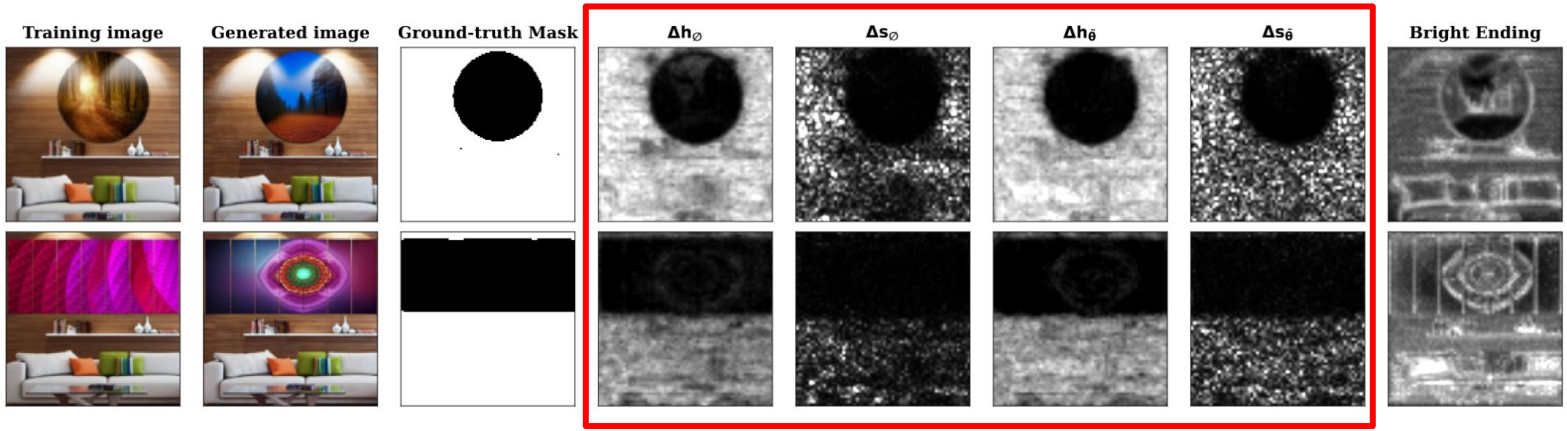
ICML
International Conference
On Machine Learning

Localizing Memorized Regions in Diffusion Models via Coordinate-Wise Curvature Differences

