to memorize their training data, posing severe privacy risks. Yet, current methods to detect training membership typically rely on the weak signals of a single forward pass. In this work, we find that training samples and unseen (held-out) data follow visibly different trajectories under recursive self-generation – repeatedly feeding a model's output back as its next input. Held-out samples *drift*: they lose the specifics of the original within a few steps. Training samples stay *anchored*, degrading far more slowly. We show that the membership signal this produces holds across model modalities, architectures and scales, spanning language, diffusion, and autoregressive vision models. Furthermore, these recursive trajectories provide a signal that raises membership inference TPR at 1% FPR for nearly every attack we evaluate, roughly doubling it on the weakest baselines and still improving the strongest, which shows that a model's behavior under recursion carries membership evidence that a single query does not.

1 INTRODUCTION

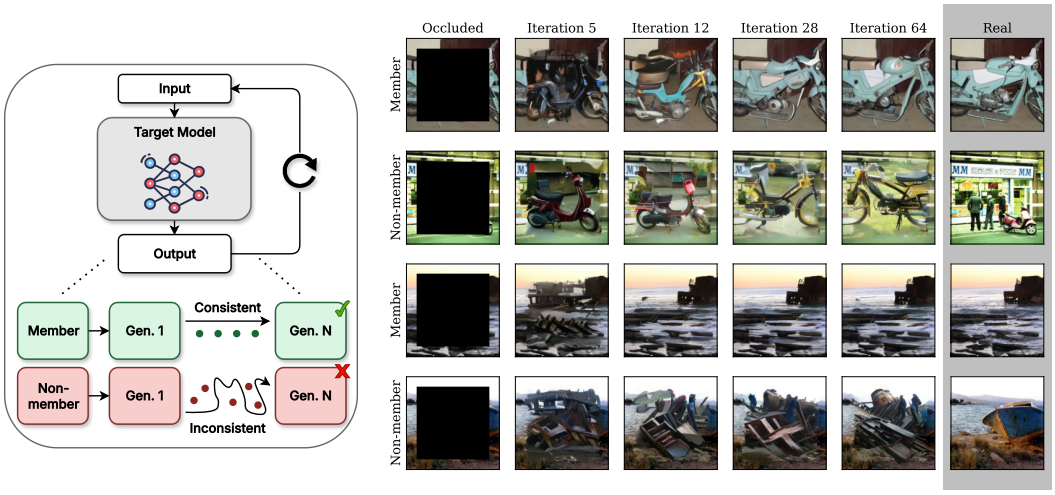
The rapid development of generative AI has triggered a pressing demand for training data, frequently leading to the unauthorized usage of private, sensitive, or copyrighted content. Determining whether specific samples or datasets were used to shape a model's parameters is therefore central to practical auditing, from protecting medical privacy (Zhang et al., 2022) to identifying licensed content (Dubiński et al., 2025) and detecting benchmark contamination (Maini et al., 2024; Singh et al., 2024; Zawalski et al., 2026). Membership inference attacks (MIAs) (Shokri et al., 2017) and dataset inference (DI) (Maini et al., 2021) formalize it, and a substantial body of work has developed increasingly refined scoring functions for these settings (Zhang et al., 2024; WU et al., 2024; Choquette-Choo et al., 2021). What has not changed is the auditor's position: the model is trained, its weights and its training data out of reach, and all that remains is how it processes the information.

Both difficulties come from the fact that what a model has taken from its training data is distributed over billions of weights during training, which is difficult or expensive to repeat. Shadow-model attacks (Carlini et al., 2022; Ye et al., 2022) ask for that repetition, training auxiliary copies of the target to calibrate against, which is unaffordable at scale. A single-query attack asks instead for all of it in a single answer, and recent evaluations show its limitations, since many MIAs degrade sharply under distribution shift, at times barely beating chance (Maini et al., 2024; Dubiński et al., 2025).

We think an audit at large scale needs two things. First, with no shadow model available, it has to be self-referential – the only thing left to compare a model against is the model itself. Second, since one answer does not exhaust what the weights hold, it has to read more than one. Both follow from a recursive self-generation procedure, in which the model generates from the sample, then from that generation, and so on for several iterations; at every iteration it has nothing to condition on but its own previous output, so the trajectory that results is shaped entirely by what its weights already encode. One number becomes a trajectory, and the trajectory is what we study.

In those trajectories we observe a consistent asymmetry. One that starts from a sample the model was trained on (a member) stays close to it, while one that starts from a sample the model has not seen (a non-member) does not. Within a few iterations, the specifics of the real sample are gone, and it is drawn toward the training data itself, settling near samples the model was fitted on rather than near the one it started from – visible in Figure 1b and measured in Figure 3. We call the two ends of this behavior *anchored* and *drifting*, and measure where a sample falls between them. The asymmetry

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.



(a) Recursive self-generation.

(b) Anchored vs. drifting, at 60% occlusion.

Figure 1: **Anchored or drifting under Recursive Self-Generation (RSG). (a)** Each iteration feeds the model's output back as its next input. With no external reference to correct toward, the trajectory is driven entirely by what the weights already encode. A training sample's trajectory stays near it; a held-out sample's trajectory is pulled toward the training manifold too, but to a neighborhood that is not its own. That difference is the signal we measure. **(b)** One example of the two behaviors, with 60% of each image removed before the trajectory starts: both populations produce coherent images of the right class, but the training samples return to the particular image that was occluded while held-out samples settle on a different instance. The rightmost column shows the true original.

holds across image autoregressive models, diffusion models, and large language models, and the gap is already measurable after a handful of iterations (Figure 2).

Recursion of this kind has been studied before, as a pathology. Model Autophagy Disorder (Alemohammad et al., 2023; Shumailov et al., 2024) describes how a *sequence* of models retrained on synthetic data loses variance and collapses toward degenerate approximations of the data distribution. Ours is a single frozen model, with no retraining and recursion applied only at inference – *what is a failure mode during training becomes a measurement instrument after it*. We read the statistics of the trajectory to improve membership (MIA) and dataset inference (DI) attacks across model families and modalities. It costs T generation calls per sample and nothing else – no shadow models, and no access the base attack does not already have.

In summary, our contributions are:

- We identify an asymmetry in recursive self-generation: a member's trajectory stays anchored near the sample it started from, while a non-member's drifts toward the training data instead. Anchoring is graded rather than binary, and we show how to read it off a trajectory. The asymmetry holds across image autoregressive, diffusion, and language models.
- We fuse trajectory features into existing attacks rather than replacing them, and the gain shows up at the low false-positive rates: TPR at 1% FPR roughly doubles, improving in 118 of 122 attack – model pairs, while AUC moves by about 5%.
- We demonstrate that the gain is not a query-budget artifact. Spending equivalent compute on independent perturbations of the sample yields inferior results, confirming that the recursive trajectory captures membership evidence missed by isolated queries.

2 Related Works

Memorization. Memorization in generative models – the tendency to reproduce training examples rather than generate novel samples – has been studied across model families, and the measurements are largely static. Early work formalized the distinction between memorization, mode collapse, and

overfitting (van den Burg & Williams, 2021); subsequent studies characterized the generalization-to-memorization transition in diffusion models (Gu et al., 2023), localized it through attention patterns (Sakarvadia et al., 2024), and showed that standard evaluation metrics fail to surface it (Bai et al., 2021). Mitigation has been proposed for both LLMs (Hans et al., 2024) and text-to-image models (Chen et al., 2025). What these measures share is that they evaluate a model in a single forward pass. We instead read what a model retains of individual samples from its behavior over repeated self-generation, which surfaces structure that is not visible in any single pass.

Model Collapse. Training a generative model on synthetic data has been studied as a training-time pathology. Alemohammad et al. (2023) show that self-consuming loops progressively degrade quality and diversity when insufficient real data is injected, a phenomenon they term Model Autophagy Disorder, and Shumailov et al. (2024) show that recursive training on model-generated data causes tails of the original distribution to disappear. Both concern a *sequence* of models retrained on synthetic data. We study a single frozen model, with no retraining, and apply recursion at inference time. Collapse is drift of a whole distribution across training generations; what we measure is whether an individual sample drifts or stays *anchored* within the inference-time dynamics of one fixed model.

Membership and Dataset Inference. Individual Membership Inference Attacks (MIA) can be confounded by distribution shifts (Maini et al., 2024), prompting a shift toward Dataset Inference (DI), which aggregates evidence across many samples (Maini & Suri; Dubiński et al., 2025; Kowalczuk et al., 2025). Shadow model approaches (Ye et al., 2022; Carlini et al., 2021) are now computationally infeasible for large architectures, so modern attacks extract signals from limited black-box outputs (Zhang et al., 2024; Chang et al., 2025; Tao & Shokri, 2025). Most relevant is Li et al. (2024), which averages independent perturbations of a target image. Because prior methods rely on isolated queries rather than a recursive feedback loop, our trajectory features capture a missing signal. Rather than competing with existing attacks, we seamlessly fuse our features into them.

3 METHOD

3.1 RECURSIVE SELF-GENERATION

One step of the procedure maps a sample to another sample of the same type – a reconstruction for image models, a continuation for language models (Table 1). We start from the real sample, and then the model's output becomes its own next input. Formally, given a sample $Z_0 = x$ and a trained generative model f, we construct

$$Z_{t+1} = \mathcal{G}(f, Z_t), \qquad t = 0, \ldots, T-1,$$

where $\mathcal{G}$ is a modality-specific generation operator: the procedure needs nothing of f beyond its ability to generate a sample of the same type, which is what makes it applicable to any generative model and across modalities. The same operator and the same settings are used at every step, and the original sample Z_0 is never reintroduced after $t = 0$. We refer to $(Z_0, \ldots, Z_T)$ as the *trajectory* of x under f.

3.2 INSTANTIATING $\mathcal{G}$

Table 1 summarizes the generation operator for each model family. In all cases, the operator is deterministic in its settings: sampling parameters, noise levels, and decoding configuration are fixed at $t = 0$ and held constant for the remainder of the trajectory.

