Tracking Memorization Geometry throughout the Diffusion Model Generative Process

Jonathan Brokman* JONATHAN.BROKMAN@FUJITSU.COM
Itay Gershon* ITAY.GERSHON@FUJITSU.COM
Omer Hofman OMER.HOFMAN@FUJITSU.COM
Guy Gilboa GUY.GILBOA@EE.TECHNION.AC.IL
Roman Vainshtein ROMAN.VAINSHTEIN@FUJITSU.COM

Editors: List of editors' names

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.

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

* Equal contribution.

approximated manifold Bortoli et al. (2022); Brokman et al. (2024b); Kadkhodaie et al. (2024); 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: [1]

- A novel **magnitude-invariant mid-generation detection criterion** based on the direction (not the size) of the 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.** Tracking κ^{Δ} over generation steps reveals a clear phase-structured behavior: memorized prompts start in an attracting region ($\kappa^{\Delta} > 0$) and grow rapidly toward a dominant mid–late concavity peak, whereas non-memorized prompts typically begin in a repelling convex regime ($\kappa^{\Delta} < 0$) before gradually crossing into a weaker concave regime that peaks around the same time. The separation is therefore governed by *early sign-geometry* and *later concavity strength*.

- **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

Prior work has shown that diffusion memorization can be revealed through mid-generation signals. Wen et al. (2024) detect memorized prompts by measuring the magnitude of the classifier-free guidance (CFG) score gap throughout the denoising process, demonstrating that memorization produces abnormally large conditional–unconditional score norms even at the first generation step. Jeon et al. (2025) further analyze memorization via the sharpness of the conditional probability landscape, using Hessian–score products. Both

1. Official implementation: https://github.com/JonathanBrok/Tracking-Memorization-Geometry-throughout-the-Diffusion-Model-Generative-Process/tree/main