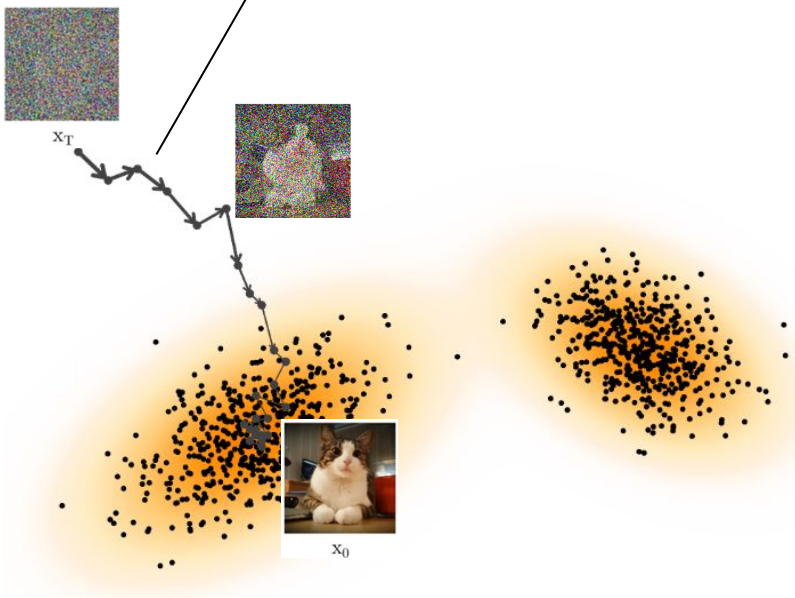
Gwangho Kim, Sungyoon Lee



Background: Diffusion Models & Memorization

score

$$s_\theta(x_t) \approx \nabla_{x_t} \log p(x_t)$$

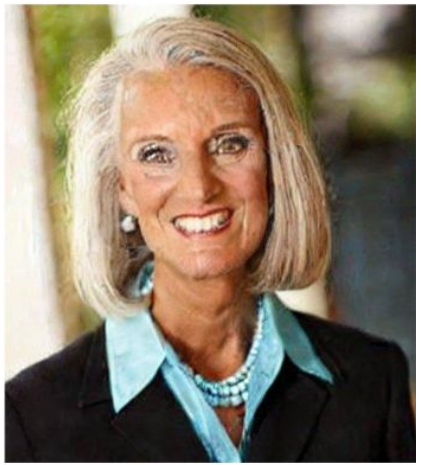


Carlini et al.

Training Set

Caption: Living in the light with Ann Graham Lotz

Generated Image

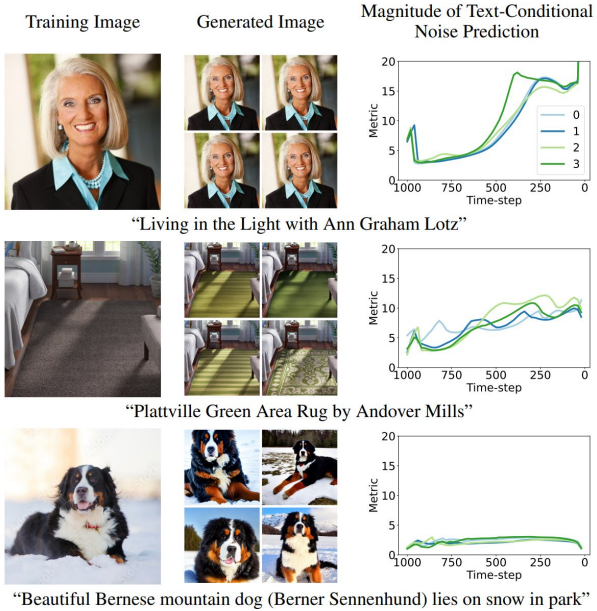


Prompt: Ann Graham Lotz

Related Works & Contributions

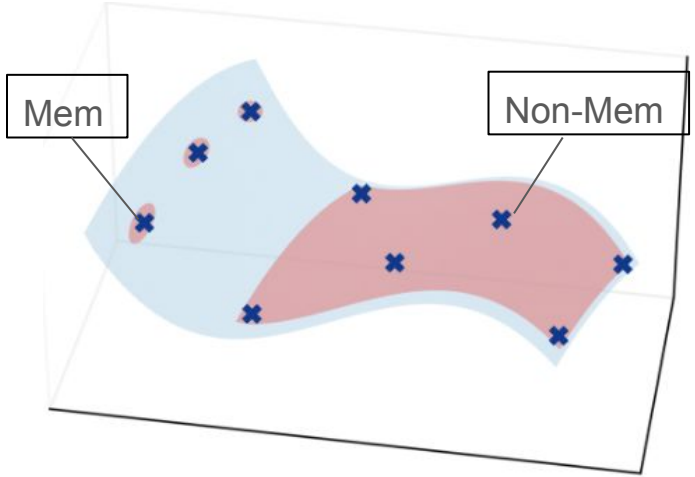
Wen et al. 2024

$$\|s_\theta^{\Delta}(\mathbf{x}_t)\| := \|s_\theta(\mathbf{x}_t, c) - s_\theta(\mathbf{x}_t)\|$$



Ross et al. 2025

Low dimensionality as signal of memorization



Jeon et al. extend this to **sharpness**, interpreting $\|s_\theta^{\Delta}(\mathbf{x}_t)\|$ as a sharpness gap.