For image autoregressive models (IAR), one step is image-to-image generation under fixed decoding. For diffusion models, one step is partial forward noising to a fixed level followed by reverse denoising under a fixed scheduler. For language models, Transformer and state-space alike, one step is a sliding-window continuation: the model generates m tokens conditioned on the current k-token prefix, and the window advances by dropping the oldest m tokens and appending the newly generated ones. The context length stays fixed at k while generated text progressively displaces the original prefix, so that after $\lceil k/m \rceil$ steps no original tokens remain.

We define *generation strength* as the magnitude of the model's intervention at each recursive step, quantifying the balance between the information preserved from the previous state and the new

Table 1: **The generation operator $\mathcal{G}$ across model families.** One step maps a sample to a new sample of the same type. Every setting, the strength s included, is fixed at $t = 0$ and held fixed for the remainder of the trajectory. The original sample is never reintroduced.

Family	One step of $\mathcal{G}$	Generation strength s	Held fixed	Access
IAR	Regenerate part of the token representation and decode to an image	Number of final scales (VAR) or fraction of tokens (RAR) regenerated	Sampling temperature, class conditioning	Black-box
Diffusion	Forward-noise to level σ_s, then reverse-denoise to $t = 0$	Re-noising level σ_s	Scheduler, step count, guidance scale	Black-box
LLM	Generate m tokens from the current k-token prefix; advance the window by m	Sampling temperature	Window length k, stride m, decoding configuration	Grey-box

information introduced from the model's learned distribution. Each operator carries a single control s governing how much of Z_t is discarded and resampled: the number of final scales (VAR) or fraction of tokens (RAR) re-generated for IARs; the re-noising level σ_s for diffusion models; and the temperature parameter, for language models. Larger s means less of the previous state is carried over.

3.3 Measuring anchoring and drift

Let $\delta_t := \delta(Z_0, Z_t)$ measure how far the trajectory has moved from the sample it started from. Under recursion δ_t rises and then flattens, so a trajectory is summarized by two numbers: the *contraction rate* $\lambda \in [0, 1]$ at which it flattens, smaller meaning stronger anchoring, and the *drift floor* δ_∞ it flattens at. A sample is *anchored* when δ_∞ is small – the model's own dynamics hold it near where it started – and *drifting* when it is large, so the two words name the ends of a scale rather than two classes, and the samples an audit finds hardest sit between them. Corollary J.5 shows these are not free parameters of the measurement: at a memorized point $\lambda = 1 - s^2 m$, with s the generation strength we choose and m the local sharpness of the model's density there. The trajectory features we feed the scorer estimate λ, δ_∞, or functions of both. The sequence $\{\delta_t\}_{t=1}^{T}$ is what the term *anchoring* refers to throughout the paper, and it is measured per sample.

We instantiate δ per modality: MSE, LPIPS (Zhang et al., 2018), and SSIM (Wang et al., 2004) for images, and 12 features for language models, e.g. Mean Reciprocal Rank, Edit distance, or Jensen-Shannon divergence. We provide detailed feature descriptions in Appendix E.1 and E.2.

3.4 From anchoring to membership inference

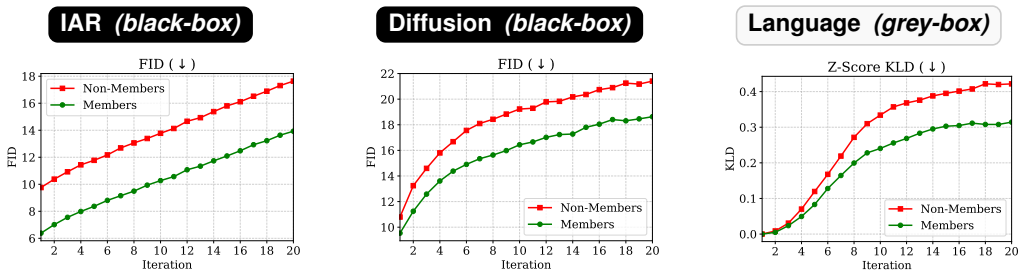
We score each sample twice under an identical procedure, once from the single-query evidence alone, and once from the same evidence together with the trajectory it generates, and compare the two. The scorer, the calibration protocol, the splits, and the base attacks are the same in both cases.

Let $z_{\text{base}} = \phi_{\text{base}}(Z_0) \in \mathbb{R}^d$ denote the score produced by any single-query attack. We form the fused representation $\tilde{z} = [z_{\text{base}} \| z_{\text{traj}}] \in \mathbb{R}^{d+k}$ and fit an L_1-regularized logistic regression h as a plug-in estimator of $P(M = 1 \mid \tilde{z})$, giving the final score $s(Z_0) = h(\tilde{z})$. For dataset inference, the procedure is identical, with the membership label replaced by dataset origin $D \in \{0, 1\}$ and per-sample scores aggregated by a fixed mean rule. Feature orientation, normalization, and the regression itself are fit on the training split alone and applied unchanged at test time. The auditor requires only the access needed by the base attack plus T generation calls per sample. Because recursive self-generation relies entirely on standard model outputs, it functions as an any-box method. The effective access level is therefore inherited directly from the fused attack.

4 Experiments

4.1 Experimental Setup

To ensure a scientifically sound evaluation across our MIA tasks, we restrict our setup to models trained on public datasets with well-defined training and test splits. We evaluate our method on three



(a) VAR-d30 (b) DiT-MoE-G (c) Pythia-6.9B

Figure 2: **Population-level divergence under recursive self-generation.** Each curve tracks a whole set of samples: at every iteration we compute FID (vision) or KLD (language) between the set of iterates and the corresponding set of originals. These are distributional distances between populations. The member population diverges more slowly than the non-member population throughout, establishing the asymmetry that the per-sample measurements exploit.

model families spanning two modalities, vision and language, to demonstrate its broad applicability. For image generation, we analyze SOTA autoregressive models (VAR-d{20, 24, 30} (Tian et al., 2024), RAR-{L, XL, XXL} (Yu et al., 2024)) and diffusion models DiT-MoE-{XL, G} (Fei et al., 2024), UViT-T2I-Deep (Bao et al., 2023)), trained primarily on the ImageNet (Deng et al., 2009) or COCO (Veit et al., 2016) datasets. We extend this evaluation to the language domain utilizing prominent LLMs (Pythia-{6.9, 12}B (Biderman et al., 2023), GPT-NeoX-20B (Black et al., 2022), GPT-Neo-2.7B (Black et al., 2021), GPT-J-6B (Wang & Komatsuzaki, 2021), and the state-space model Mamba-2.8B (Gu & Dao, 2023)). We leverage the established Mimir (Duan et al., 2024) dataset on *ngram_7_0.2* split to evaluate language models. Comprehensive details regarding all specific models and datasets used in experiments are provided in the Appendix C and D.

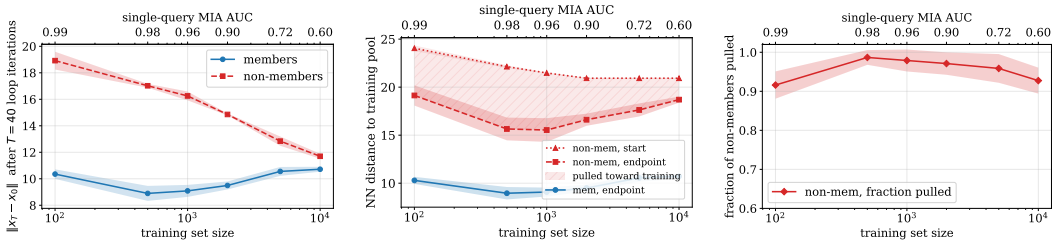
4.2 Protocol and Metrics

Metrics. In the main part of our experiments we report TPR@1%FPR as the most important metric for MIAs, since an audit is only useful at an operating point where a positive call can be acted on, and threshold-free summaries average over regimes an auditor would never use; AUC is reported beside it in the appendix (Table 10, Table 11). To measure the similarity between feature representations and their fidelity, we utilize the Fréchet Inception Distance (FID) (Heusel et al., 2017) for vision, with $\Delta\text{FID} = \text{FID}_{\text{nonmem}} - \text{FID}_{\text{mem}}$ for the population gap. For language models we use Kullback–Leibler Divergence (KLD) between the normalized average token probability distribution at the current iteration and that of the first evaluation iteration. For additional insights into trajectory dynamics we also report generative precision and recall (Kynkäänniemi et al., 2019) metrics: precision is the share of generated samples that fall inside the manifold of real ones, recall the share of real samples covered by the generated manifold. We also use Recovery@1 (Section 4.5) – the rate at which the final iterate ranks the true original first among same-class distractors.

MIA pipeline. For each labeled member/non-member sample, we generate $Z_0, \ldots, Z_T$, compute ϕ_t, S_T, and modality-specific trajectory features, form $\tilde{z}$, and fit h_{mia}. We evaluate univariate trajectory statistics (as described in Section 3.3) by direct thresholding and multivariate features by logistic-regression fusion on strictly stratified train-test splits. We report AUC and TPR@1%FPR. All descendants of the same Z_0 (all Z_t, all derived features) remain in the same partition. Thresholds, feature normalization, and LR calibration are fit on train only.

Applying $\mathcal{G}$ repeatedly to a sample produces a trajectory that either stays near where it started or does not, and which of the two occurs is governed by whether the sample was in the training set. Figure 2 shows this for one model from each family. Divergence from Z_0 grows in both cases (recursion is lossy), but it grows consistently slower for training samples, and the gap widens within the first few iterations and persists to the end of the trajectory.

The mechanism is visible in what recursion removes. Recursive self-generation has no external reference to correct toward, because at every step the model can only condition on its own previous output, so whatever survives must be sustained by its weights alone. Samples memorized by the



(a) Endpoint displacement $\|x_T - x_0\|$ (pixels). Members stay *anchored*; non-members *drift*, traveling up to $2\times$ further.

(b) Nearest training-image distance, start vs. end. Non-members are *pulled toward training*; members stay on the training manifold.

(c) Fraction of non-members ending closer to the training pool than x_0 did. Remains ≥ 0.9 across all training-set sizes.

