Proceedings Track

Tracking Memorization Geometry throughout the Diffusion Model Generative Process

Abstract

Memorization in generative text-to-image diffusion models is a phenomenon where instead of valid image generations, the model outputs near-verbatim reproductions of training images. This poses privacy and copyright risks, and remains difficult to prevent without harming prompt fidelity. We present a mid-generation, geometry-informed criterion that detects, and then helps avoid (mitigate), memorized outputs. Our method analyzes the natural image distribution manifold as learnt by the diffusion model. We analyze a memorization criterion that has a local curvature interpretation. Thus we can track the generative process, and our criterion's trajectory throughout it, to understand typical geometrical structures traversed throughout this process. This is harnessed as a geometry-aware indicator that distinguishes memorized from valid generations. Notably, our criterion uses only the direction of the normalized score field, unlike prior magnitude-based methods; combining direction and magnitude we improve mid-generation detection SOTA by $\sim 5\%$. Beyond detecting memorization, we use this indicator as a plug-in to a mitigation policy to steer trajectories away from memorized basins while preserving alignment to the text. Empirically, this demonstrates improved fidelity–memorization trade-off over the competitors. By linking memorization to magnitude-invariant geometric signatures of the generative process, our work opens a new direction for understanding—and systematically mitigating—failure modes in diffusion models. Official code: bit.ly/4ndeISd

1. Introduction

Understanding the geometry of natural-image distributions has been studied for almost a century Simoncelli (2024), where its roots lie in power-spectrum/statistical regularities and multiscale structure Field (1987); Ruderman and Bialek (1994); van der Schaaf and van Hateren (1996); Huang and Mumford (1999); Portilla and Simoncelli (2000); Simoncelli and Olshausen (2001); Zoran and Weiss (2012); Wainwright and Simoncelli (2000). In today's age of generative AI, diffusion models enable sampling from this probability with unprecedented quality. They do so by approximating the probability's score field, and simulating a multi-step reverse diffusion *generative process* Hyvärinen (2005); Vincent (2011); Song and Ermon (2019); Ho et al. (2020); Song et al. (2021); Dhariwal and Nichol (2021); Rombach et al. (2022). In accordance, a growing body of work develops theory and analyses of this approximated manifold Kadkhodaie et al. (2024); Bortoli et al. (2022); Benton et al. (2024); Sakamoto and Suzuki (2024); Tang et al. (2024a); Potaptchik et al. (2024). Both high-quality sampling as well as score-function analysis of the learned natural image manifold are prominent features of *memorization* research.

Alongside the success of diffusion models came the prominent risk of *memorization*: near-verbatim reproduction of training images, which pose copyright and privacy risks Carlini et al. (2023); Webster (2023); Hu and Pang (2023); Tang et al. (2024b); Zhai et al. (2024); Pang et al. (2025). Recently, approaches analyzing the manifold geometry around

points of memorization have shown significant potential in understanding and mitigating this phenomenon Wen et al. (2024); Jeon et al. (2025); Brokman et al. (2025). The former two emphasize the importance of geometry of memorization from *mid-generation* signals, so that it can be detected to steer away generation from memorized outcomes while sampling. Both offer methods that incorporate *magnitude* of score-functions.

This work advances the geometry-aware research of these phenomena. Our main contributions are as follows:

- A novel **magnitude-invariant mid-generation detection criterion** based on the direction (not the size) of classifier-free-guidance (CFG) guiding score field. This criterion, κ^{Δ}, approximates a high-dimensional generalization of the surface *mean curvature* - thus providing direct geometric interpretability.
- **Shifting-geometry observations.** We track the values of κ^{Δ} as they change throughout the image generation process. Memorized prompts show increasingly convex/attractive geometry, with early positive κ^{Δ} that strengthens throughgout the process, whereas non-memorized prompts start with $\kappa^{\Delta} < 0$, pass smoothly through $\kappa \approx 0$, and often culminate with a weak $\kappa^{\Delta} > 0$ (Figs. 1, 2). This clearly distinct.
- **Improving the baseline.** Since our criterion holds information about the direction of $\nabla \log p(x)$, it holds complementary information to the previous magnitude-based approaches. Thus, it is natural to combine it with such methods - raising AUC at *earliest* generation steps from 0.925 to 0.97. Additionally - we demonstrate how plugging our criterion, as a stand-alone, into a mitigation strategy improves the fidelity–memorization balance over the competitors.

2. Related Work

Several recent works probe memorization through local differentials of the score field. Wen et al. Wen et al. (2024) use the *norm* of the predicted score (and the conditional–unconditional gap under CFG) as a prompt-level detector and then mitigate by *norm minimization* via soft–prompt optimization. Jeon et al. Jeon et al. (2025) extend this magnitude-based view with a Hessian–score product (sharpness) criterion, deriving early-time tests and proposing latent–noise optimization for mitigation. Closest to our perspective, p-Laplace–based detectors apply higher-order differentials of the *final* score field Brokman et al. (2025). These are post-hoc (image-based) approaches and thus not suited for *mid-generation* steering, or mitigation in general - since these interfere during and not after the generative process.

Other approaches intervene without modeling the geometry of the probability manifold: cross-attention analyses reveal token-level triggers and enable editing-time mitigation Ren et al. (2024); neuron-level methods identify and suppress memorizing units Hintersdorf et al. (2024).