Our Contributions

1. Geometrically localize **where** memorization occurs.
2. Explaining why the gap itself and the unconditional model matter

Motivating Example

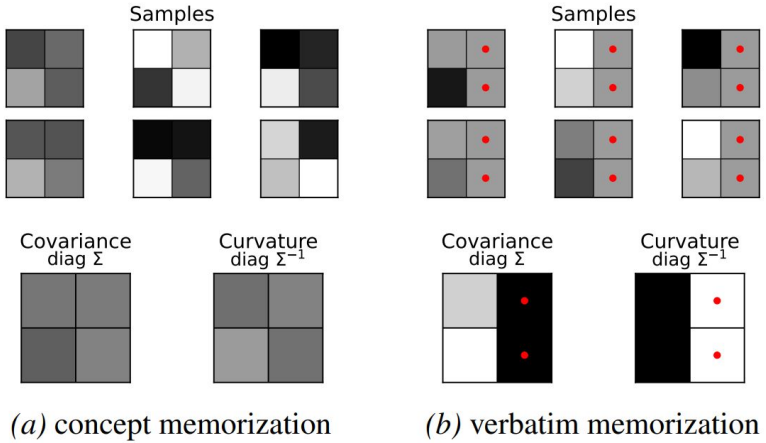


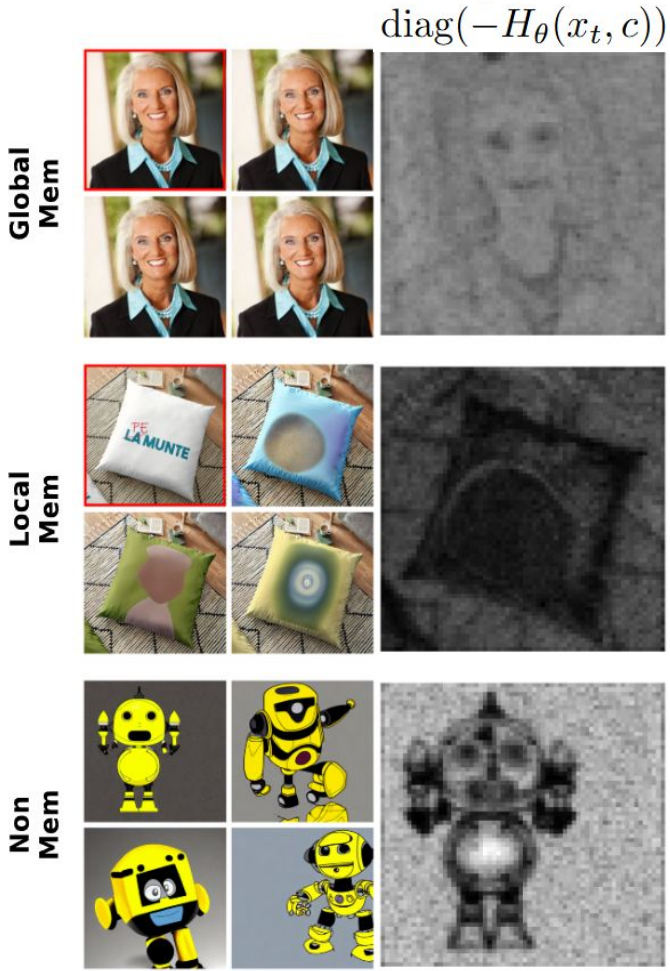
Figure 2. Samples and coordinate-wise curvature of linear Gaussian models $x = Az + \varepsilon$ with $z \sim \mathcal{N}(0, I_2)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_4)$ ($\sigma \ll 1$). Both constructions have $\mathrm{rank}(A) = 2$ and identical Frobenius norm $\|A\|_F$, and thus share the same intrinsic dimensionality.

This motivates the use of **coordinate-wise curvature** over dimensionality or sharpness.

$$H_\theta(x_t, c) := \nabla_{x_t} s_\theta(x_t, c)$$

So, how about trying $\mathrm{diag}(-H_\theta(x_t, c))$?