Figure 3: **Geometry of frozen DDPM trajectories across training-set size.** Small-CNN DDPMs trained from scratch for 400 epochs on $\{100, 500, 1000, 2000, 5000, 10000\}$ disjoint CIFAR-10 images. The architecture is identical across runs, so the degree of memorization is set by training-set size alone; the top axis of each panel reports the resulting single-query MIA AUC (pixel-MSE at $\mathrm{denoise_steps} = 10$), which declines as the training set grows. Every trajectory is iterated at $\mathrm{denoise_steps} = 2$ for $T = 40$ iterations; shaded bands are ± 1 SD across 5 seeds.

model have somewhere to be sustained, and the trajectory returns to them, while samples it has not seen have no such point, and the trajectory settles wherever the model's priors place it.

4.3 Anchoring and Drift in Controlled DDPMs

We further present a small, fully controlled experiment to test whether *repeatedly* passing an image through a generative model's own inference pipeline – rather than looking at a single reconstruction error – exposes membership in the training set. To that end, we train a family of small CIFAR-10 DDPMs from scratch at six deliberately different levels of memorization and probe each of them with recursive self-generation under a frozen SDEdit operator (i.e., a fixed-weight model whose sample is repeatedly re-noised to an intermediate diffusion step t_{edit} and denoised back, so no training takes place between iterations). Our setup lets us disentangle two questions that are usually confounded: how memorization shapes the geometry of the trajectory, and how much of the resulting membership signal can be recovered by a classical attacker versus one that reads the whole trajectory. Here the generation strength s of Table 1 is instantiated as $\mathrm{denoise_steps}$, the DDIM reverse-diffusion budget spent on each iteration, which sets how much of the input the model rewrites per pass.

4.4 Training Data Stays Anchored Under Recursive Self-Generation

Members are fixed points, non-members are transported. Recursion induces a direction asymmetry between the two populations (Fig. 3). A member image barely moves – the endpoint sits on average $\|x_T - x_0\| \approx 10$ pixels from the input, essentially flat across the memorization range – while a non-member is transported 1.5–$2\times$ further and, crucially, not in a random direction: for every training-set size, over 90% of non-member trajectories terminate closer to the training pool than they began (Fig. 3c), and their nearest-neighbour pixel distance to that pool shrinks along the trajectory, while members are on the pool from the outset and stay there (Fig. 3b).

The same trajectories also carry the membership signal. A logistic regression over the sequence of per-step origin distances lifts TPR@5%FPR from 0.43 at $K = 1$, which is the classical single-query attack, to 0.99 at $K = 15$, while AUC is already above 0.9 at $K = 1$ and merely saturates (Appendix I).

4.5 Recursion Returns to Specific Samples

Under class-conditional generation, a trajectory might stay close to Z_0 simply by producing a generic sample of the correct class. If most of a sample is destroyed before the trajectory begins, does the model reconstruct *that sample* or *its class*? We test this by heavily occluding the image and forcing the model to inpaint the missing content.

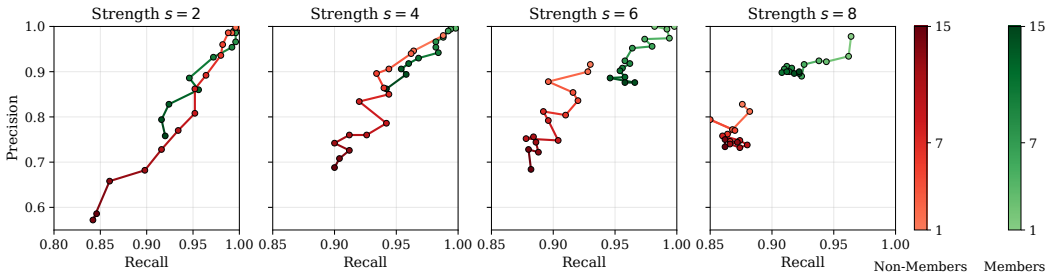


Figure 4: **Generative Precision-Recall curves for VAR-d30** across generation strengths $s \in \{2, 4, 6, 8\}$. Members (green) and Non-Members (red) are traced over 15 iterations, with color intensity indicating iteration progress. Larger s corresponds to more aggressive regeneration.

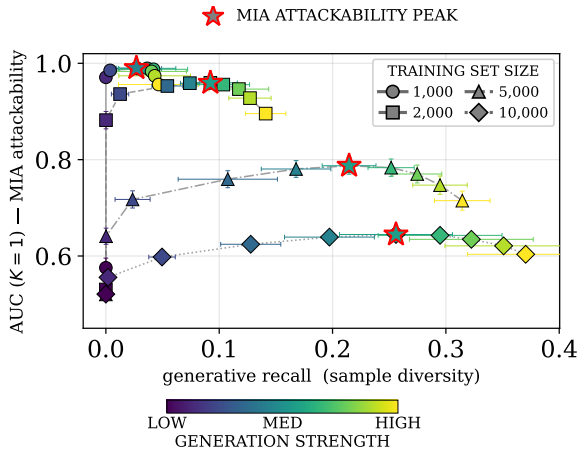


Figure 5: **MIA attackability across generation strength and memorization.** Markers denote (training-set size, denoise_steps) pairs, averaged over five seeds. Each model traces an arc with a single interior maximum.

Table 2: **Generation strength on RAR-XXL** (TPR@1%FPR; best per row in bold). Four of the five attacks peak at the interior s=0.5 on RAR-XXL rather than at the most aggressive setting.

		+ours		
Attack	Baseline	s=0.2	s=0.5	s=0.8
Loss	5.9	9.0	9.0	**9.1**
Hinge	5.7	7.8	**9.1**	8.9
Min-K%++	13.0	16.0	**17.8**	17.4
SURP	4.3	6.7	**8.3**	7.0
CAMIA	3.6	5.2	**7.0**	6.2

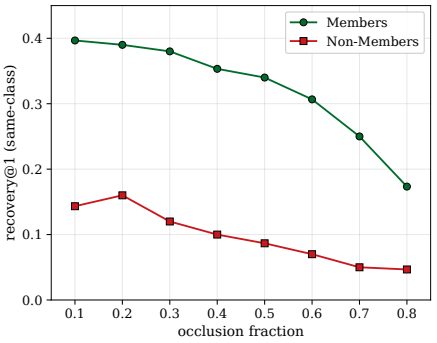


Figure 6: **Recovery of the occluded original**, on DiT-MoE-XL. The final iterate is ranked against the true original and ten same-class distractors.

Qualitatively (Figure 1b), while both populations generate coherent class images under 60% occlusion, member trajectories reconstruct the specific missing objects whereas non-member trajectories land on a different instance of the same class. We quantify this by ranking the final iterate against the true original and 10 same-class distractors, eliminating any category-level advantage. On DiT-MoE-XL (Figure 6), training samples are retrieved at 0.39 and 0.31 under 20% and 60% occlusion, compared to 0.16 and 0.07 for non-members.

4.6 Generation strength controls the drift rate

Figure 4 shows generative precision–recall curves (Kynkäänniemi et al., 2019) for VAR-d30 at $s \in \{2, 4, 6, 8\}$, where s is the number of final scales regenerated. Members separate from non-members at every s. Table 2 sweeps s for the attack itself on RAR-XXL, and Tables 14 to 17 present results for the remaining models.

More generation strength is not always more membership signal. While Figure 4 might suggest that stronger recursion always improves performance by driving populations further apart, this is not the case. Evaluated on actual attacks, performance consistently shows a single interior maximum rather than peaking at either extreme (Figure 5 and Table 2). This tradeoff arises because separating

Table 3: **MIA on IARs** (TPR@1%FPR).

	VAR-d30		VAR-d24		RAR-XXL		RAR-XL	
Attack	Baseline	+ours	Baseline	+ours	Baseline	+ours	Baseline	+ours
Loss	3.5	6.6 (+3.1)	1.8	2.3 (+0.5)	5.9	9.0 (+3.1)	3.2	5.0 (+1.8)
Hinge	2.8	5.0 (+2.2)	1.8	2.0 (+0.2)	5.7	9.1 (+3.4)	3.4	5.0 (+1.6)
Min-K%++	9.5	15.5 (+6.0)	2.7	3.4 (+0.7)	13.0	17.8 (+4.8)	6.0	8.6 (+2.6)
SURP	3.6	6.4 (+2.8)	1.8	2.3 (+0.5)	4.3	8.3 (+4.0)	2.4	3.7 (+1.3)
CAMIA	15.2	14.8 (−0.4)	2.3	3.9 (+1.6)	3.6	7.0 (+3.4)	2.2	3.8 (+1.6)
SOTA	92.0	92.5 (+0.5)	22.8	25.2 (+2.4)	44.4	47.3 (+2.9)	25.5	27.1 (+1.6)

broad populations differs from ranking individual samples. Weak regeneration yields a trajectory too subtle to measure, whereas strong regeneration overwrites the input with a fresh prior sample. Under strong regeneration, populations still separate but individual trajectories lose their sample-specific information. Generation strength is thus an operating point to tune rather than a parameter to maximize. Theory mirrors this shape in Corollary J.5 and Remark J.6, showing that raising s accelerates contraction but lifts the floor $\varepsilon_s/(1-\lambda)$ and narrows the valid contraction region.

4.7 Trajectories recover membership signal where single-forward attacks fail

Anchoring is therefore real, instance-specific, and measurable at a tunable strength (Sections 4.4 to 4.5). To test whether this signal is practically useful, we take the least invasive approach available. We leave existing attacks unmodified and simply concatenate their scores with our trajectory features before applying a calibrated scorer. Any improvement is therefore attributable to information in the trajectory that the base attack does not already extract.

Table 4: **MIA on diffusion models** (TPR@1%FPR).

	DiT-MoE-XL		UViT-T2I-Deep	
Attack	Base	+ours	Base	+ours
SecMI	2.0	17.9 (+15.9)	2.3	3.3 (+1.0)
Den. Loss	1.4	26.7 (+25.3)	2.0	3.3 (+1.3)
PIA	2.9	24.5 (+21.6)	4.3	7.0 (+2.7)
PIAN	11.9	17.3 (+5.4)	0.9	1.7 (+0.8)
CDI	74.8	77.0 (+2.2)	5.9	8.7 (+2.8)

We report TPR at 1% FPR, and include a matched-compute control in which the same query budget is spent on independent perturbations of Z_0 rather than on recursion. We benchmark language models against Loss (Yeom et al., 2018), Zlib (Carlini et al., 2021), Hinge (Bertran et al., 2023), Min-K% (Shi et al., 2023), Min-K%++ (Zhang et al., 2024), ReCaLL (Xie et al., 2024), SURP (Zhang & Wu, 2024), and CAMIA (Chang et al., 2025), alongside their adapted versions for image autoregressive models (Kowalczuk et al., 2025), whose strongest attack per model we report as SOTA. For diffusion models, our benchmark encompasses SecMI (Duan et al., 2023), Denoising Loss (Carlini et al., 2023), PIA (Kong et al., 2023), PIAN, and CDI (Dubiński et al., 2025).