In contrast to magnitude-dependent norms or post-hoc image tests, our geometry-aware criterion is both *magnitude-invariant* and *trajectory-aware*: we normalize the conditional-unconditional gap and track and approximate a well-known high-dimensional generalization of the surface mean curvature along the generative path. hence - tracking its signed value along the generative process has direct geometrical interpretations.

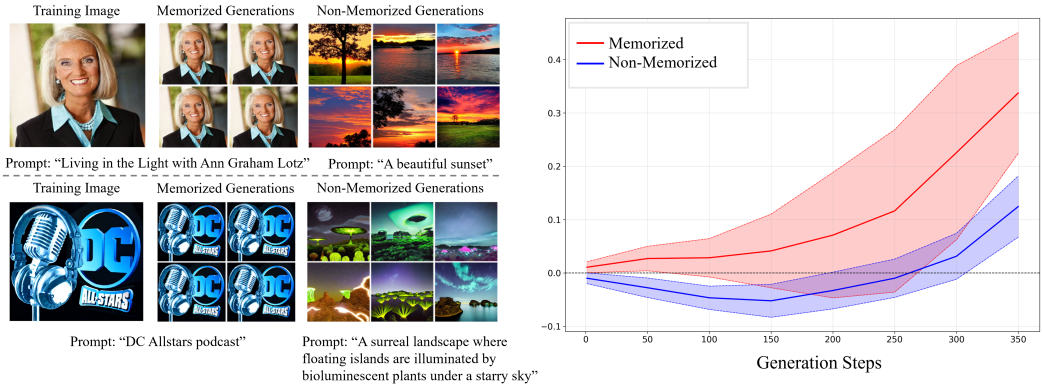

Figure 1: **Memorized vs non-memorized prompts, and the observed shifting geometry signature.** *Left:* Two prompts that trigger memorization: for each we show the associated training image and four near-verbatim generations produced by the memorized prompt. *Middle:* Two non-memorized prompts with six valid generations each (diverse, non-matching). *Right:* Temporal evolution of the curvature signal κ^{Δ} over generation steps; solid lines are the mean across generations from 2×250 prompts (250 memorized, 250 non-memorized), shaded bands denote ± 1 s.t.d. Memorized prompts (red) enter convex/attractive geometry, with early $\kappa^{\Delta} > 0$ that strengthen, whereas non-memorized prompts (blue) start in saddle-like/repelling geometry ($\kappa^{\Delta} < 0$), pass near-flat regions, and only later become weakly positive. This geometry underpins our mid-generation detector and steering policy.

3. Method

3.1. Diffusion model settings

Let $X \subset \mathbb{R}^d$ be data drawn from an unknown distribution with density p. Diffusion models learn the *score function* $s(x) := \nabla_x \log p(x)$ at noise level t, typically via noise prediction Song and Ermon (2019); Nichol and Dhariwal (2021); Miyasawa et al. (1961): Given a noising process $q(x_t \,|\, x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1 - \bar{\alpha}_t) I\big)$, the MMSE noise predictor aligns with $-s_t(x_t)$. A neural network trained to predict noise thus estimates the score over time steps.

Forward and reverse processes. The forward process progressively corrupts x_0 to noise; the reverse process denoises from $x_T \sim \mathcal{N}(0, I)$ using the learned scores to reconstruct x_0 Ho et al. (2020); Song et al. (2021). In text-conditional settings, classifier-free guidance (CFG) provides both unconditional and conditional scores, $s_\theta(x_t)$ and $s_\theta(x_t, c)$, respectively Ho and Salimans (2022).

SDE view and latent diffusion. The reverse dynamics can be written as an SDE or ODE driven by the score Song et al. (2021). In latent diffusion Rombach et al. (2022), the process runs in a learned latent z_t, with decoding $x_0 = D(z_0)$. All quantities below apply identically in latent space by replacing x with z.

