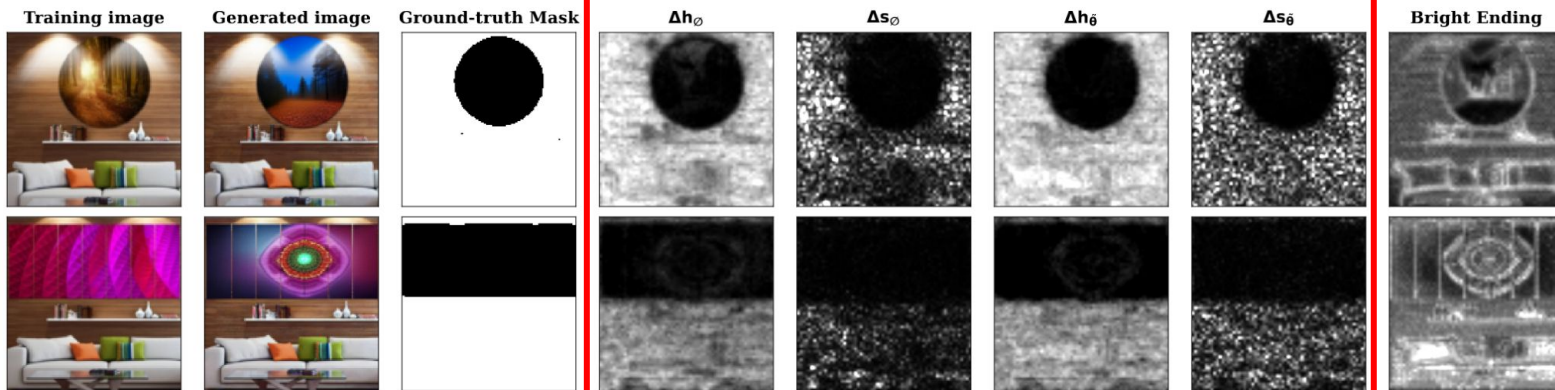


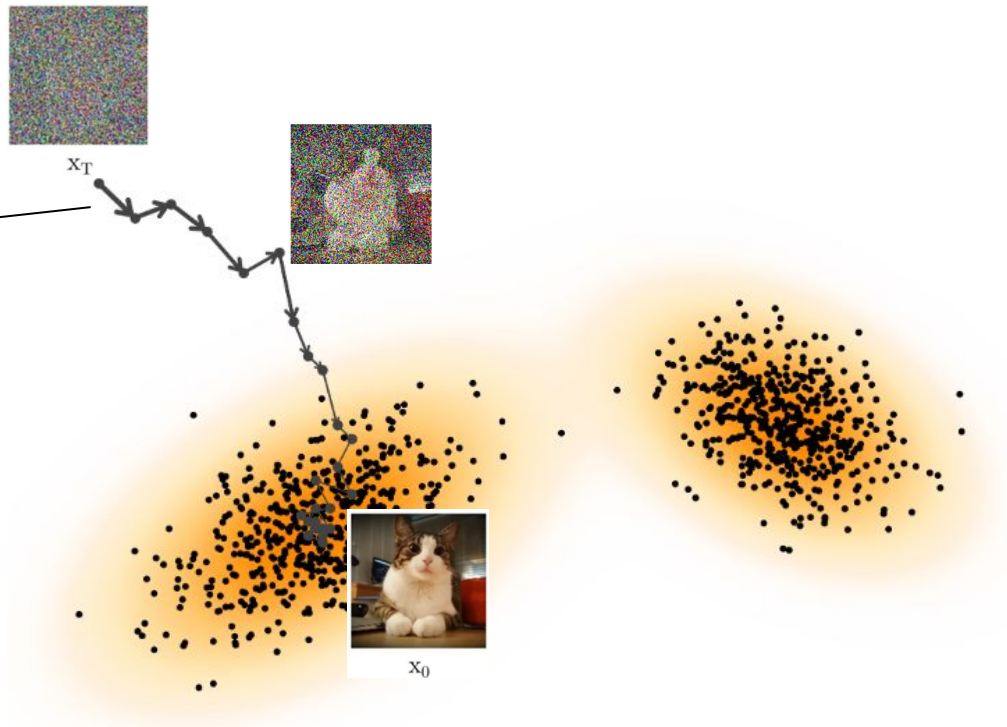
Localizing Memorized Regions in Diffusion Models via Coordinate-Wise Curvature Differences

