Tracking Memorization Geometry throughout the Diffusion Model Generative Process

FUJITSU TECHNION NEURAL INFORMATION PROCESSING SYSTEMS

Jonathan Brokman*, Itay Gershon*, Omer Hofman, Guy Gilboa, Roman Vainshtein

*Equal contributors

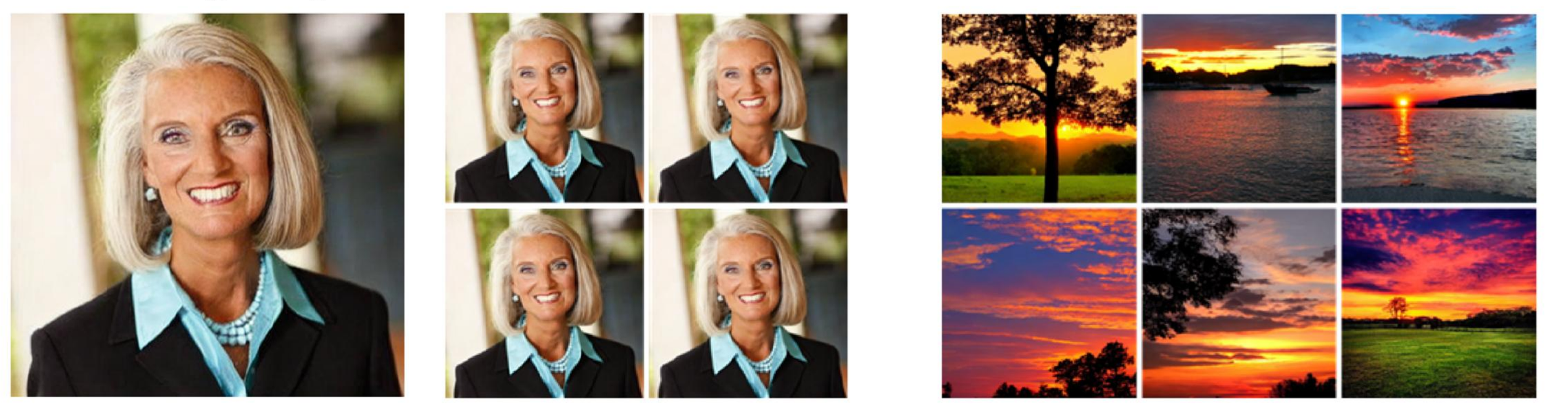
Introduction

Diffusion models may reproduce training images verbatim, creating privacy and copyright risks. Existing detectors depend on score magnitude.

Contribution:

- Magnitude-invariant criterion κ^Δ, expressing curvature of $\log\left(\frac{p(z_t\mid c)}{p(z_t)}\right)$
- Captures memorization throughout the generative process
- Complements the existing magnitude-based approaches