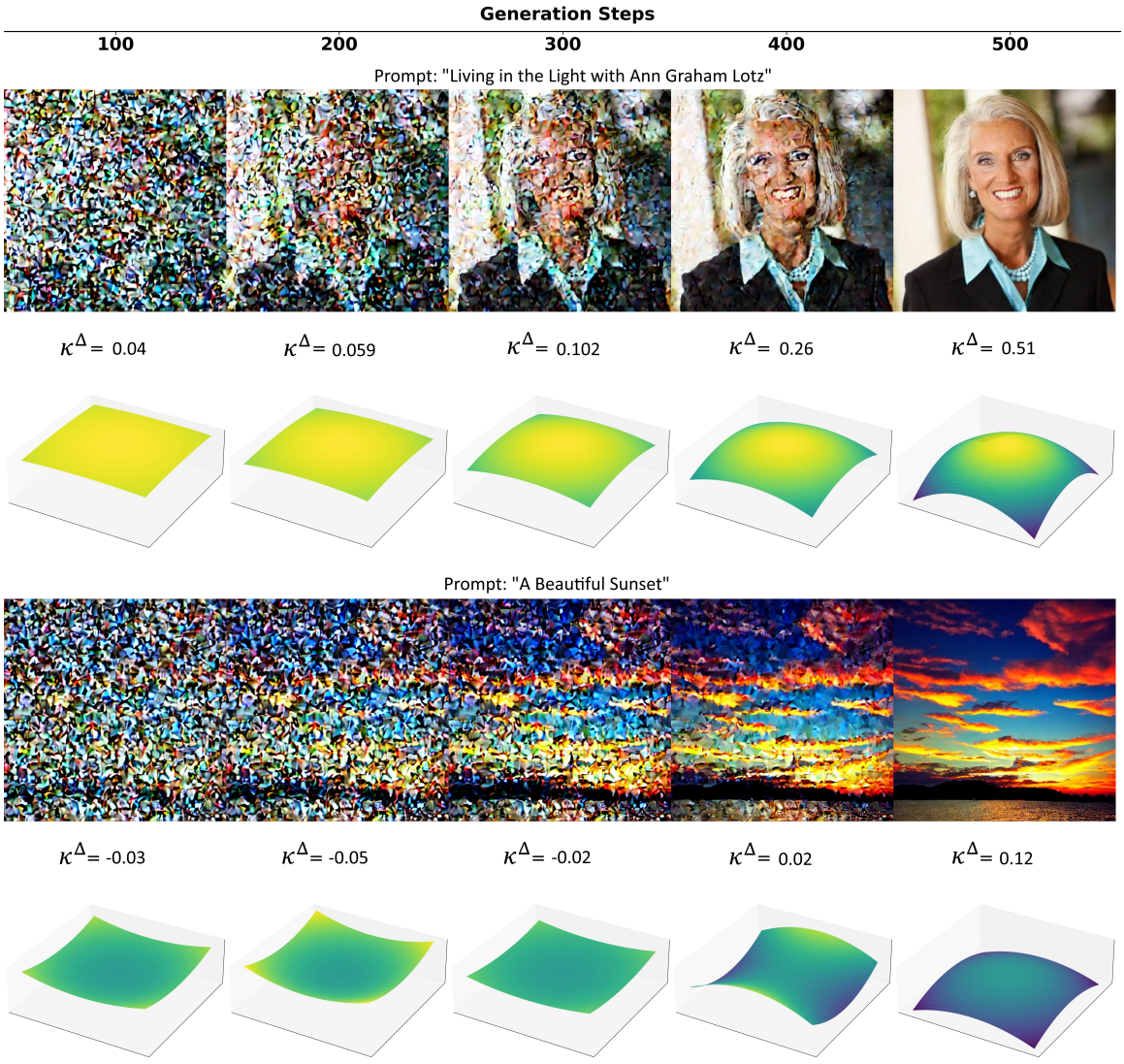

Figure 2: **Demonstrating the generative process' mid-generation geometry-shift.** Each row shows a 500-step generative process (snapshots every 100 steps) with our criterion κ^{Δ} computed per snapshot, and a surface that reflects the exact same mean curvature: Our criterion is based on a known high-dimensional generalization of surface mean curvature - enabling interpretability of its values. 4

3.2. Method: Normalized boundary flux via classifier-free guidance

We study the *prompt-induced* change in the score and its local geometry.

CFG score difference and normalization. Define the CFG difference, or the CFG guiding score field as

$$s^{\Delta}(x_t) \;:=\; s_\theta(x_t, c) \;-\; s_\theta(x_t),$$

and its unit regularized field $\hat{s}^{\Delta}(x_t) := \frac{s^{\Delta}(x_t)}{\|s^{\Delta}(x_t)\|+\delta}$, with $\delta > 0$ a regularizer that prevents division by zero. In standard CFG sampling, the effective guidance is $s_\theta(x_t) + \omega\, s^{\Delta}(x_t)$ for a guidance scale ω; here we isolate the *directional* effect via $\hat{s}^{\Delta}$.

Normalized boundary flux (mean-curvature analogue). For a small ball $B_R(x_0) \subset \mathbb{R}^d$ and its boundary $\partial B_R(x_0)$ with outward unit normal n, we define the *normalized boundary flux divergence*

$$\kappa^{\Delta}(x_0) \;:=\; -\frac{1}{|B_R(x_0)|}\int_{\partial B_R(x_0)} \hat{s}^{\Delta}(x)\cdot n\, ds.$$

As $R \to 0$, $\kappa^{\Delta}(x_0) = \nabla\cdot\hat{s}^{\Delta}(x_0)$. This mirrors the classical level-set expression of mean curvature of a surface (2D embedded in 3D) as the divergence of a unit normal field, $\kappa = \nabla\cdot\frac{\nabla\phi}{\|\nabla\phi\|+1}$ Sochen et al. (1998); thus, since the image probability manifold is high-dimensional (and not a mere surface), κ^{Δ} serves as a curvature-like measure of how prompt conditioning bends the learned log-density geometry along $\hat{s}^{\Delta}$.

Algorithm: Numeric calculation of our criterion. Sample N points y_i uniformly on $\partial B_R(x_0)$, set $n_i = (y_i - x_0)/R$. Using $|\partial B_R|/|B_R| = d/R$, a Monte Carlo boundary estimator is

$$\widehat{\kappa}^{\Delta}(x_0) \;\approx\; \frac{d}{RN}\sum_{i=1}^{N} \hat{s}^{\Delta}(y_i)\cdot n_i.$$

In practice, we take small R, evaluate in the native sampling domain (pixel or latent), and (optionally) track $\kappa^{\Delta}(x_t)$ along the generation trajectory to profile prompt-induced curvature over time. This estimator requires only $\hat{s}^{\Delta}(\cdot)$ evaluations (no volume integrals), aligns with CFG, and is directly compatible with standard diffusion pipelines.