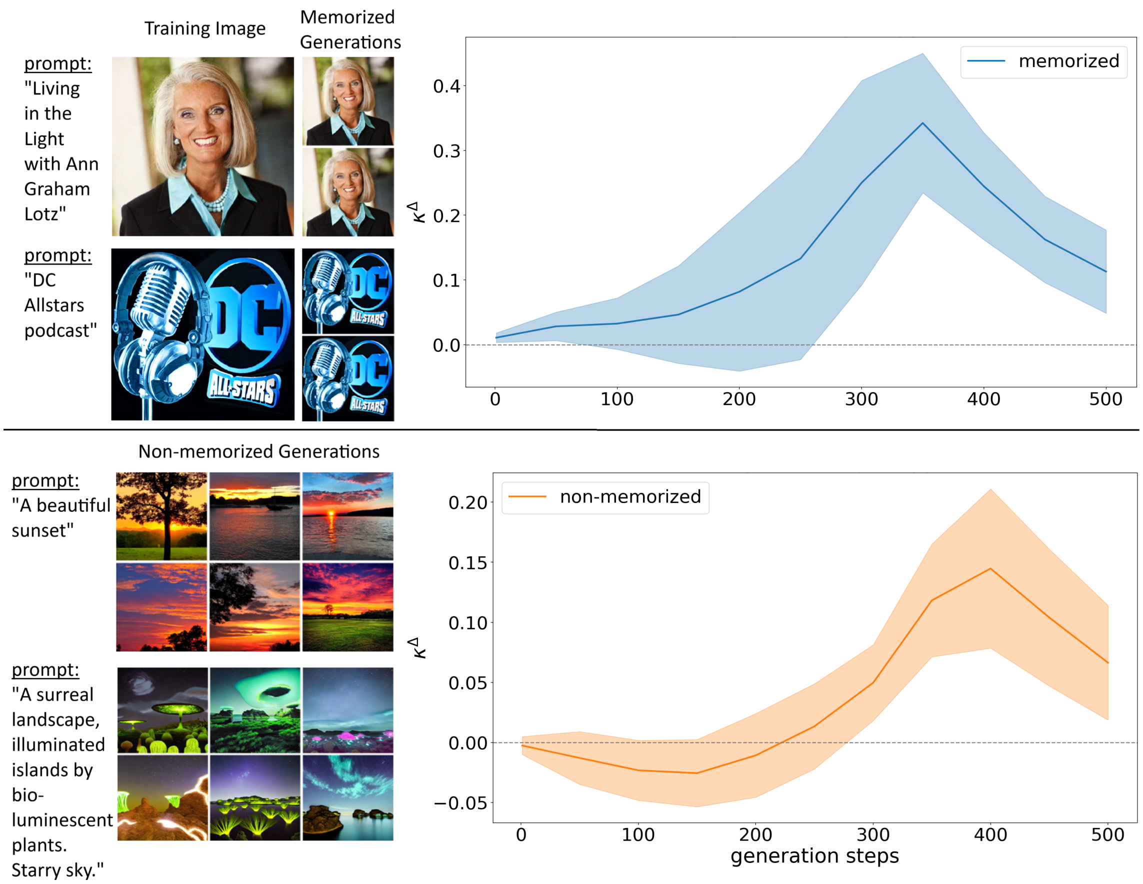

Figure 1: **Left:** Memorized prompts generate near-verbatim copies of a training images, unlike non-memorized prompts. **Right:** Temporal evolution of the curvature signal κ^Δ across generation steps. Memorized prompts immediately enter an attracting (concave) regime and steadily rise to a strong mid–late concavity peak. Non-memorized prompts typically begin in a repelling (convex) regime with $\kappa^\Delta < 0$, but later transition into a weaker positive concavity peak occurring in a similar time window. The distinction therefore lies in **early-phase geometry (sign)** and **mid-phase concavity strength**.

approaches rely on magnitude-dependent differentials of the score field. In contrast, our method introduces a magnitude-invariant, geometry-aware criterion based solely on the *direction* of the CFG difference. This yields a signed curvature surrogate, enabling us to track how prompt conditioning bends the generative trajectory's local geometry and to combine this signal with magnitude-based methods for improved early-step detection and mitigation. Closest to our perspective, Brokman et al. (2025) propose p-Laplace–based detectors that apply higher-order differentials of the *final* score field, including the total-variation–based mean-curvature surrogate that we employ, in the spirit of classical image and shape-processing methods Sochen et al. (1998); Brokman and Gilboa (2021); Brokman et al. (2024a). However, it is a post-hoc approach and is not suited for *mid-generation* steering (demonstrated in experiments below), or mitigation in general - since mitigation requires interfering during and not after the generative process. Other less related 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 a direct geometrical interpretations.

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: Given a noising process $q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)I\big)$, the MMSE noise predictor aligns with $-s_t(x_t)$. A neural network trained to predict noise thus estimates the score over time steps. See Miyasawa et al. (1961); Nichol and Dhariwal (2021).

Forward and reverse processes. The forward process progressively corrupts x_0 into noise, and its evolving densities q_t satisfy the Fokker–Planck equation associated with the chosen diffusion. The reverse process inverts this evolution: starting from $x_T \sim \mathcal{N}(0, I)$ ($\alpha_T = 0$), it uses the learned scores to reconstruct x_0 ($\alpha_0 = 1$) by following the reverse-time diffusion whose drift is defined by the Fokker–Planck time-reversal formula. See Anderson (1982); Ho et al. (2020); Song et al. (2021).

SDE view in text-conditional latent diffusion. In text-conditional settings, classifier-free guidance (CFG) provides unconditional and conditional scores, $s_\theta(x_t)$ and $s_\theta(x_t, c)$. The reverse dynamics can be written as an SDE driven by the score. In latent diffusion, the SDE 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. See Rombach et al. (2022); Ho and Salimans (2022).

3.2. Method: Decode–encode boundary flux of classifier-free guidance

We measure how prompt conditioning deforms the log-probability landscape by estimating a mean-curvature analogue of $\log \frac{p(z_t|c)}{p(z_t)}$. This is done through a normalized boundary-flux divergence of the classifier-free guidance score gap in latent space.

CFG score gap. For a latent z_t and prompt c, let $s_\theta(z_t, c)$ and $s_\theta(z_t)$ be the conditional and unconditional scores. Their classifier-free guidance gap, and its normalized version are

$$s_t^\Delta(z_t, c) := s_\theta(z_t, c) - s_\theta(z_t), \qquad \hat{s}_t^\Delta(z) := \frac{s_t^\Delta(z, c)}{\|s_t^\Delta(z, c)\|_\delta},$$

with $\|v\|_\delta = \sqrt{\delta + \|v\|^2}$ ensuring stability ($\delta > 0$ ensures a strictly positive denominator).