Table 3 reports results on IARs. Trajectory features lift TPR in 23 of the 24 attack–model cells, by a median factor of 1.5× (e.g. Min-K%++ on VAR-d30 from 9.5 to 15.5). The improvement persists even against the strongest available baseline: on RAR-XXL, SOTA improves from 44.4 to 47.3. The effect is largest on diffusion models (Table 4), where several base attacks are close to uninformative and the trajectory supplies most of the signal. On DiT-MoE-XL, all baseline attacks rise significantly, e.g. Denoising Loss TPR@1%FPR rises from 1.4 to 26.7. On the other hand, Table 5 reports TPR on Pythia-6.9B across the seven MIMIR domains. Averaged over domains, every attack improves, and the spread between attacks narrows sharply. All attacks converge to a similar fused value within each domain, as if the trajectory supplied a common signal the base scores capture only partially.

Corresponding AUC values (Tables 10, 11 and 13 in Appendix F) move by points rather than factors, matching expectations for a method acting on the low-FPR tail. Dataset inference follows suit (Appendix G). Using trajectory features reduces the samples required to reach $p \leq 0.01$ in 61 of 86 settings. The savings peak exactly where baselines struggle most, driving several settings from 300–500 samples down to under 150.

Recursion is not simply more queries. The natural objection to any multi-query attack is that the gains come from the query budget rather than from the structure of the queries. The matched-compute control tests this directly: we generate T independent perturbations of Z_0, run the base attack on

Table 5: **MIA on Pythia-6.9B, per MIMIR domain** (TPR@1%FPR).

	arXiv		DM Math		GitHub		HackerNews	
Attack	Baseline	+ours	Baseline	+ours	Baseline	+ours	Baseline	+ours
Loss	3.1	11.0 (+7.9)	30.8	75.9 (+45.1)	31.1	33.4 (+2.3)	2.4	5.2 (+2.8)
Zlib	2.7	10.4 (+7.7)	11.6	74.2 (+62.6)	43.5	38.8 (−4.7)	2.4	4.9 (+2.5)
Hinge	0.6	9.1 (+8.5)	20.3	75.2 (+54.9)	23.4	24.8 (+1.4)	2.2	5.0 (+2.8)
Recall	8.5	15.5 (+7.0)	21.6	76.2 (+54.6)	38.8	33.3 (−5.5)	3.2	6.4 (+3.2)
Min-K%	4.6	15.5 (+10.9)	57.5	74.8 (+17.3)	30.5	32.4 (+1.9)	2.3	5.6 (+3.3)
Min-K%++	2.9	16.3 (+13.4)	29.2	75.4 (+46.2)	22.0	29.9 (+7.9)	2.7	6.6 (+3.9)
SURP	4.3	14.0 (+9.7)	57.9	72.7 (+14.8)	31.2	32.0 (+0.8)	2.4	5.8 (+3.4)
CAMIA	11.3	14.0 (+2.7)	19.0	74.8 (+55.8)	26.0	25.8 (−0.2)	4.7	7.2 (+2.5)

	Pile CC		PubMed		Wikipedia		Average	
Attack	Baseline	+ours	Baseline	+ours	Baseline	+ours	Baseline	+ours
Loss	6.3	11.3 (+5.0)	8.2	16.9 (+8.7)	10.6	13.9 (+3.3)	13.2	23.9 (+10.7)
Zlib	6.1	7.3 (+1.2)	6.1	16.6 (+10.5)	6.6	13.4 (+6.8)	11.3	23.7 (+12.4)
Hinge	4.0	9.9 (+5.9)	3.4	10.2 (+6.8)	8.1	11.4 (+3.3)	8.9	20.8 (+11.9)
Recall	4.6	7.2 (+2.6)	8.5	19.6 (+11.1)	8.8	14.6 (+5.8)	13.4	24.7 (+11.3)
Min-K%	6.4	10.4 (+4.0)	14.9	21.7 (+6.8)	10.7	12.5 (+1.8)	18.1	24.7 (+6.6)
Min-K%++	5.0	7.8 (+2.8)	8.9	16.3 (+7.4)	6.6	8.6 (+2.0)	11.0	23.0 (+12.0)
SURP	7.2	11.5 (+4.3)	14.8	21.2 (+6.4)	10.9	13.3 (+2.4)	18.4	24.3 (+5.9)
CAMIA	4.5	15.1 (+10.6)	5.9	14.3 (+8.4)	8.9	11.7 (+2.8)	11.5	23.3 (+11.8)

Table 6: **Recursion is not a query-budget effect** (TPR@1%FPR, best in bold). The matched-compute *+aug* control spends the same budget on T independent perturbations of Z_0 instead of recursion. **(a)** It degrades accuracy against the strongest attacks per family. **(b)** Even on attacks it benefits, the trajectory still outperforms it. Language results average the seven MIMIR domains.

(a) Strongest attack available.

Family Model	IAR RAR-XXL	Diffusion DiT-MoE-XL	Language Pythia-6.9B
Attack	SOTA	CDI	SURP
Baseline	44.4	74.8	18.4
+aug	43.6 (−0.8)	72.3 (−2.5)	10.2 (−8.2)
+ours	**47.3** (+2.9)	**77.0** (+2.2)	**24.3** (+5.9)

(b) Attack that gains most from *+aug*.

Family Model	IAR RAR-XXL	Diffusion DiT-MoE-XL	Language Pythia-6.9B
Attack	Min-K%++	SecMI	CAMIA
Baseline	13.0	2.0	11.5
+aug	16.4 (+3.4)	2.8 (+0.8)	15.5 (+4.0)
+ours	**17.8** (+4.8)	**17.9** (+15.9)	**23.3** (+11.8)

each, and aggregate before scoring. This does not recover the gains. It usually performs at or below the single-query baseline, and on several attacks it degrades performance substantially (e.g. SURP on Pythia-6.9B falls from 18.4 to 10.2, Table 6). The direction of this result is what the anchoring account predicts. Membership evidence is concentrated at the exact sample: the model's response is sharp at Z_0 and unremarkable after even light augmentation, since the perturbed copies were themselves never trained on. Averaging over a neighborhood therefore integrates the signal away rather than denoising it. Recursion avoids this because the auditor does not choose where the trajectory goes. The model does, and if a memorized point exists it is pulled toward it.

Extended evaluations and ablations. The pattern generalizes broadly. Tables 12 and 13 extend our evaluation to four additional LLMs – including Mamba-2.8B – confirming the behavior holds in non-Transformer state-space models. Additionally, Section H.1 ablates scale by tracking the membership gap across eight vision models, spanning three architectural families and an order of magnitude in size.

5 Conclusions

Recursive self-generation exposes a fundamental asymmetry in how generative models process data. Over multiple iterations, unseen samples drift toward the broader training manifold, while seen samples remain anchored. This trajectory gap is a persistent phenomenon across language, diffusion, and autoregressive vision models. By repurposing this recursive behavior into an inference-time measurement, we extract a continuous membership signal without the cost of shadow models. Fusing trajectory features into existing attacks improves 118 of 122 evaluated configurations: a model's behavior under recursion carries evidence missed by a single query.

AI USAGE STATEMENT

Large language models (LLMs) were used as an assistive tool to improve the readability, copy-editing, and structural flow of this manuscript, including LaTeX formatting and caption drafting. We also utilized LLMs as an interactive aid to refine our hypotheses, integrate attack code for baselines, improve plots clarity, fix mathematical notation errors and discuss results (the full-length manuscript were also checked for possible errors and inconsistencies).

The core intellectual contributions remain entirely our own: generative AI was not used to propose hypotheses, design conceptual framework or process datasets. All empirical results and numerical values reported were generated exclusively by our own experimental code. The authors thoroughly reviewed all AI-assisted work, verified all citations and technical statements, and assume full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our findings, all experiments in this work exclusively utilize open-source models, and publicly available, identically distributed datasets. We conduct all evaluations under strict, deterministic settings – enforcing greedy decoding and standardized sequence lengths – to eliminate confounding variance and isolate genuine membership signals. All experiments ran on a single machine with an NVIDIA A100 GPU and an AMD EPYC 7742 64-core CPU. A complete codebase – containing our modular evaluation framework, data processing pipelines, and exact implementations of our method alongside all baseline attacks will be made publicly available upon the acceptance of this manuscript.

IMPACT STATEMENT

This work advances methods for auditing generative models by improving membership and dataset inference through recursive self-generation (RSG). The primary positive impact is stronger accountability: RSG can help detect memorization of sensitive, proprietary, or benchmark data, supporting privacy audits, copyright verification, and unlearning validation across model families and modalities.

While enhanced inference capabilities can assist in model auditing and transparency, they also require responsible application to avoid potential misuse. We frame RSG as a tool for research evaluation, compliance monitoring, and internal red-teaming. It is important to note that our method provides statistical evidence rather than a definitive proof of data inclusion; therefore, results should be interpreted alongside additional forensic and procedural evidence within a broader data governance framework.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*, 2023.

Ching-Yuan Bai, Hsuan-Tien Lin, Colin Raffel, and Wendy Chi-wen Kan. On training sample memorization: Lessons from benchmarking generative modeling with a large-scale competition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 2534–2542, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467198. URL https://doi.org/10.1145/3447548.3467198.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023.

Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 36:314–330, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022. URL https://arxiv.org/abs/2204.06745.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022.

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)*, pp. 5253–5270, 2023.

Hongyan Chang, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Reza Shokri. Context-aware membership inference attacks against pre-trained large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7299–7321, 2025.