Interpretable and practical surface curvature aspect. For interpretability, we discuss the observed dynamics as if we are traversing a *2D surface in 3D*. Then $\kappa^{\Delta} = \nabla\cdot\hat{s}^{\Delta}$ acts as a signed mean-curvature surrogate, and we can analyze whether it aligns with our observed high-dimensional behavior, as in Fig. 2. Let the principal curvatures be κ_1, κ_2 so

$$M := \tfrac{1}{2}(\kappa_1 + \kappa_2), \qquad G := \kappa_1\kappa_2.$$

Moreover, note that κ^{Δ} gives *mean-curvature* information (M), while *Gaussian curvature* (G) is not directly available; nevertheless - empirical observations of the generative process bridge some of these gaps.

Memorized. Typical behavior:

$\kappa^{\Delta}(t) > 0\ \forall t,\ \exists\, t_{\text{dip}}:\ \kappa^{\Delta}(t_{\text{dip}}) < \kappa^{\Delta}(t_0)$, Empirical observation: narrow stable set collapse.

The empirical observation is consistent with *dome-like* (elliptic) geometry,

$$M > 0, \qquad \text{and we infer } G > 0 \text{ (both } \kappa_i \text{ share sign)},$$

even though G is not measured.

Non-memorized. Typical behavior:

$$\kappa^{\Delta}(t_0) < 0 \;\longrightarrow\; \exists\, \tau :\; \kappa^{\Delta}(\tau) = 0 \;\longrightarrow\; \kappa^{\Delta}(t) \gtrsim 0 \text{ (small)},$$

this is compatible with boundary-trap/absorbing dynamics and *non-collapsed* outputs; A *boundary trap* means solution traverses into and remains in a compact absorbing set, and is a typical case of curvature sign-switch as observed here. Note that here G may have either sign - resulting in a flat or saddle point at the end of the process - indeed allowing for diverse output images.

Takeaway. Memorization shows *early, increasing* $M > 0$ *with empirically inferred* $G > 0$, i.e. a dome geometry that steepens as generation progresses; valid generations show $M < 0 \to 0 \to M \gtrsim 0$ with diversity preserved.

4. Experiments

In this section, we demonstrate the capabilities of our method. We evaluate the ability of κ^{Δ} and related metrics to distinguish memorized from non-memorized generations.

4.1. Settings and frameworks

We evaluate on *Stable Diffusion v1.4* with classifier-free guidance (CFG) scale 7.5. Unless stated otherwise, all detectors are computed at the first denoising step ($t{=}T$), which is the most actionable point for mitigation.

Prompt sets. Memorized prompts follow Wen et al. (2024); non-memorized prompts are LLM-generated to avoid overlap with training data Wen (2025).

Detection baselines. We compare our direction-only criterion κ^{Δ} against: (i) CFG-MAG—the score-norm gap under CFG Wen et al. (2024); (ii) HCP-MAG—the Hessian–score product magnitude Jeon et al. (2025); and (iii) PLAP—a post-hoc, image-based p-Laplace detector Brokman et al. (2025). The first two are mid-generation; PLAP is designed for post-generation.

Hybrid detectors. To test complementarity, we learn a linear SVM over the 2D feature space [κ^{Δ}, baseline] on a small held-out split and report test AUC. We also compare against a hybrid of CFG-MAG and HCP-MAG.

Mitigation framework. Following Wen et al. (2024), we perform soft-prompt optimization at $t{=}T$ for memorized prompts:

$$L(c) \;=\; L_{\text{miti}}(x_T, c) \;+\; \lambda \|c - c_0\|_2,$$

where c_0 is the initial text embedding and $x_T \sim \mathcal{N}(0, I)$. We instantiate L_{miti} with κ^{Δ} (ours) or each baseline criterion and compare fidelity–memorization trade-offs. Other generation hyperparameters follow the respective works.

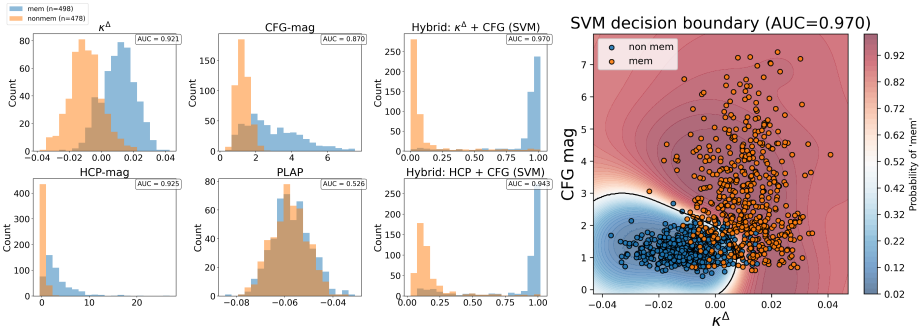

Figure 3: **Early-step detection at t=T (first generation step): univariate histograms and 2D hybrid.** Left: per-criterion histograms for memorized (orange) and non-memorized (blue) generations with AUCs overlaid. Our κ^Δ separates well (AUC = 0.922), CFG magnitude is competitive (AUC = 0.885), HCP magnitude is slightly higher in isolation (AUC = 0.932), while the post-hoc PLAP signal is uninformative at this early generation step (AUC = 0.550). Right: scatter of x=κ^Δ vs. y=CFG-mag with a linear SVM and regions. The hybrid (κ^Δ+CFG-mag) gives the best performance (AUC = 0.976), exceeding either component alone and a magnitude-only hybrid (HCP-mag + CFG-mag, AUC = 0.948). *Directional* geometry is highly informative at the first step and complements magnitude-based criteria to achieve SOTA.