Boundary-flux curvature. Let $B_{R_t}(c_t)$ be a small latent ball centered at c_t with radius R_t, and let n be the outward unit normal on $\partial B_{R_t}(c_t)$. The curvature-like quantity we estimate is the normalized flux

$$\kappa^{\Delta}(c_t) := \frac{1}{|B_{R_t}|} \int_{\partial B_{R_t}(c_t)} \hat{s}_t^{\Delta}(z) \cdot n \, ds,$$

which limits to the mean curvature analogue $\nabla \cdot \hat{s}_t^{\Delta}(c_t)$ as $R_t \to 0$. Using N Monte Carlo samples $y_i \in \partial B_{R_t}(c_t)$ with normals $n_i = (y_i - c_t)/R_t$,

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

Scaled centers and decode–encode refinement. In latent diffusion, $z_t \approx \sqrt{\bar{\alpha}_t}\, z_0 +$ noise, suggesting the geometry be probed at

$$c_t = \sqrt{\bar{\alpha}_t}\, z_0, \qquad R_t = \sqrt{1 - \bar{\alpha}_t}\, R_0.$$

In Brokman et al. (2024b) they observed, in a different context, that processing SD latent space via a decode–encode process may improve separability. To retain this benefit in latent space, we optionally replace z_0 by a decode–encode refinement

$$\hat{z}_0 := E(D(z_t)), \qquad c_t := \sqrt{\bar{\alpha}_t}\, \hat{z}_0. \tag{1}$$

The estimator applies equally to z_0 or $\hat{z}_0$, requiring only evaluations of $\hat{s}_t^{\Delta}$ in latent space.

Interpretable and practical curvature aspect. Since $s_\theta(z_t, c) \approx \nabla_{z_t} \log p(z_t \mid c)$ and $s_\theta(z_t) \approx \nabla_{z_t} \log p(z_t)$, our estimator satisfies

$$\kappa^{\Delta}(c_t) \;\approx\; \nabla_{z_t} \cdot \left(\frac{\nabla_{z_t} \log\left(\frac{p(z_t|c)}{p(z_t)}\right)}{\left\| \nabla_{z_t} \log\left(\frac{p(z_t|c)}{p(z_t)}\right) \right\|_{\delta}} \right),$$

i.e. it measures a **time-sensitive** mean-curvature analogue of the prompt-induced deformation of the latent probability geometry. Being t-sensitive, we can track $\kappa^{\Delta}(z_t)$ along the generation trajectory to profile how prompt-induced curvature evolves over time. Such trajectories were gathered in Fig. 1.

For interpretability, we can discuss the observed dynamics (Fig. 1) as if we are traversing a 2D surface in 3D. Then $\kappa^{\Delta} = \nabla \cdot \hat{s}^{\Delta}$ acts as a signed mean-curvature. Let the principal curvatures be κ_1, κ_2 so

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

are the *mean-curvature* (M), and *Gaussian curvature* (G). Usually, in simple 2-surface analysis (embedded in 3D) we perform full surface analysis by combining M and G to classify elliptic, hyperbolic, or parabolic points, but since κ^{Δ} captures only the mean-curvature component it cannot provide such classification; nonetheless, its boundary-integral form exposes whether the prompt induces elliptic-like behavior (consistent inward/outward bending