Chen Chen, Daochang Liu, Mubarak Shah, and Chang Xu. Enhancing privacy-utility trade-offs to mitigate memorization in diffusion models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8182–8191, 2025. URL https://api.semanticscholar.org/CorpusID:278129333.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International conference on machine learning*, pp. 1964–1974. PMLR, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pp. 8717–8730. PMLR, 2023.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

Jan Dubiński, Antoni Kowalczuk, Franziska Boenisch, and Adam Dziedzic. Cdi: Copyrighted data identification in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18674–18684, 2025.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Scaling diffusion transformers to 16 billion parameters, 2024. URL https://arxiv.org/abs/2407.11633.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

Abhimanyu Hans, Yuxin Wen, Neel Jain, John Kirchenbauer, Hamid Kazemi, Prajwal Singhania, Siddharth Singh, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, et al. Be like a goldfish, don't memorize! mitigating memorization in generative llms. *Advances in Neural Information Processing Systems*, 37:24022–24045, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. *arXiv preprint arXiv:2305.18355*, 2023.

Antoni Kowalczuk, Jan Dubiński, Franziska Boenisch, and Adam Dziedzic. Privacy attacks on image autoregressive models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=7SXXczJCWP.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Jingwei Li, Jing Dong, Tianxing He, and Jingzhao Zhang. Towards black-box membership inference attack for diffusion models. *CoRR*, abs/2405.20771, 2024. doi: 10.48550/arXiv.2405.20771. URL https://doi.org/10.48550/arXiv.2405.20771.

Pratyush Maini and Anshuman Suri. Reassessing emnlp 2024's best paper: Does divergence-based calibration for mias hold up? In *The Fourth Blogpost Track at ICLR 2025*.

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. *arXiv preprint arXiv:2104.10706*, 2021.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM dataset inference: Did you train on my dataset? *CoRR*, abs/2406.06443, 2024. doi: 10.48550/arXiv.2406.06443. URL https://doi.org/10.48550/arXiv.2406.06443.

Mansi Sakarvadia, Aswathy Ajith, Arham Mushtaq Khan, Nathaniel C Hudson, Caleb Geniesse, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W Mahoney. Mitigating memorization in language models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631:755–759, 07 2024. doi: 10.1038/s41586-024-07566-y.

Aaditya K Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, Gergely Szilvasy, and Dieuwke Hupkes. Evaluation data contamination in llms: how do we measure it and (when) does it matter? *arXiv preprint arXiv:2411.03923*, 2024.

Jiashu Tao and Reza Shokri. (token-level) InfoRMIA: Stronger membership inference and memorization assessment for LLMs. *CoRR*, abs/2510.05582, 2025. doi: 10.48550/arXiv.2510.05582. URL https://doi.org/10.48550/arXiv.2510.05582.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024. URL https://arxiv.org/abs/2404.02905.

Gerrit van den Burg and Chris Williams. On memorization in probabilistic deep generative models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27916–27928. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/eae15aabaa768ae4a5993a8a4f4fa6e4-Paper.pdf.

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=84n3UwkH7b.

YUTONG WU, Han Qiu, Shangwei Guo, Jiwei Li, and Tianwei Zhang. You only query once: An efficient label-only membership inference attack. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7WsivwyHrS.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. Recall: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8671–8689, 2024.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pp. 3093–3106, 2022.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation, 2024. URL https://arxiv.org/abs/2411.00776.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. Detecting data contamination in LLMs via in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YlpaaYxx4t.

Anqi Zhang and Chaofeng Wu. Adaptive pre-training data detection for large language models via surprising tokens. *arXiv preprint arXiv:2407.21248*, 2024.

Jingyang Zhang, Jingwei Sun, Eric C. Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Helen Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. *CoRR*, abs/2404.02936, 2024. doi: 10.48550/arXiv.2404.02936. URL https://doi.org/10.48550/arXiv.2404.02936.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Ziqi Zhang, Chao Yan, and Bradley A Malin. Membership inference attacks against synthetic health data. *Journal of biomedical informatics*, 125:103977, 2022.

A LIMITATIONS

We evaluate three model families; recursion applies wherever a model maps samples to samples of the same type, and extending it to audio, video, and diffusion language models is left to future work. While recursion incurs a linear overhead of T forward passes per sample at inference time, this cost is orders of magnitude lower than the training requirements of shadow model baselines. Furthermore, we leave the optimization of this overhead – such as implementing adaptive early-stopping criteria for highly confident trajectories to future work.

B EXTENDED RELATED WORKS

Our work builds upon three intersecting lines of prior research: the characterization of data memorization in generative models, the evolution of membership inference, and the dynamics of model collapse during recursive generation.

B.1 MEMORIZATION

Memorization – the tendency of generative models to reproduce training examples rather than generate novel samples – has been studied across multiple model families and from both measurement and mitigation perspectives. van den Burg & Williams (2021) formalized the problem for probabilistic generative models such as VAEs, showing that memorization differs fundamentally from mode collapse and overfitting and is not captured by commonly-used nearest-neighbor tests. For diffusion models, Gu et al. (2023) show that the denoising score matching objective has a closed-form optimum that can only replicate training samples, and introduces the EMM metric to quantify how dataset size and model configuration govern the generalization-to-memorization transition. Sakarvadia et al. (2024) localize this phenomenon through bright-ending cross-attention patterns, while the sharpness-based framework of (Wen et al., 2024) justifies score-difference memorization metrics and proposes mitigation via sharpness-aware regularization of the initial noise. The benchmarking study (Bai et al., 2021) demonstrated that standard evaluation metrics fail to surface memorization even in competitive settings. Mitigation has been tackled both for LLMs, where Hans et al. (2024) propose the goldfish loss that excludes randomly sampled token subsets from the training objective to prevent verbatim reproduction, and for text-to-image diffusion models, where Chen et al. (2025) address the privacy–utility tension by combining prompt re-anchoring with semantic prompt search to improve both dimensions simultaneously.

B.2 MEMBERSHIP AND DATASET INFERENCE

A second line of work investigates whether specific examples or datasets can be identified from model behavior. Because individual Membership Inference Attacks (MIAs) can be confounded by distribution shifts (Maini et al., 2024), recent literature often favors Dataset Inference (DI), which aggregates feature evidence across many samples to statistically detect training data usage (Maini & Suri; Dubiński et al., 2025; Kowalczuk et al., 2025). Concurrently, individual MIA methods must adapt to increasingly restrictive black-box deployments. Furthermore, approaches based on training multiple shadow models to learn membership distributions (Ye et al., 2022; Carlini et al., 2021) are now computationally infeasible for massive modern architectures. Consequently, modern attacks must extract signals using only limited outputs rather than internal weights or gradients (Zhang et al., 2024; Chang et al., 2025; Tao & Shokri, 2025).

In these restricted settings, recent black-box attacks heavily rely on output variations. For example, Li et al. (2024) perform MIAs on diffusion models by repeatedly perturbing a target image via an API, averaging the results, and comparing them to the original sample. However, in an interrogation analogy, this approach merely asks multiple paraphrased versions of the exact same question. Because the target sample is perturbed each time independently, the query does not dynamically evolve in response to the model's previous answers, leaving deeper structural memorization unexploited.

B.3 Model Collapse

The last, but very important point is the literature on recursive self-training in generative models. Alemohammad et al. (2023) showed that self-consuming generative loops lead to progressive degradation in quality or diversity when insufficient fresh real data is injected at each generation, a phenomenon they term Model Autophagy Disorder. Their analysis is especially important for our setting because it frames repeated regeneration not as a neutral operation, but as a process that can magnify latent properties of the learned distribution. Closely related, Shumailov et al. (2024) showed that recursively training on model-generated data causes model collapse, where tails of the original distribution disappear and learned behavior drifts toward degenerate approximations. Taken together, these works suggest that iterative generation is structurally revealing: under repeated reuse, memorized or high-density regions may persist differently from non-member examples, while generic outputs may drift or collapse. Our method turns this insight into a privacy-auditing mechanism: rather than studying recursive generation as a training-time pathology, **we exploit recursive self-generation at inference time** to improve the membership-relevant differences.

C Model Details

In our experiments, we consider two vision model families: image autoregressive models (IARs) and diffusion models. The IAR category includes VAR Tian et al. (2024) and RAR Yu et al. (2024) variants, while the diffusion category includes DiT-MoE Fei et al. (2024) and UViT-T2I Bao et al. (2023). Furthermore, we evaluate five large language models (LLMs): four Transformer-based – Pythia (Biderman et al., 2023), GPT-NeoX (Black et al., 2022), GPT-Neo (Black et al., 2021) and GPT-J (Wang & Komatsuzaki, 2021) – and one state-space model (SSM), Mamba (Gu & Dao, 2023). Across all settings, we focus on representative, high-performing model variants.

Table 7: **Vision models used in evaluation.**

	IAR Models						Diffusion Models		
	VAR-d30	VAR-d24	VAR-d20	RAR-XXL	RAR-XL	RAR-L	DiT-MoE-G	DiT-MoE-XL	UViT-T2I-Deep
Model parameters	2.1B	1.0B	600M	1.5B	955M	462M	16.5B	4.1B	141M
Training epochs	350	300	250	400	400	400	—	—	—
FID	1.92	2.33	2.95	1.48	1.50	1.70	1.72	2.10	5.48

Table 8: **Language models used in evaluation.**

	Pythia	Pythia	GPT-NeoX	GPT-Neo	GPT-J	Mamba
Parameters	6.9B	12B	20B	2.7B	6B	2.8B
Training tokens	300B	300B	472B	420B	402B	300B

D Dataset Details

For vision models we use ImageNet Deng et al. (2009) and COCO Veit et al. (2016) datasets that have publicly known and available train/test splits. For language models, we use the established MIA benchmark Mimir Duan et al. (2024) with the *ngram_7_0.2* split.

Table 9: **Member and non-member sources, per model family.**

	VAR	RAR	DiT-MoE	UViT-T2I	Pythia	GPT-NeoX	Mamba	GPT-Neo	GPT-J
Dataset	ImageNet	ImageNet	ImageNet	COCO	Mimir	Mimir	Mimir	Mimir	Mimir

E Metrics Details

The following metrics are computed over the sequence of model outputs collected across RSG iterations, capturing how the model's generative behavior evolves under repeated generation.

E.1 Features for Language Models

As in the vision case, every candidate document is run through a self-consuming, sliding-window trajectory of T iterations, and each raw signal below is recorded once per iteration and reduced over the trajectory via its mean, final-iteration value, or linear slope (noted per feature), except where the signal is already a single whole-trajectory scalar.