4.2. Results

First generation step (t=T) separability. Figure 3 summarizes detection at the first denoising step, the most critical regime for on-the-fly mitigation. Our curvature surrogate κ^Δ already yields strong separability between memorized and non-memorized prompts (AUC = 0.922), outperforming the CFG magnitude baseline (CFG-mag, AUC = 0.885). The Hessian–score magnitude (HCP-mag) is slightly higher in isolation (AUC = 0.932), while the post-hoc 1-Laplace proxy (PLAP) is not informative at t=T (AUC = 0.550), as expected for a criterion designed for post-generation.

Complementarity of direction and magnitude. Because κ^Δ depends on *direction* (normalized field) and CFG/HCP depend on *magnitude*, the signals are complementary. A simple linear SVM trained on a small held-out calibration split combining (κ^Δ, CFG-mag) achieves the best mid-generation discrimination (AUC = 0.976), improving over either component alone and also over the hybrid (HCP-mag + CFG-mag) (AUC = 0.948). The 2D scatter and decision regions in Figure 3 (right) visualize this complementarity.

Temporal signature at scale. The population trajectories in Figures 1–2 show a consistent *shifting-geometry* pattern across hundreds of prompts per class: memorized prompts enter and remain in attractive (convex) regions with κ^Δ>0 that strengthen with time; non-memorized prompts begin with κ^Δ<0, pass through near-flat κ^Δ≈0, and only later reach weakly positive values. This gap persists over a wide range of generation steps and explains why early-time κ^Δ is predictive.

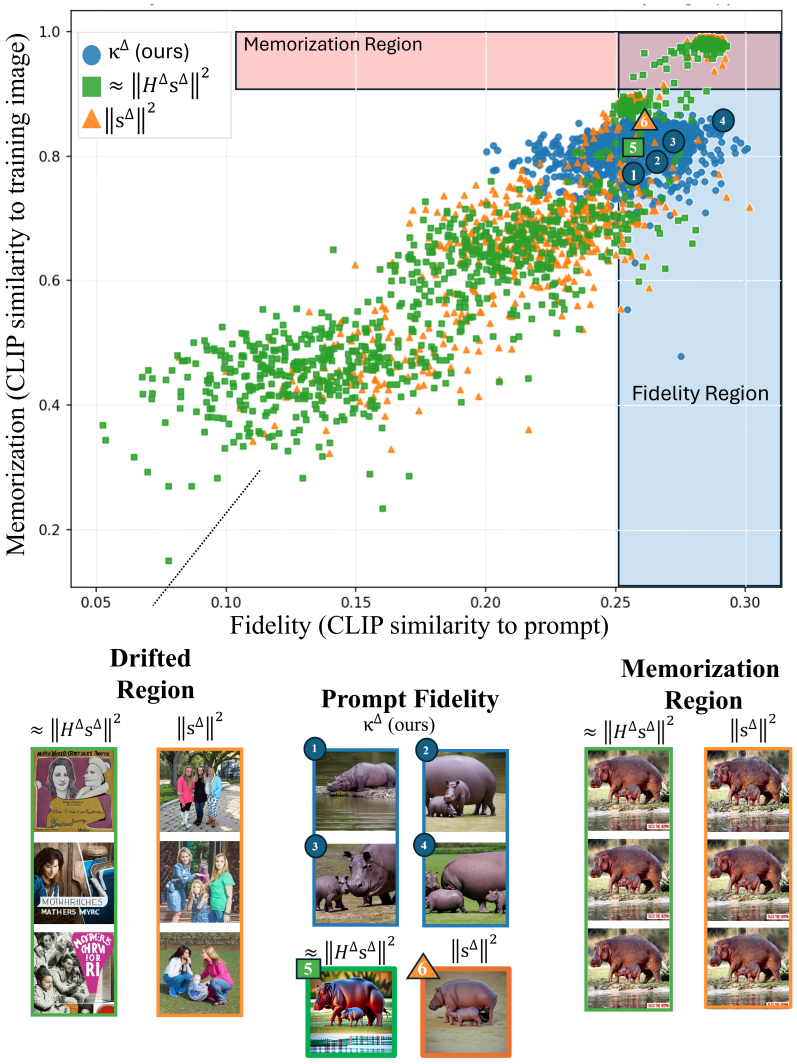

Figure 4: **Soft-prompt mitigation on a memorized prompt: A CLIP×CLIP of the fidelity–memorization trade-off .** Scatter shows *Fidelity* (CLIP cosine to the prompt) vs. *Memorization* (CLIP cosine to the training image) for 3×1000 generations using three criteria inside Wen et al.'s soft-prompt optimizer: Our κ^{Δ} (blue), the SAIL criterion $\approx\|H^{\Delta}s^{\Delta}\|^2$ (green), and the gap norm $\|s^{\Delta}\|^2$ (orange). *Memorization Region:* The horizontal red band containing the top-right cluster - we verified that it contains all memorized images. *Fidelity Region*: the vertical band matching the fidelity of memorized copies. The desired mitigation outcome is in the blue rectangle (bottom-right): matching the prompt fidelity of the memorized image, with low similarity to the actual training image. Thumbnails below illustrate typical outcomes by region: *Drifted* (low fidelity), *Prompt-fidelity / non-memorized* (target), and *Memorization*. Numbered markers correspond to the highlighted examples. Compared with the competing criteria, our κ^{Δ} concentrates samples in the target region and reduces both drift and memorization.