Gwangho Kim, Sungyoon Lee



Background: Diffusion Models

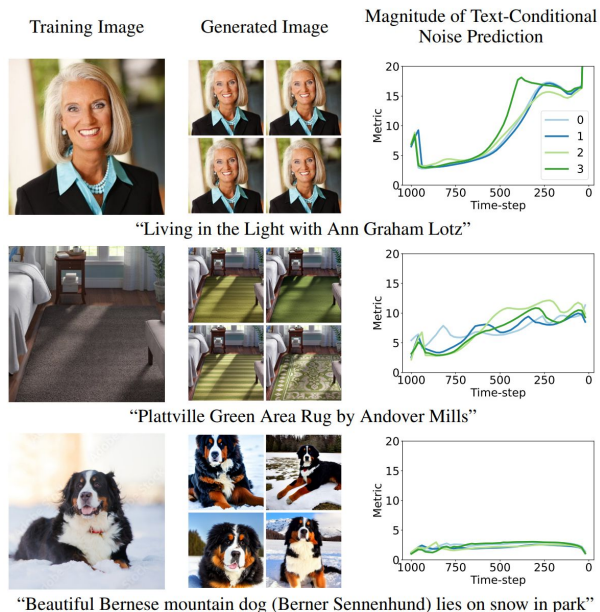
$$s_{\theta}(x_t) \approx \overset{\text{score}}{\nabla_{x_t} \log p(x_t)}$$



Related Works & Contributions

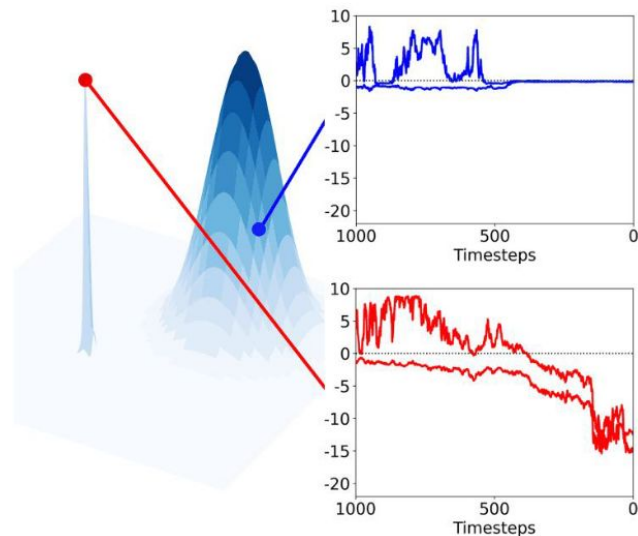
Wen et al. 2024

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



Jeon et al. 2025

Sharpness $\text{tr}(\nabla_{x_t}^2 \log p(x_t))$ as memorization signal



Interpret $\|s_{\theta}^{\Delta}(\mathbf{x}_t)\|$ as sharpness gap

Our Contributions

1. Moving beyond global metrics to identify **where** memorization occurs.
2. Explaining why the **gap** and the **unconditional** model matter

Motivating Example

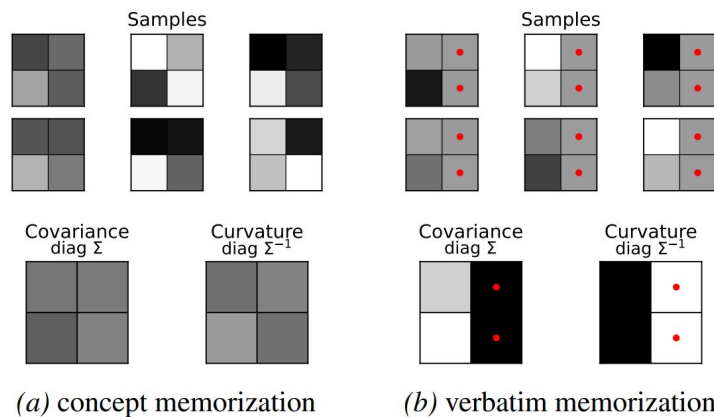


Figure 2. Samples and coordinate-wise curvature of linear Gaussian models $x = Az + \varepsilon$ with $z \sim \mathcal{N}(0, I_2)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_4)$ ($\sigma \ll 1$). Both constructions have $\text{rank}(A) = 2$ and identical Frobenius norm $\|A\|_F$, and thus share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