Token Match Ratio and Verbatim Recall: At each iteration the model regenerates n tokens from a fixed window; we compare the regenerated continuation $\hat{y}^{(t)}$ against the true continuation y token-by-token. The match ratio is the fraction of positions reproduced exactly, reduced via its final value, its slope across the trajectory, and its leading *depth* — the run of iterations from the start of the trajectory for which the match ratio stays at or above 0.5. Verbatim recall is the length of the longest exact-match streak observed anywhere in the trajectory, tracked at token granularity across iteration boundaries. A soft variant relaxes exact equality to "the true token was among the model's own top-k candidates," giving a soft-verbatim streak that is more forgiving of near-misses. A high, stable match ratio or a long exact-match streak indicates the model can reconstruct the true continuation from a truncated context, a strong sign of memorization.

$$\mathrm{MatchRatio}_t = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[\hat{y}_i^{(t)} = y_i\right],$$

$$\mathrm{Verbatim} = \max_t \ \ell_t, \quad \ell_t = \text{longest exact-match run ending at step } t.$$

Mean Reciprocal Rank (MRR): At each generation step, the true token y_i is assigned a rank r_i under the model's own predicted distribution (rank 1 = the model's top choice). MRR aggregates this rank into one confidence score per iteration; values close to 1 mean the ground-truth continuation was consistently the model's most likely prediction.

$$\mathrm{MRR}_t = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i}, \qquad r_i = |\{v \in \mathcal{V} : \ell_v > \ell_{y_i}\}| + 1$$

where ℓ_v is the logit assigned to vocabulary token v and $\mathcal{V}$ is the model's vocabulary. Reduced via its mean across the trajectory.

Log-Probability: Complementary to the rank-based scores above, the mean log-probability the model assigns to the true continuation measures calibrated confidence rather than ordinal preference.

$$\mathrm{LogProb}_t = \frac{1}{n}\sum_{i=1}^{n}\left(\ell_{y_i} - \log\sum_{v\in\mathcal{V}} e^{\ell_v}\right)$$

Reduced via its mean across the trajectory.

Predictive Entropy and Logit Margin: Beyond agreement with the ground truth, we track how confidently the model commits to its *own*, greedily generated continuation at each step, via the entropy of its output distribution and the gap between its top-2 candidate logits. A model reproducing a memorized sample tends to generate with lower entropy and a wider top-1/top-2 margin than one improvising.

$$H_t = -\sum_{v\in\mathcal{V}} p_v \log p_v, \qquad \mathrm{Margin}_t = \ell_{(1)} - \ell_{(2)}$$

where $p_v = \mathrm{softmax}(\ell)_v$ and $\ell_{(1)} \geq \ell_{(2)}$ are the model's own top two logits at that step. Both reduced via their mean across the trajectory.

Jaccard Similarity: A coarser, string-level similarity between the regenerated continuation and the ground truth, computed over word trigrams and anchored to the true continuation at the current window offset. Unlike the exact match ratio, it remains informative even when regeneration diverges token-by-token but preserves the same local phrasing.

$$\mathrm{Jaccard}_t = \frac{|\mathcal{T}(\hat{y}^{(t)}) \cap \mathcal{T}(y)|}{|\mathcal{T}(\hat{y}^{(t)}) \cup \mathcal{T}(y)|}$$

where $\mathcal{T}(\cdot)$ denotes the set of word trigrams in a text. Reduced via its mean across the trajectory.

Edit Distance: The token-level Levenshtein distance between the regenerated and true continuations — a distance-based sibling of the match-ratio family that additionally penalizes insertions and deletions, not just substitutions.

$$\mathrm{EditDist}_t = \mathrm{Lev}\Big(\hat{y}^{(t)}, y\Big)$$

Reduced via its mean across the trajectory.

Jensen–Shannon Divergence (JSD): Rather than comparing to the ground truth, JSD measures how much the model's own output distribution drifts between *consecutive* regeneration steps, comparing the mean vocabulary-probability profile at iteration t to that at $t-1$ (not to the original prefix, which the sliding window eventually evicts). A stable, low-divergence trajectory suggests the model has settled into reproducing a fixed, memorized continuation rather than continuing to explore alternatives.

$$\mathrm{JSD}(P_t, P_{t-1}) = \tfrac{1}{2}\,\mathrm{KL}(P_t \,\|\, M) + \tfrac{1}{2}\,\mathrm{KL}(P_{t-1} \,\|\, M), \qquad M = \tfrac{1}{2}(P_t + P_{t-1})$$

where P_t is the mean softmax probability vector over the vocabulary at iteration t. Reduced by taking the mean across the trajectory.

E.2 Features for Vision Models

Mean Squared Error (MSE): Measures the average pixel-level reconstruction error between the generated image at a given iteration and the original input. Lower MSE indicates that the model consistently reproduces fine-grained pixel details across iterations, which is a strong signal of memorization.

$$\mathrm{MSE}(x, \hat{x}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2$$

Structural Similarity Index Measure (SSIM) (Wang et al., 2004): Evaluates perceptual similarity between the reconstructed image $\hat{x}$ and the original x by jointly comparing luminance, contrast, and structural information across local image patches. Unlike MSE, SSIM is sensitive to perceptual distortions that are meaningful to human observers, and its stability across iterations serves as a complementary signal to pixel-level metrics.

$$\mathrm{SSIM}(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)}$$

where μ_x, $\mu_{\hat{x}}$ are local means, σ_x^2, $\sigma_{\hat{x}}^2$ are local variances, $\sigma_{x\hat{x}}$ is the cross-covariance, and c_1, c_2 are stabilization constants.

Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018): Quantifies perceptual dissimilarity between x and $\hat{x}$ using deep feature representations extracted from a pretrained network ϕ. By operating in a learned feature space rather than pixel space, LPIPS captures high-level semantic and textural differences that are invisible to MSE or SSIM, making it particularly sensitive to cases where a model reproduces semantic content while varying low-level details.

$$\mathrm{LPIPS}(x, \hat{x}) = \sum_{l}\frac{1}{H_l W_l}\sum_{h,w}\|w_l \odot (\phi_l(x)_{hw} - \phi_l(\hat{x})_{hw})\|_2^2$$

where ϕ_l denotes the feature map at layer l of the pretrained network and w_l are learned channel-wise weights.

F Additional MIA Results

Table 10: **MIA on diffusion models** (AUC).

	DiT-MoE-XL		UViT-T2I-Deep	
Attack	Baseline	+ours	Baseline	+ours
SecMI	65.9	90.1 (+24.2)	68.9	69.6 (+0.7)
Denoising Loss	52.0	91.7 (+39.7)	60.6	63.5 (+2.9)
PIA	63.9	92.4 (+28.5)	70.4	71.4 (+1.0)
PIAN	61.8	90.0 (+28.2)	51.3	54.2 (+2.9)
CDI	98.3	98.5 (+0.2)	74.3	74.7 (+0.4)

Table 11: **MIA on IARs** (AUC).

	VAR-d30		VAR-d24		RAR-XXL		RAR-XL	
Attack	Baseline	+ours	Baseline	+ours	Baseline	+ours	Baseline	+ours
Loss	74.4	80.4 (+6.0)	58.3	60.9 (+2.6)	73.7	76.5 (+2.8)	64.8	67.2 (+2.4)
Hinge	75.3	79.8 (+4.5)	61.9	63.5 (+1.6)	81.2	82.8 (+1.6)	70.7	72.2 (+1.5)
Min-K%++	84.8	86.5 (+1.7)	67.1	67.9 (+0.8)	84.5	86.7 (+2.2)	75.5	77.1 (+1.6)
SURP	74.4	80.3 (+5.9)	58.3	60.9 (+2.6)	76.7	80.0 (+3.3)	66.7	69.6 (+2.9)
CAMIA	79.8	81.5 (+1.7)	64.8	66.0 (+1.2)	66.0	71.2 (+5.2)	58.5	63.2 (+4.7)
SOTA	99.6	99.6 (+0.0)	87.4	87.6 (+0.2)	97.3	97.4 (+0.1)	92.1	92.2 (+0.1)

Table 12: **MIA on LLMs** (TPR@1%FPR, averaged over MIMIR domains).

	Pythia-12B		GPT-J-6B		GPT-Neo-2.7B		Mamba-2.8B	
Attack	Baseline	+ours	Baseline	+ours	Baseline	+ours	Baseline	+ours
Loss	12.9	22.7 (+9.8)	14.7	23.3 (+8.6)	14.6	24.5 (+9.9)	12.4	22.3 (+9.9)
Zlib	10.8	23.1 (+12.3)	10.5	22.8 (+12.3)	9.6	24.0 (+14.4)	11.1	22.6 (+11.5)
Hinge	8.7	19.6 (+10.9)	8.2	20.2 (+12.0)	7.9	21.3 (+13.4)	10.0	20.0 (+10.0)
Recall	12.2	23.8 (+11.6)	10.6	22.5 (+11.9)	10.5	23.8 (+13.3)	10.8	21.8 (+11.0)
Min-K%	14.8	24.3 (+9.5)	15.6	23.5 (+7.9)	16.0	24.9 (+8.9)	13.6	22.5 (+8.9)
Min-K%++	11.5	22.2 (+10.7)	11.5	22.2 (+10.7)	8.6	22.3 (+13.7)	9.8	21.6 (+11.8)
SURP	14.6	24.0 (+9.4)	18.8	23.7 (+4.9)	16.3	24.7 (+8.4)	13.8	22.9 (+9.1)
CAMIA	11.2	21.6 (+10.4)	10.5	21.0 (+10.5)	10.7	21.8 (+11.1)	12.3	21.8 (+9.5)

Table 13: **MIA on LLMs** (AUC, averaged over MIMIR domains).