Mitigation outcomes. Plugging κ^Δ into the soft-prompt optimizer of Wen et al. (2024) shifts samples toward the target *high-fidelity, non-memorized* region (Fig. 4, blue), reducing both outright memorization (high training-image similarity) and fidelity drift (low prompt alignment). Qualitatively, κ^Δ steers trajectories away from convex basins associated with verbatim copies while preserving text fidelity, in line with the observed temporal geometry.

Evaluation. *Fidelity* measures CLIP cosine similarity between the generated image and the text prompt. *Memorization* denotes CLIP cosine similarity between the generated image and the specific training image for the prompt. We visualize the joint distribution of these metrics and show representative samples at different operating regions. We perform this analysis on SD 1.4, on a known memorized prompt, "Mothers influence on her young hippo."

Findings. Figure 4 summarizes the trade-offs:

- κ^Δ **(ours):** concentrates samples in the *high-fidelity / low-memorization* band. Qualitatively, outputs remain on-prompt (hippos) without reproducing the training instance.
- **SAIL sharpness** $\|H^\Delta s^\Delta\|^2$**:** reduces memorization but at a substantial cost in prompt fidelity; many samples drift off the target semantics.
- **Gap norm** $\|s^\Delta\|^2$**:** either fails to suppress memorization for a non-trivial subset or, when it succeeds, often degrades fidelity.

Overall, κ^Δ delivers the most favorable fidelity–memorization balance among the three, aligning with our hypothesis that a magnitude-invariant, geometry-aware signal is better suited for mid-generation steering.

5. Summary

We introduced a magnitude–invariant, geometry–aware detector for diffusion memorization based on the *normalized boundary flux* of the CFG gap, denoted κ^Δ. The quantity is interpretable as a mean–curvature surrogate of the learned log–density along the generation trajectory, and it isolates the directional effect. Empirically, κ^Δ reveals a consistent *shifting-geometry* signature: memorized prompts quickly move into and remain in attractive (positive-curvature) regimes, while non-memorized prompts begin negative curvature regimes, pass through near-flat geometry, and only later reach weak attraction. In terms of limitations: While κ^Δ is a novel mean-curvature proxy - if we want a complete analysis of surfaces, incorporating a generalization of the Gaussian curvature would be fitting; Additionally, we use the SD-1.4 model; broader validation across diverse architectures remains for future work. In terms of performance - we show across thousands of image generations that κ^Δ separates mem/non-mem already at the earliest step with strong AUC, and sets a new SOTA when combined with norm-based criteria such as $\|s^\Delta\|^2$. Turning the signal into a plug-in policy for soft-prompt optimization improves the fidelity–memorization trade-off relative to these baselines. Overall, this work proposes novel geometric signatures and advances memorization research of generative models - improving practical tasks with geometrical observations.

References

Joe Benton, Carles Domingo-Enrich, Stephan Hölzl, Matthew J. Johnson, Emile Mathieu, and Arnaud Doucet. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=ifS8R0ZQei. arXiv:2308.03686.

Valentin De Bortoli, Émile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/105112d52254f86d5854f3da734a52b4-Paper-Conference.pdf.

Jonathan Brokman, Amit Giloni, Omer Hofman, Roman Vainshtein, Hisashi Kojima, and Guy Gilboa. Identifying memorization of diffusion models through p-laplace analysis. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 295–307. Springer, 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.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://papers.nips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A*, 4(12):2379–2394, 1987. doi: 10.1364/JOSAA.4.002379. URL https://opg.optica.org/josa/abstract.cfm?uri=josaa-4-12-2379.

Dominik Hintersdorf, Lennart Struppek, Hendrik Schuff, Benjamin Hilprecht, and Kristian Kersting. Localizing neurons responsible for memorization in diffusion models. *arXiv preprint arXiv:2406.02366*, 2024. URL https://arxiv.org/abs/2406.02366. NeurIPS 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020. doi: 10.48550/arXiv.2006.11239. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Hailong Hu and Jun Pang. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*, 2023. URL https://arxiv.org/abs/2301.09956. Also published as: *Loss and Likelihood Based Membership Inference of Diffusion Models*, ESORICS 2023 Workshops, Springer LNCS.

Jinggang Huang and David Mumford. Statistics of natural images and models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 541–547, 1999.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding and mitigating memorization in generative models via sharpness of probability landscapes. In *Forty-second International Conference on Machine Learning*, 2025.

Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ANvmVS2Yr0.

Koichi Miyasawa et al. An empirical bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist*, 38(181-188):1–2, 1961.