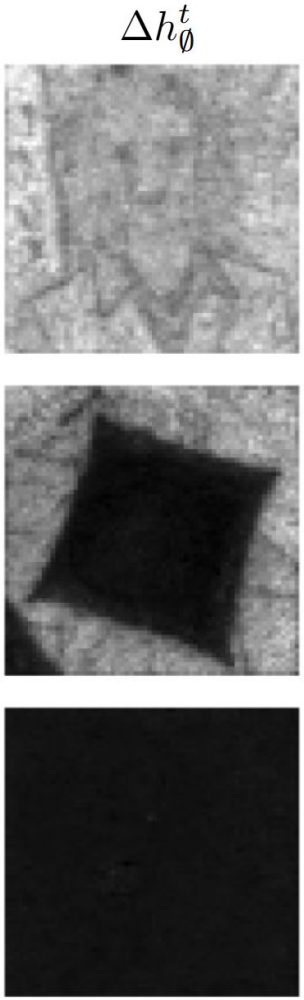
Why Gap Matters: Suppressing Intrinsic Curvature



What if **subtracting** the curvature of the **unconditional** model?

$$\Delta h_\emptyset^t := \mathrm{diag}(-H_\theta(x_t, c)) - \mathrm{diag}(-H_\theta(x_t)),$$

Subtracting the unconditional model suppresses **intrinsic data curvature**, revealing **overfitting-driven memorization**

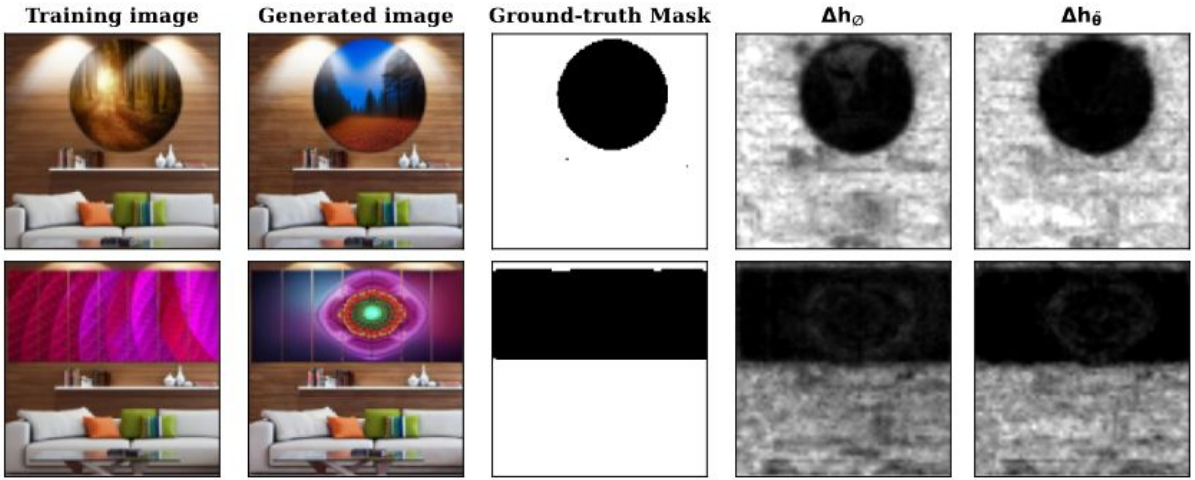


Interpretation of unconditional model as underfitted model

> “It has to generate from all classes at once, whereas conditional model can focus on a single class for any specific sample… the network attains a worse fit to the data.” (Karras et al. 2024)

Let’s try another underfitted model (less-trained version)

$$\Delta h^t_{\tilde{\theta}} := \mathrm{diag}(-H_\theta(x_t, c)) - \mathrm{diag}(-H_{\tilde{\theta}}(x_t, c))$$



Interpretation of score-difference metric $\|s_\theta(x_t, c) - s_\theta(x_t)\|_2$ Wen et al. (2024)

Now, we understand the geometric significance of the **gap** and the **unconditional** reference.