	Pythia-12B		GPT-J-6B		GPT-Neo-2.7B		Mamba-2.8B	
Attack	Baseline	+ours	Baseline	+ours	Baseline	+ours	Baseline	+ours
Loss	71.7	68.1 (−3.6)	71.8	70.1 (−1.7)	71.1	69.8 (−1.3)	71.2	68.6 (−2.6)
Zlib	70.5	68.6 (−1.9)	70.4	70.2 (−0.2)	69.7	70.0 (+0.3)	70.1	68.9 (−1.2)
Hinge	63.3	65.5 (+2.2)	64.0	67.0 (+3.0)	62.9	67.0 (+4.1)	63.4	66.3 (+2.9)
Recall	67.6	68.9 (+1.3)	67.0	69.0 (+2.0)	66.0	69.1 (+3.1)	67.2	67.7 (+0.5)
Min-K%	72.3	69.7 (−2.6)	72.2	71.8 (−0.4)	71.3	71.2 (−0.1)	71.6	69.6 (−2.0)
Min-K%++	66.3	66.7 (+0.4)	66.7	68.8 (+2.1)	65.1	68.0 (+2.9)	64.8	66.5 (+1.7)
SURP	72.4	70.2 (−2.2)	72.2	71.9 (−0.3)	71.3	71.3 (+0.0)	71.6	70.2 (−1.4)
CAMIA	67.4	68.0 (+0.6)	67.3	69.6 (+2.3)	66.7	69.0 (+2.3)	66.5	68.5 (+2.0)

G Dataset Inference Results

DI pipeline. The DI pipeline is identical to MIAs, replacing the target label with dataset-origin variable D. The same Z_t, ϕ_t, and trajectory-fusion machinery is used; only label semantics and calibration change. For DI, splitting/evaluation are performed at the dataset or source-group level, and per-sample logits are aggregated by a fixed mean rule into a dataset-level score. Both MIA and DI setups inherit standard generative privacy-audit conventions (Kowalczuk et al., 2025; Dubiński et al., 2025; Zhang et al., 2024).

Results. Dataset inference asks the same question at the level of a corpus rather than a sample, and the trajectory features stay unchanged. Only the label semantics and the aggregation rule differ

(Section 3.4). The relevant measure is sample efficiency – how many samples are needed before the test reaches significance. Figure 7 compares the minimum n required to reach $p \leq 0.01$ with and without trajectory features, over 86 attack–model settings. Trajectory features help in 61 of them, leave 23 unchanged and cost slightly in 2, and the advantage is largest precisely where the baseline is most expensive, with several settings requiring 300 – 500 samples at baseline dropping to under 150.

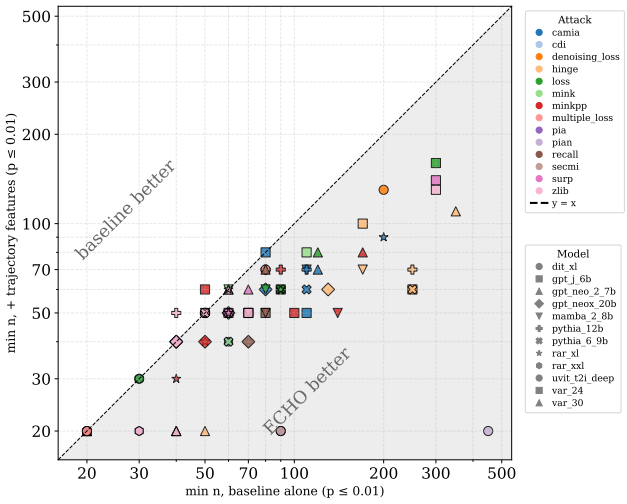


Figure 7: **Dataset inference reaches significance from fewer samples.** Each point is one attack on one model: the smallest n at which the test attains $p \leq 0.01$, for the base attack alone against the same attack fused with trajectory features. Points below the diagonal are settings where the trajectory helps; 61 of 86 fall there, 23 sit on it, and 2 above. Color is the attack, shape describes the model.

H Precision and Recall for Generative Models

Figure 8 shows Generative Precision and Recall across iterations for VAR-d30 and DiT-MoE-XL. In both models and both metrics, members consistently score higher than non-members throughout all iterations, confirming that the membership signal is stable and model-agnostic. Notably, the gap between members and non-members widens as iterations progress, indicating that recursive self-generation amplifies the underlying asymmetry rather than merely preserving it.

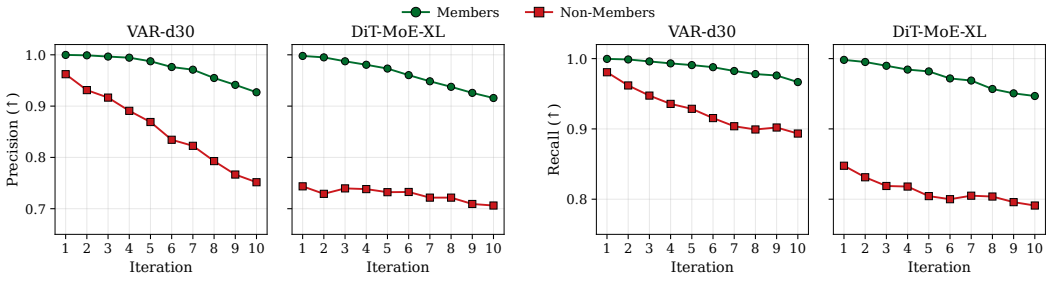


Figure 8: **Generative Precision and Recall across iterations.** Members stay above non-members in both models and both metrics. For VAR-d30 the gap widens significantly with iteration.

H.1 The asymmetry persists across model scale

As illustrated in Figure 9, the membership separation $\Delta\text{FID} = \text{FID}_{\text{nonmem}} - \text{FID}_{\text{mem}}$ stays positive at every iteration for every model we test, so in all of them the gap between members and non-members keeps accumulating along the trajectory. How it depends on scale is family-specific. In VAR the three sizes separate cleanly and the gap grows with the model, sitting near 0.6 for d20, 1.4 for d24 and 3.4 for d30, and it is flat in t. In RAR the three sizes track each other closely and all decay over the

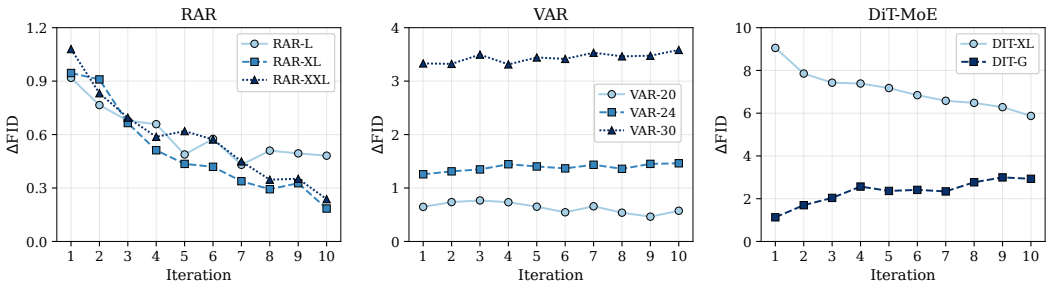
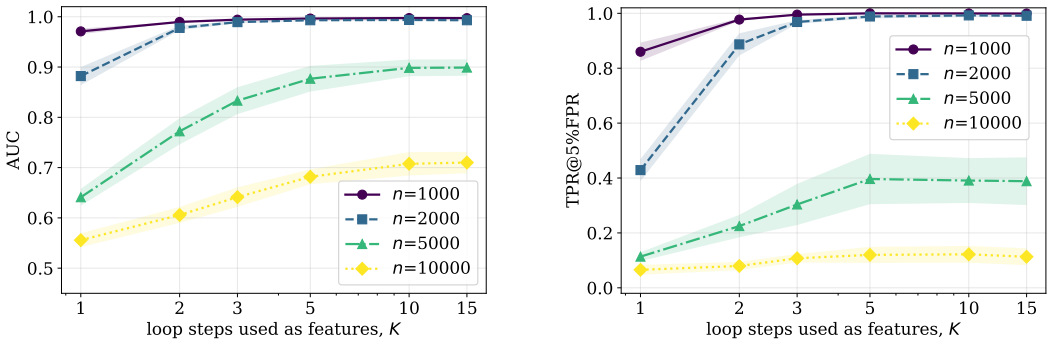


Figure 9: **Ablation: trajectory asymmetry across model scale.** Membership separation $\Delta\text{FID} = \text{FID}_{\text{nonmem}} - \text{FID}_{\text{mem}}$ over the first ten iterations, for every size in each family. Panels carry independent y axes, since the separation differs by an order of magnitude between families; darker curves are larger models.

trajectory, so scale makes little difference. In DiT-MoE the two sizes differ in shape rather than in height alone – the gap decays for DiT-MoE-XL and grows for DiT-MoE-G – which is too little to read a trend from. What is robust is that the gap is present and accumulates in every family at every size, not the direction in which it scales.

I Trajectory Classifier on Controlled DDPMs

On the controlled DDPMs of Section 4.3 we also read the trajectory directly as a membership signal. A 5-fold cross-validated logistic regression is fit to the K-dimensional vector of origin distances $\left[\|x(k) - x(0)\|^2/D\right]_{k=1}^{K}$, with K interpolating between the classical single-query attack at $K = 1$ and a reader of the whole trajectory prefix at $K = 15$. Increasing K never hurts AUC, since the classifier can ignore later coordinates, but at the boundary of memorization it changes the operating-point metric sharply: for the $n = 2000$ model TPR@5%FPR climbs from 0.43 to 0.99, while AUC is already above 0.9 at $K = 1$ and merely saturates. An averaged metric therefore hides that a single-query attacker would still miss more than half the members at a 5% false-positive budget. The signal is in the shape of the trajectory, not in any single reconstruction error.



(a) Trajectory-classifier AUC as a function of K, the number of trajectory steps used as classifier features. $K = 1$ recovers the classical single-query MIA; $K = 15$ uses the full trajectory prefix.

(b) TPR@5%FPR at the same points as a. Where AUC already saturates at $K = 1$, the low-FPR regime shows trajectory features lifting the TPR (e.g., training-set size $= 2000$: $0.43 \to 0.99$).

Figure 10: **Cumulative trajectory-classifier MIA quality at** denoise_steps $= 2$**.** A 5-fold cross-validated logistic regression is trained on the K-dimensional vector of origin-distance features $\left[\|x(k) - x(0)\|^2/D\right]_{k=1}^{K}$ per sample, where K is the number of loop steps used as features. For weakly memorized models the loop-classifier recovers signal that a single-query attack ($K = 1$) misses, particularly at strict operating points. Bands are ± 1 SD across five seeds.