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

Yixin Pang, Haoran Sun, Rui Zhou, Jian Zhang, Yuhui Shi, Yinzhi Cao, and Wei Yang. White-box membership inference attacks against diffusion models. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, (3):149–170, 2025. URL https://petsymposium.org/popets/2025/popets-2025-0068.php.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–71, 2000. doi: 10.1023/A:1026553619983. URL https://www.cns.nyu.edu/pub/lcv/portilla99-reprint.pdf.

Andrei Potaptchik, Ryan Murray, Raanan Fattal, Amit Singer, and Joan Bruna. Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*, 2024. URL https://arxiv.org/abs/2410.09046.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*, pages 340–356. Springer, 2024.

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)*, pages 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01041.

Daniel L. Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994. doi: 10.1103/PhysRevLett.73.814. URL https://link.aps.org/doi/10.1103/PhysRevLett.73.814.

Kota Sakamoto and Taiji Suzuki. Geometry of diffusion models: Tubular neighborhoods and singularities. In *Proceedings of the ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling (GRaM)*, volume 251 of *Proceedings of Machine Learning Research*, 2024. URL https://proceedings.mlr.press/v251/sakamoto24a.html.

Eero P. Simoncelli. Geometry of the distribution of natural images. Invited lecture at the NeurIPS 2024 Workshop "Symmetry and Geometry in Neural Representations" (NeurReps 2024), Vancouver, Canada, 2024. URL https://slideslive.com/39030660/geometry-of-the-distribution-of-natural-images. Video recording on SlidesLive. Accessed 2025-09-02.

Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001. doi: 10.1146/annurev.neuro.24.1.1193. URL https://www.cns.nyu.edu/pub/eero/simoncelli01-reprint.pdf.

Nir Sochen, Ron Kimmel, and Ravi Malladi. A general framework for low level vision. *IEEE transactions on image processing*, 7(3):310–318, 1998.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2021. URL https://arxiv.org/abs/2011.13456.

Chengchun Tang, Jianfeng Yang, Hang Zhao, and Yaodong Yu. Adaptivity of diffusion models to manifold structures. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, 2024a. URL https://proceedings.mlr.press/v238/tang24a.html.

Shuai Tang, Steven Wu, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership inference attacks on diffusion models via quantile regression. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 47819–47829, 2024b. URL https://proceedings.mlr.press/v235/tang24g.html.

A. van der Schaaf and J. H. van Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770, 1996. doi: 10.1016/0042-6989(96)00002-8.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Martin J. Wainwright and Eero P. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems (NeurIPS) 12*, pages 855–861, 2000. URL https://proceedings.neurips.cc/paper/1999/file/6a5dfac4be1502501489fc0f5a24b667-Paper.pdf.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023. doi: 10.48550/arXiv.2305.08694. URL https://arxiv.org/abs/2305.08694.

Yuxin Wen. Diffusion memorization examples. https://github.com/YuxinWenRick/diffusion_memorization/blob/main/examples/sdv1_500_memorized.jsonl, 2025. Accessed: 2025-09-03.

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.

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 (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/874411a224a1934b80d499068384808b-Abstract-Conference.html.

Daniel Zoran and Yair Weiss. Natural images, gaussian mixtures and dead leaves. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1736–1744, 2012.

Appendix A. Mid-Generation Detection

We provide snapshots of the metric evaluated at 3 for generation steps $t = 50, 100, 150, 200, 250, 300$.

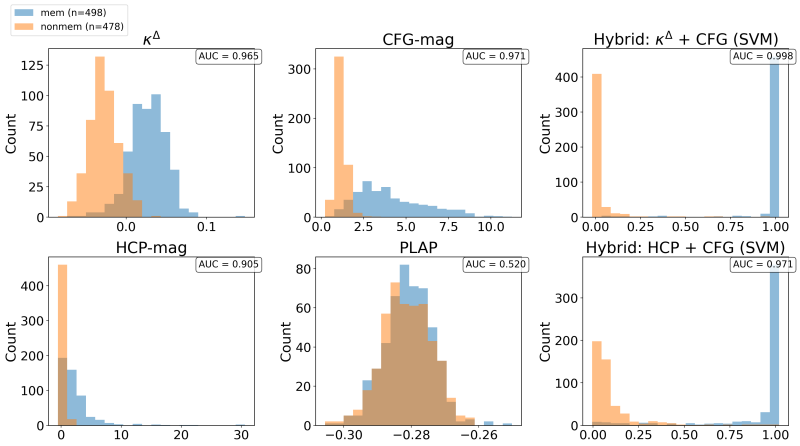

Figure 5: Generation step $t = 50$.

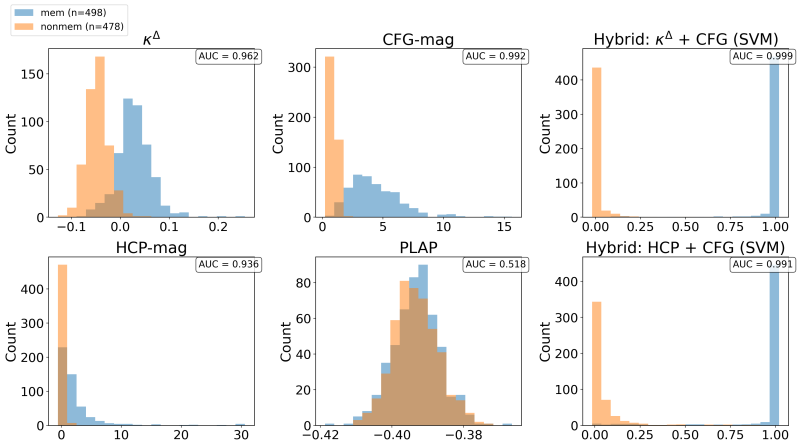

Figure 6: Generation step $t = 100$.