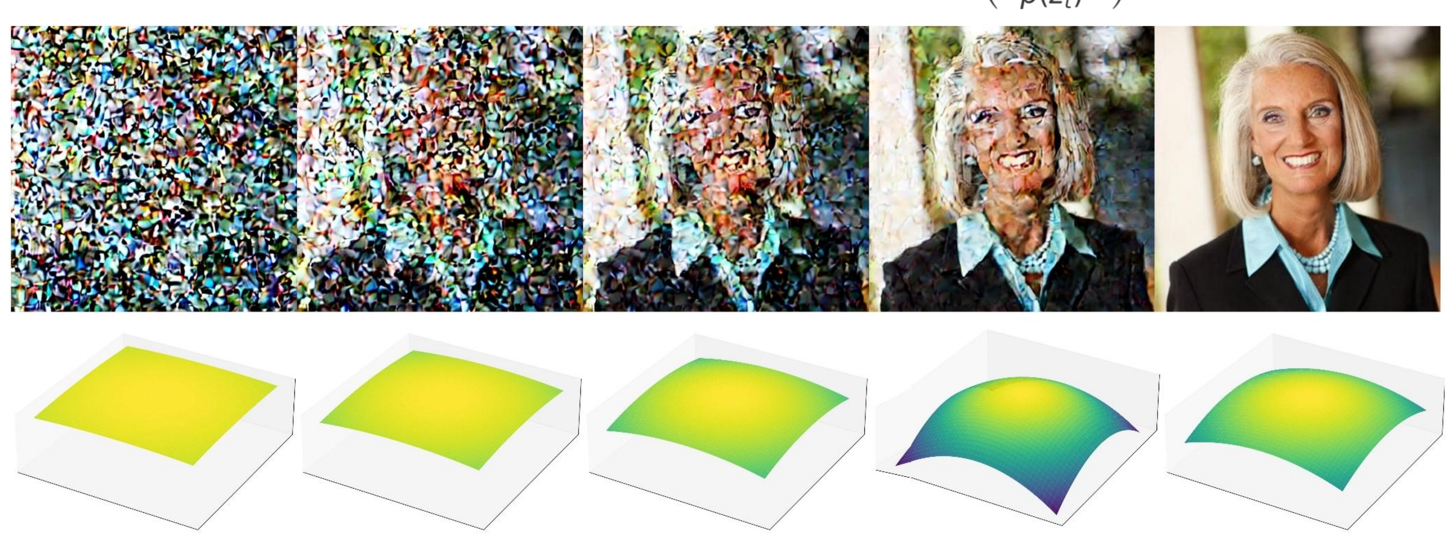
Training Image | Memorized Generations | Non-Memorized Generations



Prompt: "Living in the Light with Ann Graham Lotz" Prompt: "A beautiful sunset"

Method

We measure how prompt conditioning deforms the log-probability landscape by estimating a mean-curvature analogue for $\log\left(\frac{p(z_t\mid c)}{p(z_t)}\right)$.



Background: Stable Diffusion models operate in latent space Z. They estimate unconditioned $s_\theta(z_t) \approx \nabla_{z_t}\log p(z_t)$ and prompt c conditioned $s_\theta(z_t, c) \approx \nabla_{z_t}\log p(z_t \mid c)$ to reverse the nosing process

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon$$

We use:

$$s_\Delta(z_t, c) = s_\theta(z_t, c) - s_\theta(z_t), \qquad \hat{s}_\Delta(z_t) = \frac{s_\Delta(z_t)}{\|s_\Delta(z_t)\|}$$