For Gaussian,

$$\log p(x) = -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) + C$$

$$\nabla_x \log p(x) = -\Sigma^{-1}(x - \mu)$$

$$\nabla_x^2 \log p(x) = -\Sigma^{-1}$$

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

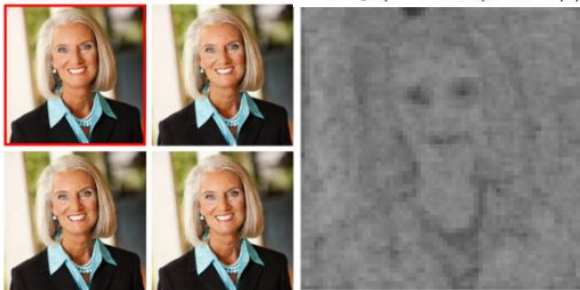
$$H_\theta(x_t, c) := \nabla_{x_t} s_\theta(x_t, c)$$

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

Why Gap Matters: Suppressing Intrinsic Curvature

$$\text{diag}(-H_{\theta}(x_t, c))$$

Global
Mem

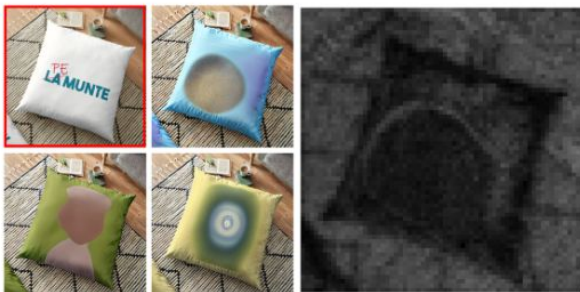


What if **subtracting** the curvature of the **unconditional** model?

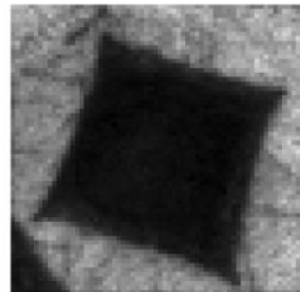
$$\Delta h_{\emptyset}^t$$



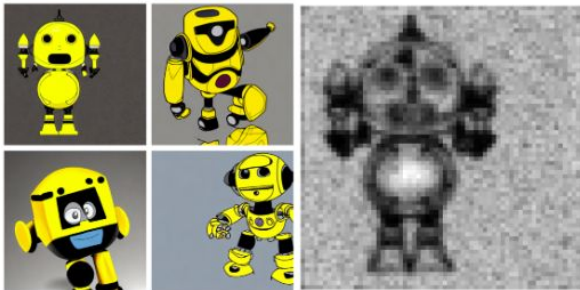
Local
Mem



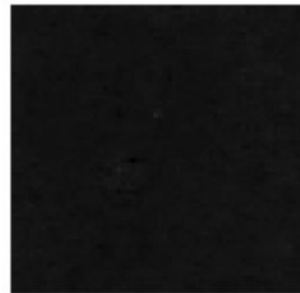
$$\Delta h_{\emptyset}^t := \text{diag}(-H_{\theta}(x_t, c)) - \text{diag}(-H_{\theta}(x_t)),$$