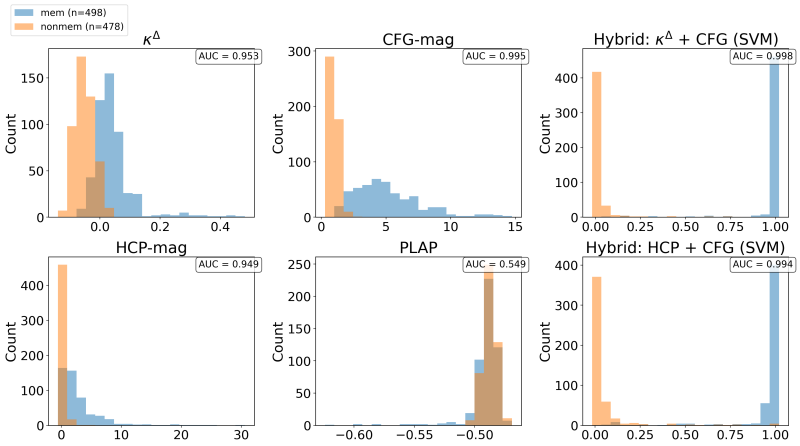

Figure 7: Generation step $t = 150$.

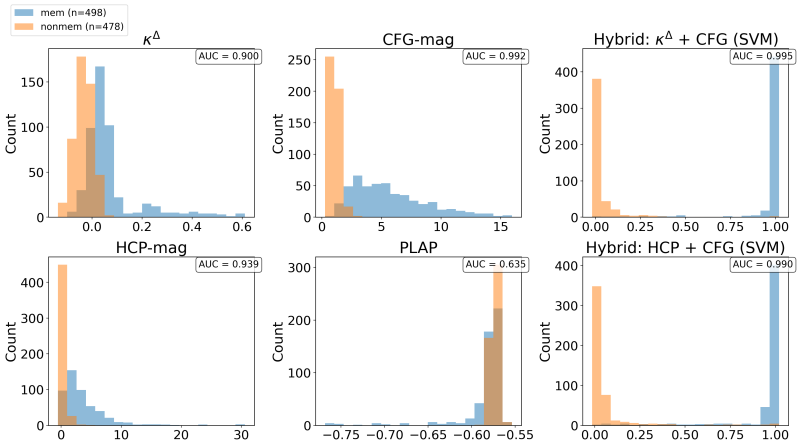

Figure 8: Generation step $t = 200$.

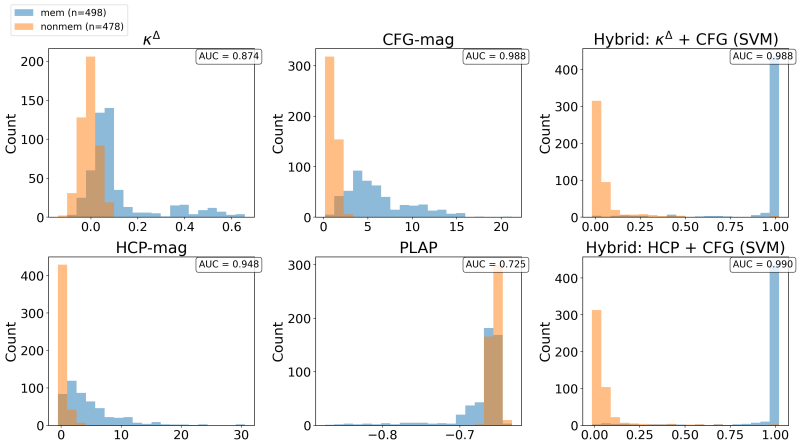

Figure 9: Generation step $t = 250$.

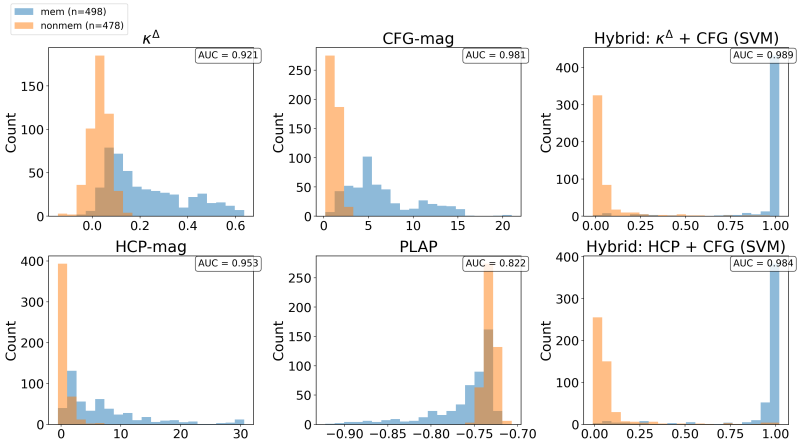

Figure 10: Generation step $t = 300$.