To define an easy-to-obtain numeric criterion

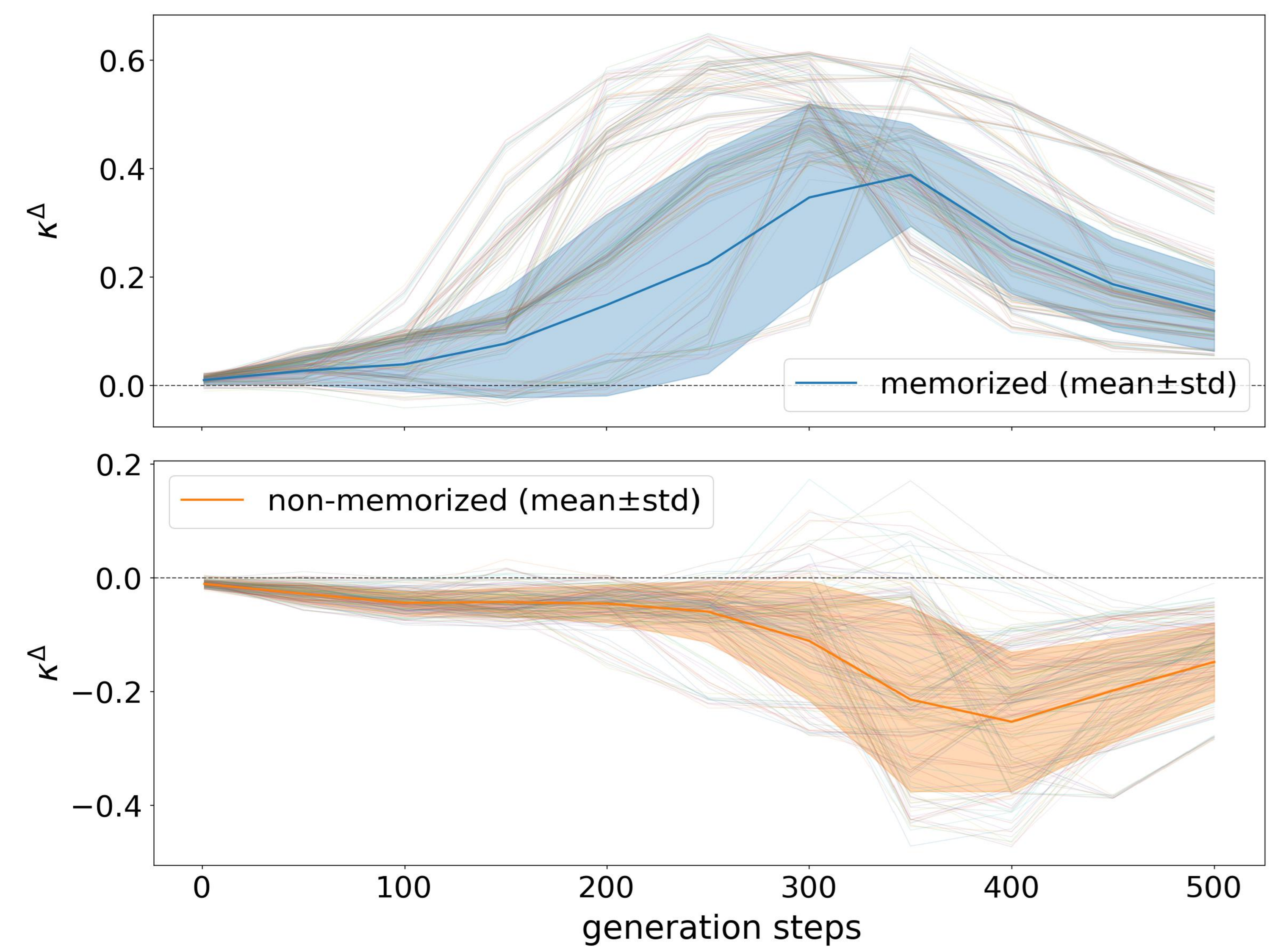
$$\kappa^\Delta(c_t) = \frac{1}{|\partial B_{R_t}(c_t)|}\int_{\partial B_{R_t}(c_t)} \hat{s}_\Delta(z)\cdot n(z)\, dS \approx \frac{d}{R_t N}\sum_{i=1}^{N}\hat{s}_\Delta(y_i)\cdot n_i$$

It is important to select c_t, R_t in accordance with the noise level at t. Namely - **∂B_t should be highest probability hyper-sphere** of the noised sample $p(z_t|z_{t+1})$ - namely $c_t = \sqrt{\bar{\alpha}_t}\, z_0$ and $R_t = \sqrt{d\,(\alpha_t - 1)}$. This ensures using s_θ on points where it was trained. y_i are uniformly sampled on ∂B_t and the sum is providing a monte-carlo estimate of the integral.