Non
Mem



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

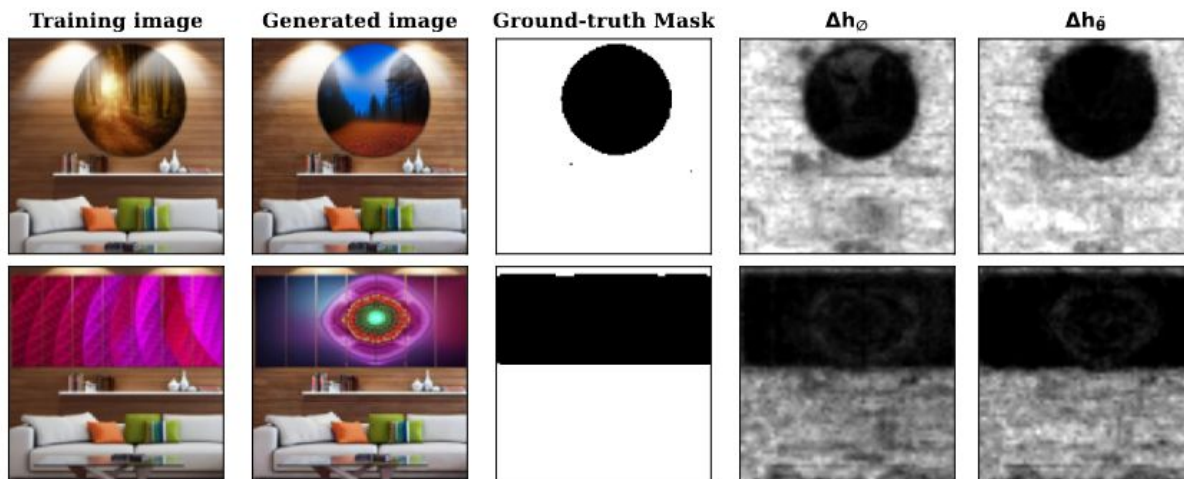


Interpretation of unconditional model as underfitted model

“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)

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



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*

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

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

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

Moreover, by taking the diagonal terms:

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

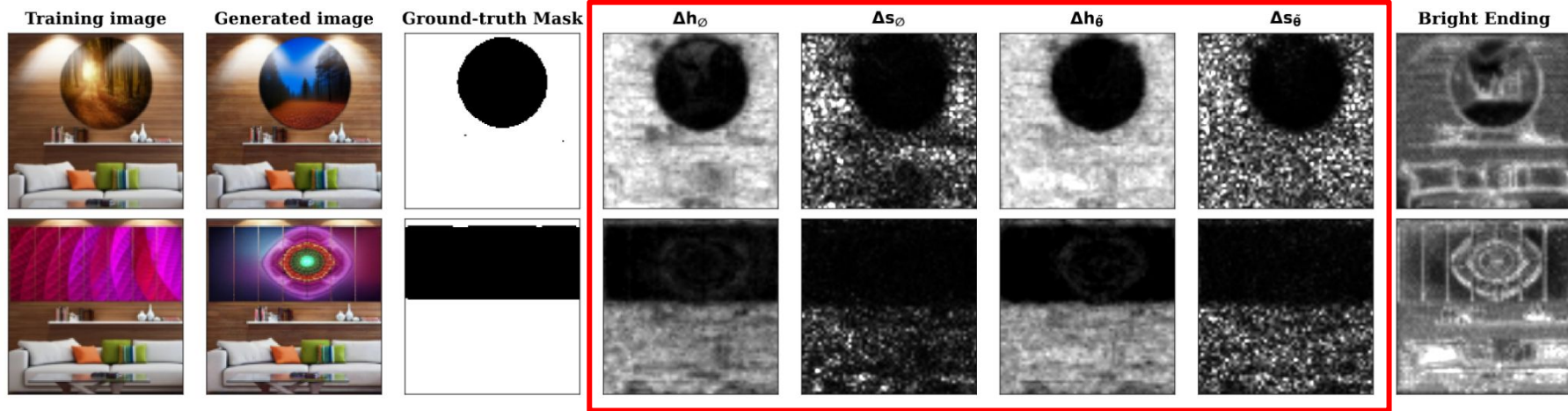
where $\odot 2$ denotes element-wise squaring.

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

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

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. Wen et al. (ICLR 2024) - [Detecting, Explaining, and Mitigating Memorization in Diffusion Models](#)
2. Chen et al. (ICLR 2025) - [Exploring Local Memorization in Diffusion Models via Bright Ending Attention](#)
3. Jeon et al. (ICML 2025) - [Understanding and Mitigating Memorization in Generative Models via Sharpness of Probability Landscapes](#)
4. Karras et al. (Neurips 2024) - [Guiding a Diffusion Model with a Bad Version of Itself](#)