Proposition 4.2. *Let $p(c \mid x)$ be a conditional likelihood that is twice continuously differentiable with respect to x. Define the Fisher information matrix with respect to x as*

$$\mathcal{I}(x) := \mathbb{E}_{c\sim p(c|x)}\left[\nabla_x \log p(c \mid x)\, \nabla_x \log p(c \mid x)^\top\right].$$

Then $\mathcal{I}(x)$ satisfies the Fisher information identity:

$$\mathcal{I}(x) = \mathbb{E}_{c\sim p(c|x)}\left[-\nabla_x^2 \log p(c \mid x)\right]. \tag{3}$$

Moreover, by taking the diagonal terms:

$$\mathbb{E}_{c\sim p(c|x)}\left[\operatorname{diag}\left(-\nabla_x^2 \log p(x|c) + \nabla_x^2 \log p(x)\right)\right] = \mathbb{E}_{c\sim p(c|x)}\left[\left(\nabla_x \log p(x|c) - \nabla_x \log p(x)\right)^{\odot 2}\right].$$

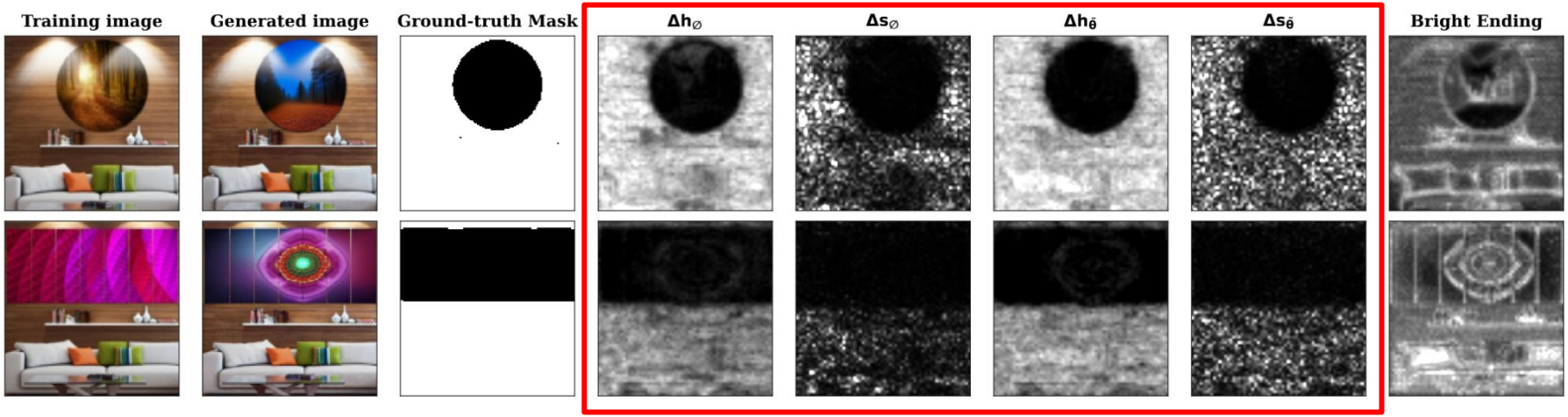
where $\odot 2$ denotes element-wise squaring.

$$\Delta s_\emptyset^t := \left(s_\theta(x_t, c) - s_\theta(x_t)\right)^{\odot 2}$$

$$\Delta s_{\tilde\theta}^t := \left(s_\theta(x_t, c) - s_{\tilde\theta}(x_t, c)\right)^{\odot 2}$$

Summary

Bright Ending : Cross-attention based method (Chen et al. 2025)



1. Accurate memorization localization using a geometric method, using ***coordinate-wise curvature differences***.

2. a novel interpretation of the score-difference metric, explaining **why subtracting the unconditional model** is important.

Thank you!

References

1. Carlini et al. (USENIX 2023) - Extracting Training Data from Diffusion Models

2. Wen et al. (ICLR 2024) - Detecting, Explaining, and Mitigating Memorization in Diffusion Models

3. Ross et al. (ICLR 2025) - A Geometric Framework for Understanding Memorization in Generative Models

4. Chen et al. (ICLR 2025) - Exploring Local Memorization in Diffusion Models via Bright Ending Attention

5. Jeon et al. (ICML 2025) - Understanding and Mitigating Memorization in Generative Models via Sharpness of Probability Landscapes

6. Karras et al. (Neurips 2024) - Guiding a Diffusion Model with a Bad Version of Itself