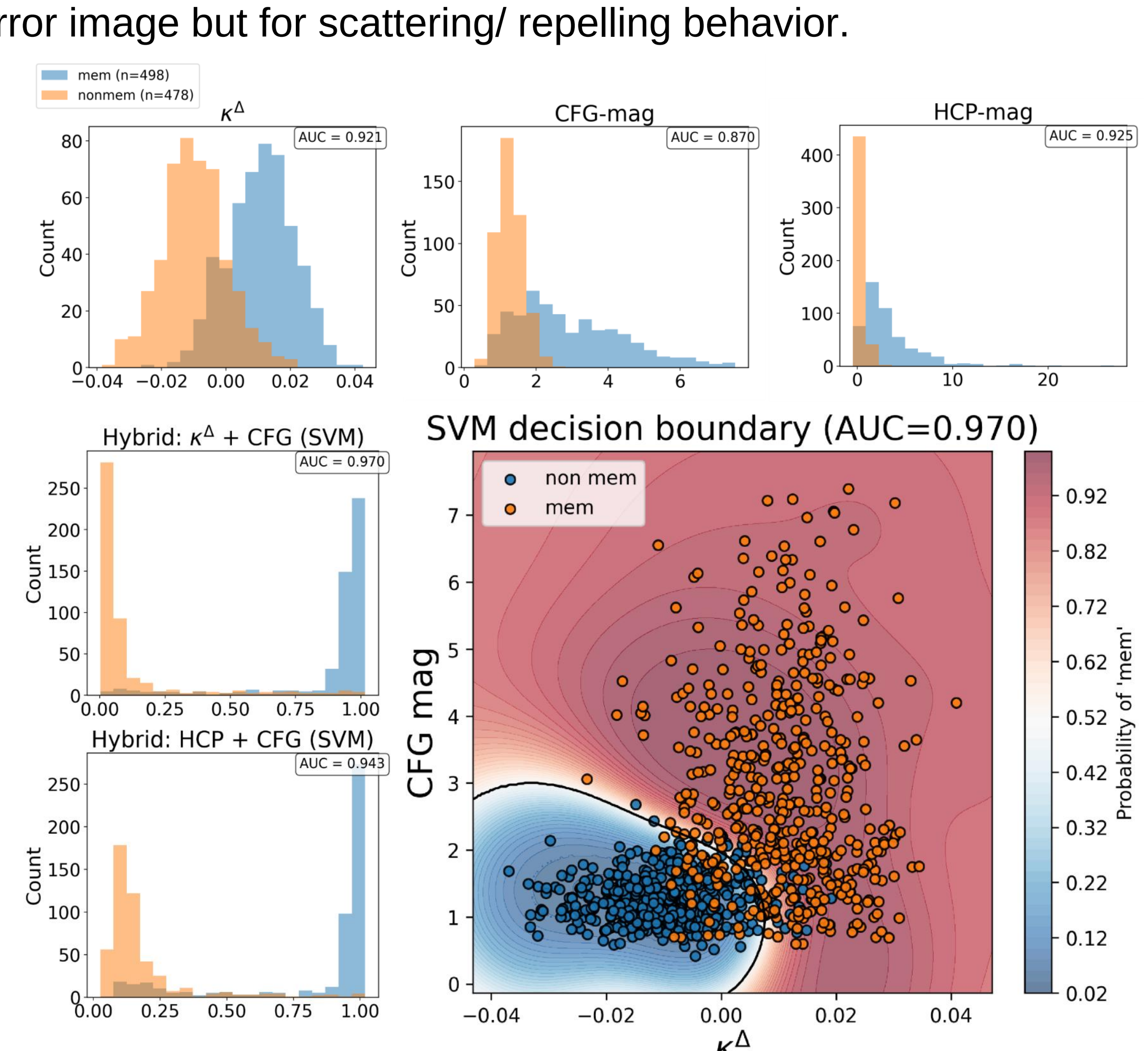
Our criterion **converges to the following mean curvature**, by gauss divergence theorem and monte-carlo estimation considerations

$$\kappa^\Delta(c_t) \to \nabla\cdot\left(\frac{\nabla\log\left(\frac{p(z_t\mid c)}{p(z_t)}\right)}{\left\|\nabla\log\left(\frac{p(z_t\mid c)}{p(z_t)}\right)\right\|}\right)$$

Findings



The geometry revealed: Memorized prompts: Zero → steep positive rise → mid-generation peak. Strong concave geometry. Non-memorized prompts: A mirror image but for scattering/ repelling behavior.