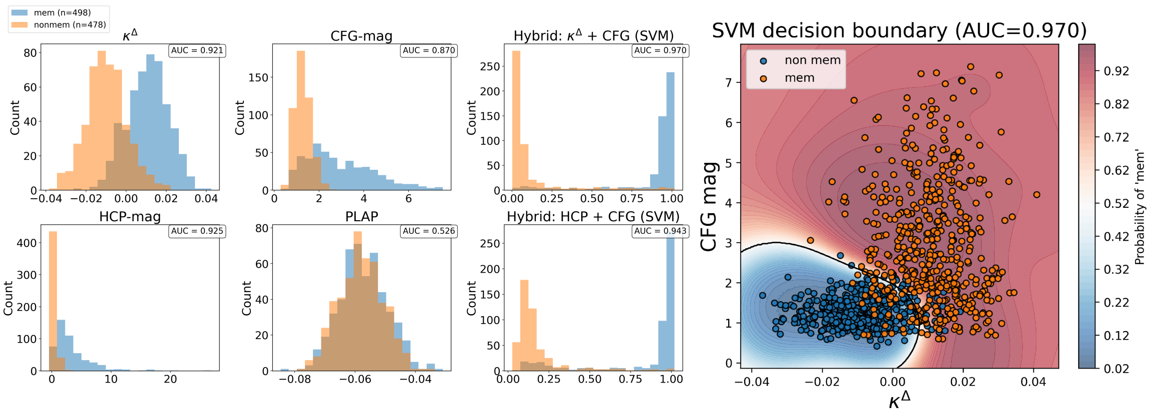

Figure 2: **Methods comparison - detection at t=T (first generation step): Univariate and 2D hybrid.** Left: Histograms for memorized (orange) and non-memorized (blue) generations with AUCs overlaid. Our κ^Δ separates well (AUC = 0.922), CFG magnitude is competitive (AUC = 0.885), HCP magnitude is slightly 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). Take-away: *Directional* geometry is highly informative at the first step and complements magnitude-based criteria to achieve SOTA.

of normalized flux) or hyperbolic-like behavior (mixed bending directions), with the sign of κ^Δ indicating whether local behavior is more towards convex vs. concave bending of probability mass in latent space.

Numerical meaning. Numerically, κ^Δ aggregates how much the (normalized) score gap points *inward* to the local center - effectively summing the inward-facing component of the approximated probability gradient over a small spherical neighborhood. Positive values therefore indicate that conditional guidance bends probability mass inward (an attracting geometry), while negative values correspond to outward-bending or repelling behavior.

Memorized. κ^Δ is positive from the earliest steps, reflecting an attracting concave-leaning geometry. It rises steadily and forms a pronounced peak in mid–late generation, then decreases slightly while remaining positive.

Non-memorized. κ^Δ begins negative, corresponding to a repelling (convex-leaning) geometry. Over time it gradually increases, crosses zero, and enters a weaker concave-leaning regime with a smaller positive peak.

Bottom line. Memorized and non-memorized trajectories tend to have opposite signs in early steps, are *concave and convex-leaning respectively at the start.* Later non-memorized transitions into the attracting behaviour, but with weaker concavity compared to the memorized case, where both exhibit their strongest concavity in a similar time window.

4. Experiments

Let us evaluate the ability of κ^{Δ} and related metrics to detect 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. Here we advocate for the decode–encode refinement (Eq. (1). 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–CFG score product magnitude Jeon et al. (2025); and (iii) PLAP—a p-Laplace detector designed for post-hoc detection ($t = 0$, last generation step, unlike here) Brokman et al. (2025).

Hybrid detectors. To test complementarity, we learn a 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. Here we keep it simple, without encode–decode refinement. 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.

4.2. Results

First generation step ($t=T$) separability. Figure 2 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 2 (right) visualize this complementarity.

Temporal signature at scale. The population trajectories in Figure 1 show a consistent *geometric* pattern across hundreds of prompts per class: for memorized prompts, κ^{Δ} rapidly departs from a narrow band near zero and climbs to a pronounced positive peak at mid–late