Table 14: **Generation strength on VAR-d30** (TPR@1%FPR; best per row in bold).

		+ours		
Attack	Baseline	s=2	s=5	s=8
Loss	3.5	6.7	6.6	**6.9**
Hinge	2.8	3.8	**5.0**	3.9
Min-K%++	9.5	14.7	**15.5**	15.4
SURP	3.6	6.6	6.4	**6.8**
CAMIA	15.2	**17.2**	14.8	15.0

Table 15: **Generation strength on VAR-d24** (TPR@1%FPR; best per row in bold).

		+ours		
Attack	Baseline	s=2	s=5	s=8
Loss	1.8	2.2	**2.3**	**2.3**
Hinge	1.8	1.8	**2.0**	1.9
Min-K%++	2.7	3.2	**3.4**	3.3
SURP	1.8	2.2	**2.3**	**2.3**
CAMIA	2.3	3.4	3.9	**4.2**

Table 16: **Generation strength on Pythia-6.9B** (decoding temperature t; TPR@1%FPR, MIMIR average; best per row in bold).

		+ours		
Attack	Baseline	t=0.0	t=0.5	t=1.0
Loss	13.2	**23.9**	22.6	20.6
Hinge	8.9	**20.8**	19.9	17.0
Min-K%++	11.0	**23.0**	22.2	18.8
SURP	18.4	**24.3**	24.1	23.6
CAMIA	11.5	**23.3**	21.6	19.3

Table 17: **Generation strength on Pythia-12B** (decoding temperature t; TPR@1%FPR, MIMIR average; best per row in bold).

		+ours		
Attack	Baseline	t=0.0	t=0.5	t=1.0
Loss	12.9	22.7	**23.5**	20.9
Hinge	8.7	**19.6**	18.0	18.0
Min-K%++	11.5	**22.2**	20.3	20.6
SURP	14.6	24.0	**24.9**	23.9
CAMIA	11.2	**21.6**	20.3	20.1

J THEORY

We define memorization as local sharpness of the model's density, and prove that under this definition the trajectory's contraction rate is an estimator of memorization strength.

Two consequences organize what follows. First, a trajectory is characterized by two numbers, a contraction rate λ and a floor $\varepsilon_\sigma/(1-\lambda)$ that it cannot settle below; every trajectory feature we compute is an estimate of one of them, or of a function of both. Second, the same analysis says why spending the query budget on independent perturbations of Z_0 instead fails: averaging a base score over a neighborhood introduces a bias that grows with the local curvature, and so penalizes exactly the samples that are most memorized.

We analyze recursive self-generation as a dynamical system on the sample space. The central claim is that memorization, understood as local sharpness of the model's own density, makes a training point an attractor of the model's generation map, with a contraction rate that factorizes into a parameter we control and the quantity we wish to measure.

J.1 SETUP AND ASSUMPTIONS

One step of recursive self-generation at strength σ is

$$\mathcal{G}_\sigma(z) \;=\; D_\theta(z+\sigma\xi), \qquad \xi\sim\mathcal{N}(0,I),$$

where D_θ is the model's denoiser and $z\in\mathbb{R}^d$. Let p_θ denote the distribution represented by the trained model and $p_\theta^{(\sigma)} := p_\theta * \mathcal{N}(0,\sigma^2 I)$ its smoothing at scale σ. We write $s(y) := \nabla \log p_\theta^{(\sigma)}(y)$ for the score and $B(x,r) := \{z : \|z-x\| \le r\}$.

Assumption J.1 (Denoiser)**.** D_θ approximates the optimal denoiser for p_θ at level σ, so that by Tweedie's identity $\|D_\theta(y) - (y+\sigma^2 s(y))\| \le \delta_\theta$ for all y.

Assumption J.2 (Local memorization)**.** A point $x^\star$ is *memorized at scale σ with strength $m>0$* if there exist $r_0>0$ and $M\ge m$ such that

$$mI \;\preceq\; -\nabla^2 \log p_\theta^{(\sigma)}(z) \;\preceq\; MI \qquad \forall z\in B(x^\star,r_0), \qquad \text{and} \qquad s(x^\star)=0.$$

A point is *unmemorized* if no such $m>0$ and $r_0>0$ exist.

Assumption J.3 (Regularity)**.** $\log p_\theta^{(\sigma)}$ is three times differentiable on $B(x^\star,r_0)$ with $\|\nabla^3 \log p_\theta^{(\sigma)}\| \le L$.

Assumption J.2 is the substantive one, and we do not claim it follows from training; that a model memorizes some samples and not others is the empirical premise of this work. What we claim is that it is *measurable*: the contraction rate it implies is estimable from observed trajectories.

J.2 ONE-STEP CONTRACTION

Proposition J.4 (Contraction at memorized points)**.** *Let $x^\star$ satisfy Assumption J.2 with constants m, M, r_0, and let Assumptions J.1 and J.3 hold. Suppose*

$$\sigma^2(M+m) \le 2$$

and let z satisfy $\|z-x^\star\| \le r$ with $r + c\sigma\sqrt{d} \le r_0$. Then

$$\mathbb{E}\big\|\mathcal{G}_\sigma(z) - x^\star\big\| \;\le\; \lambda\,\|z-x^\star\| \;+\; \varepsilon_\sigma, \qquad \lambda \;=\; 1-\sigma^2 m,$$

with

$$\varepsilon_\sigma \;=\; \lambda\,\sigma\sqrt{d} \;+\; \tfrac{1}{2}\sigma^2 L\big(r^2+\sigma^2 d\big) \;+\; \delta_\theta,$$

up to an event of probability $e^{-\Omega(d)}$ on which $z+\sigma\xi$ leaves $B(x^\star,r_0)$.

Proof of Proposition J.4. Write $u := z-x^\star$, $\eta := \sigma\xi$ and $y := x^\star + u + \eta$. By Gaussian concentration $\|\eta\| \le c\sigma\sqrt{d}$ with probability $1-e^{-\Omega(d)}$; on this event $y\in B(x^\star,r_0)$ and the expansions below are valid.

Since $s(x^\star) = 0$, Taylor's theorem with $H := \nabla^2 \log p_\theta^{(\sigma)}(x^\star)$ and Assumption J.3 give

$$s(y) \;=\; H(u+\eta) + R, \qquad \|R\| \le \tfrac{L}{2}\|u+\eta\|^2.$$

By Tweedie's formula the optimal denoiser is $D^\star(y) = y + \sigma^2 s(y)$, so by Assumption J.1,

$$\mathcal{G}_\sigma(z) - x^\star = \big(I + \sigma^2 H\big)(u+\eta) + \sigma^2 R + \Delta_\theta, \qquad \big(\mathbb{E}\|\Delta_\theta\|^2\big)^{1/2} \le \delta_\theta.$$

Linear term. Assumption J.2 gives $-MI \preceq H \preceq -mI$, so the eigenvalues of $A := I + \sigma^2 H$ lie in $[1-\sigma^2 M,\, 1-\sigma^2 m]$ and

$$\|A\|_{\mathrm{op}} \le \max\{|1-\sigma^2 m|,\, |1-\sigma^2 M|\} = 1 - \sigma^2 m = \lambda,$$

where both equalities use the hypothesis $\sigma^2(M+m) \le 2$: it is equivalent to $\sigma^2 M - 1 \le 1 - \sigma^2 m$, and, since $m \le M$, it also gives $\sigma^2 m \le 1$, so $1 - \sigma^2 m \ge 0$ and $\lambda \in [0, 1)$. Hence, by the triangle inequality and $\mathbb{E}\|\eta\| \le \sigma\sqrt{d}$,

$$\mathbb{E}\|A(u+\eta)\| \;\le\; \|Au\| + \|A\|_{\mathrm{op}}\,\mathbb{E}\|\eta\| \;\le\; \lambda\|u\| + \lambda\sigma\sqrt{d}.$$

Remainder. $\mathbb{E}\|u+\eta\|^2 = \|u\|^2 + \sigma^2 d \le r^2 + \sigma^2 d$, so $\sigma^2 \mathbb{E}\|R\| \le \frac{1}{2}\sigma^2 L(r^2 + \sigma^2 d)$. By Jensen, $\mathbb{E}\|\Delta_\theta\| \le \delta_\theta$. Combining the three bounds gives the claim. □

Corollary J.5 (Drift floor)**.** *Under the hypotheses of Proposition J.4, if the iterates remain in $B(x^\star, r)$ then*

$$\mathbb{E}\big\|Z_t - x^\star\big\| \;\le\; \lambda^t \,\|Z_0 - x^\star\| \,+\, \varepsilon_\sigma \frac{1-\lambda^t}{1-\lambda}, \qquad \textit{so} \qquad \limsup_{t\to\infty} \mathbb{E}\big\|Z_t - x^\star\big\| \;\le\; \frac{\varepsilon_\sigma}{1-\lambda} \;=\; \frac{\varepsilon_\sigma}{\sigma^2 m}.$$

This limit is the drift floor δ_∞ *of Section 3.3: the trajectory contracts toward $x^\star$ at rate λ but cannot settle closer than the noise and denoiser error allow.*

Proof of Corollary J.5. Iterate Proposition J.4 and sum the geometric series. □

Remark J.6 (Two regimes in σ)**.** The rate $\lambda = 1 - \sigma^2 m$ improves with σ, but two constraints bound it from above: the step-size condition $\sigma^2 \le 2/(M+m)$ of Proposition J.4, beyond which the denoising step overshoots in the sharpest direction and the bound λ no longer holds, and the requirement $r + c\sigma\sqrt{d} \le r_0$, which shrinks the admissible starting region as σ grows. The first is tight: at $\sigma^2 = 2/(M+m)$ the rate bottoms out at $\lambda = (M-m)/(M+m)$, the condition number of the local curvature. Recursion therefore has an interior optimum in generation strength, which we observe empirically for image models in Section 4.4.

Remark J.7 (Unmemorized points)**.** For a point admitting no $m > 0$ in Assumption J.2, the same expansion gives $\|I + \sigma^2 H\|_{\mathrm{op}} \ge 1$ and no contraction occurs. The iteration is then governed by s elsewhere in the space and converges to regions where $p_\theta^{(\sigma)}$ has mass, independently of the starting point. This is the drift we observe for non-members in Section 4.4.