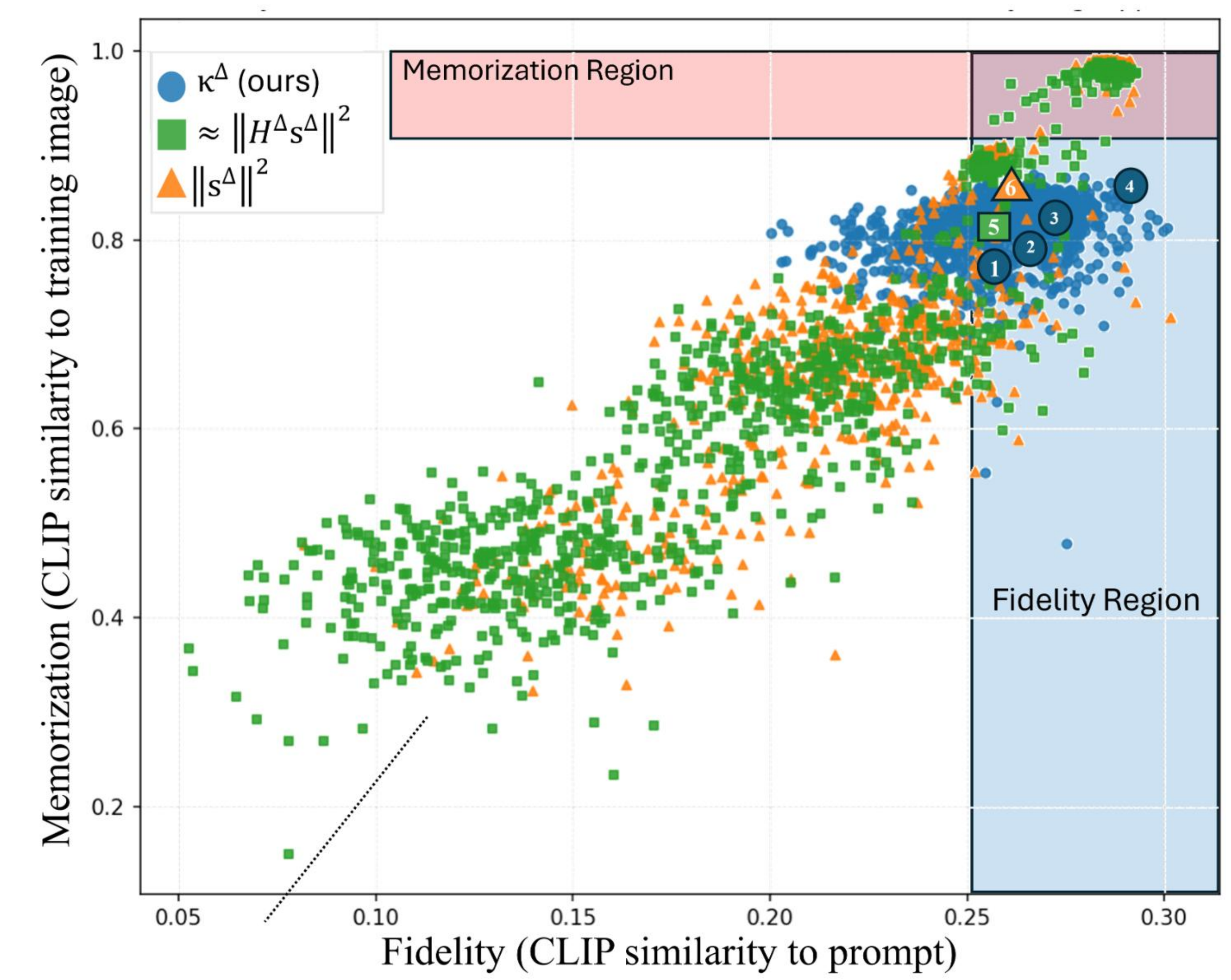
Hybrid detector, combining CFG magnitude with κ^Δ achieves SOTA AUC, showing angle information is complementary to magnitude