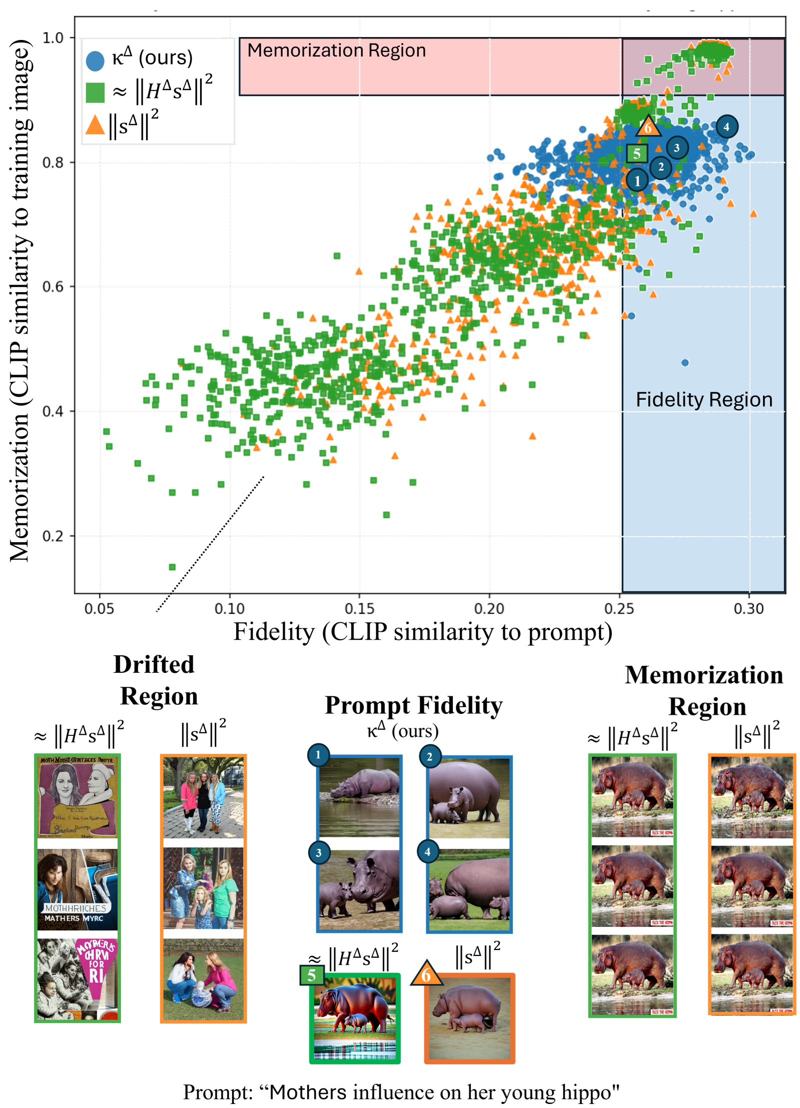

Figure 3: **Soft-prompt mitigation on a memorized prompt: A CLIP×CLIP of the fidelity–memorization trade-off .** Scatter shows *Fidelity* (CLIP cosine of generated image to the prompt) vs. *Memorization* (CLIP cosine of the generated image to the training image) for 3×1000 generations, plugging three criteria to 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 (valid non-memorized generations)*, and *Memorization Region.* 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.

steps, whereas for non-memorized prompts it symmetrically dips to a negative trough at roughly the same time indices. Importantly, the histograms in Figure 2 reveal that the two populations are already well separated even when κ^{Δ} is close to zero, so the subsequent zero-to-peak excursion sharpens—rather than creates—the distinction between memorized and non-memorized trajectories.

Mitigation outcomes. Figure 3 shows performance for the "Mothers influence on her young hippo" memorized prompt, showing clear 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** $\|s^{\Delta}\|^2$**:** exhibits a similar trade-off, with a more prominent possible semantic drift. 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 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}, with an optional decode–encode refinement step. 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 *geometric signature*: memorized prompts start in an attracting regime with $\kappa^{\Delta} > 0$ that strengthens until a mid–late peak, whereas non-memorized prompts begin in a repelling regime with $\kappa^{\Delta} < 0$ and only later transition into a weaker concave regime. Importantly, this entire behavior is expressed on the log-density ratio $\log\left(\frac{p(z_t|c)}{p(z_t)}\right)$. 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

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

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 and Guy Gilboa. Nonlinear spectral processing of shapes via zero-homogeneous flows. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 40–51. Springer, 2021.

Jonathan Brokman, Martin Burger, and Guy Gilboa. Spectral total-variation processing of shapes—theory and applications. *ACM transactions on graphics*, 43(2):1–20, 2024a.

Jonathan Brokman, Amit Giloni, Omer Hofman, Roman Vainshtein, Hisashi Kojima, and Guy Gilboa. Manifold induced biases for zero-shot and few-shot detection of generated images. In *The Thirteenth International Conference on Learning Representations*, 2024b.

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.