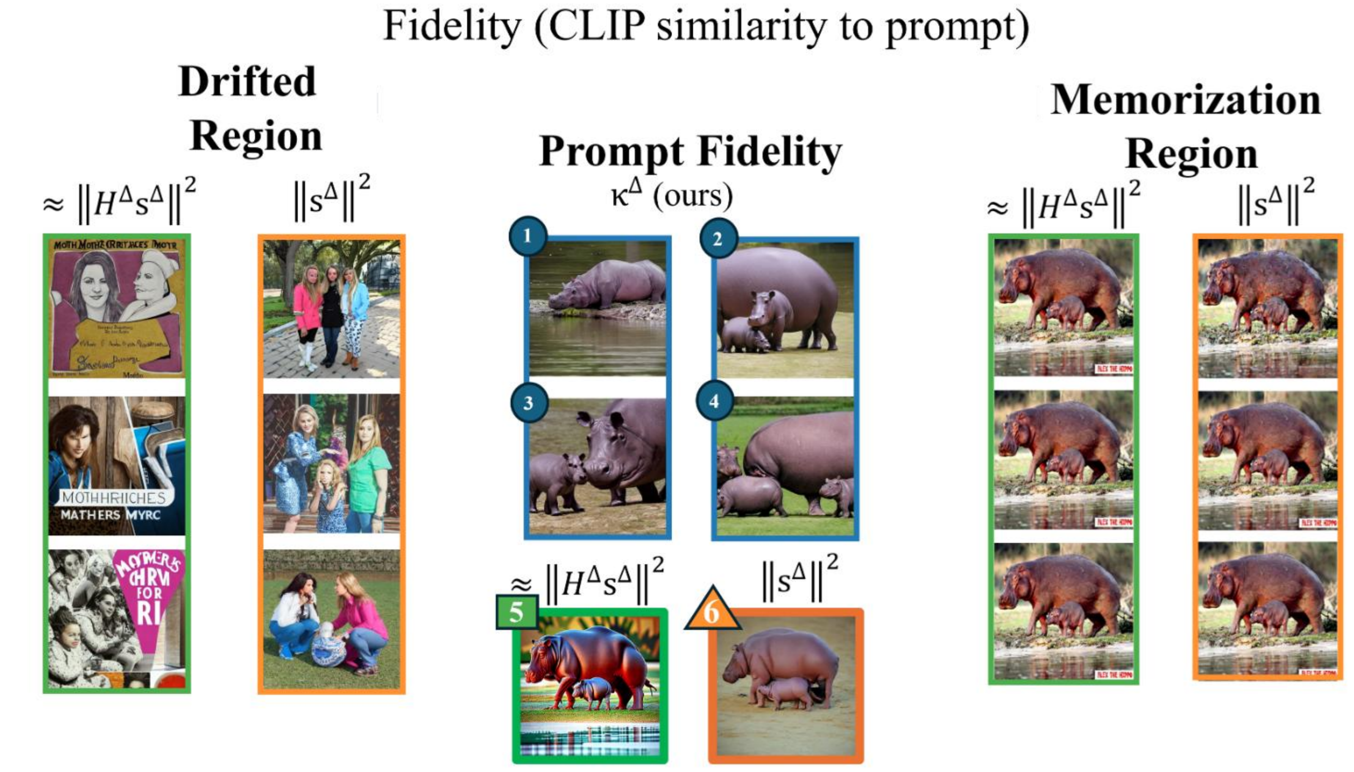
Mitigation experiment

A known mitigation framework (Wen et. al, ICLR'25) optimizes a soft prompt to reduce memorization via a loss function.

$$L_{\text{miti}} = \mathcal{L}_{\text{fid}} + \lambda\, C(z_t)$$

Where L_{fid} makes sure the prompt does not stray. We plug different criteria C into the same optimization framework to compare our κΔ to prominent competitors.





Prompt: "Mothers influence on her young hippo"

Conclusions

- Memorization yields a directional geometric signature.
- κΔ is magnitude-invariant, interpretable, and effective at the earliest step.
- Natural decision boundary: κΔ = 0.
- Hybrid detectors set new SOTA.
- κΔ enables geometry-aware mitigation improving fidelity–memorization balance.