



# Understanding Memorization in Generative Models via Sharpness in Probability Landscapes

Dongjae Jeon\*, Dueun Kim\*, Albert No

# Preliminary

## ■ Generative Modeling:

Sampling from data density  $p(\mathbf{x})$

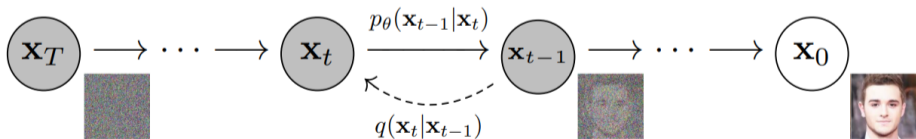
Directly constructing  $p(\mathbf{x})$  is hard!

Diffusion models construct log gradients,  $r_{\mathbf{x}} \log p(\mathbf{x})$

## ■ Denoising Diffusion (DDPM):

### Denoising Process:

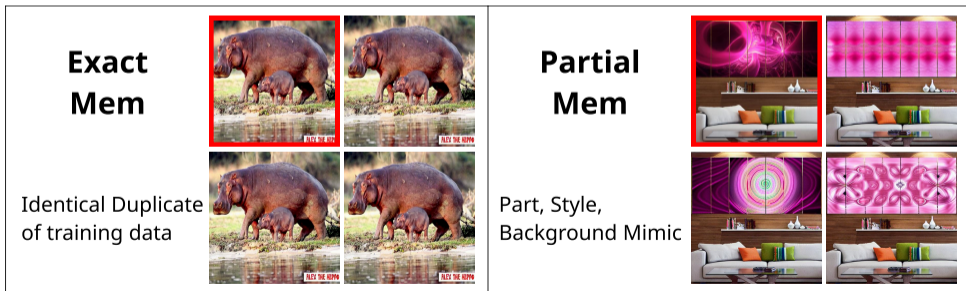
- Gradually correct  $p(\mathbf{x}_t)$  from Gaussian noise using  $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$



<sup>0</sup>Ho et al, "Denoising diffusion probabilistic models.", *NeurIPS*, 2020.

# What is Memorization in Diffusion Models?

- **Definition:** A phenomenon in which a model nearly replicates training data.
- **Risks:**
  - Copyright & Privacy issues
  - Degradation in utility
- **Memorization Categories:**
  - Training data in **Red** outline



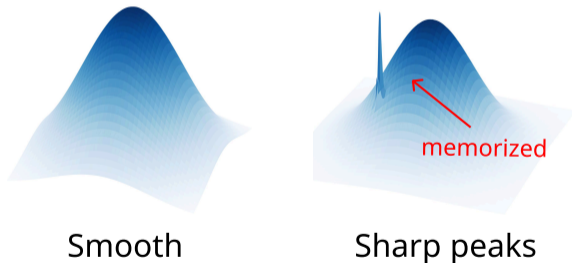
# Memorization in Probability Density Perspective

## ■ Geometric view of Memorization

**Local Intrinsic Dimensionality (LID):** Exact Memorization / 0 dimensionality<sup>1</sup>

**Probability Density (Ours):** Sharp peaks in distribution

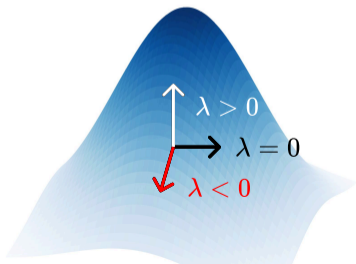
\*Enable analyzing entire denosing timesteps



<sup>1</sup>Ross et al. "A geometric framework for understanding memorization in generative models." *ICLR*. 2025.

# Sharpness interpreted via Hessian Eigenvalues

- **Score Function:**  $s_{\theta}(\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$
- **Jacobian of Score Function (Hessian):**  $H_{\theta}(\mathbf{x}_t) = \nabla_{\mathbf{x}_t}^2 \log p_t(\mathbf{x}_t)$
- **Conditional case:**  $s_{\theta}(\mathbf{x}_t; c); H_{\theta}(\mathbf{x}_t; c)$
- Hessian Eigenvalues tell Curvature:
  - $\lambda = 0$ : Concave downward or Flat
  - $\lambda < 0$ : Concave upward (Key for finding peaks)

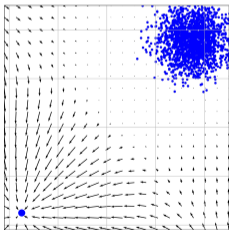


# Eigenvalue Analysis in Toy Data

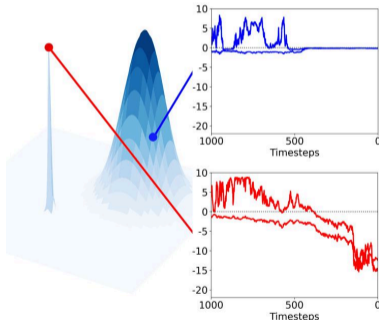
## ■ 2D Gaussian:

Duplicated single data point for a sharp peak

Sharp peak shows large negative  $\lambda$  over timesteps



(a) Learned Scores



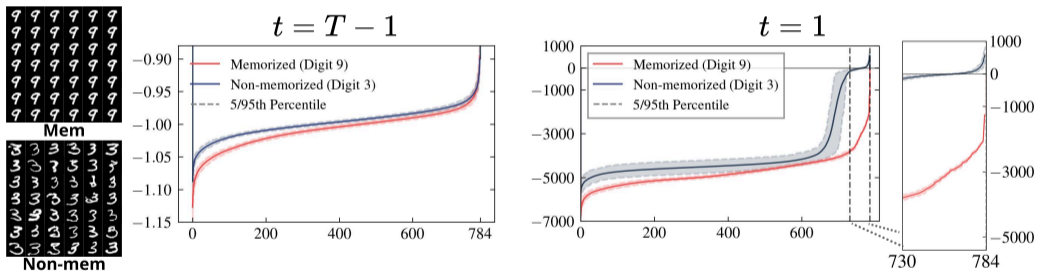
(b) Eigenvalues over Timestep

# Eigenvalue Analysis in Toy Data (Cont'd)

## ■ MNIST:

Digit 3 for Non-mem, digit 9 for Mem

Memorized samples consistently show large negative  $\lambda$  even at the initial step



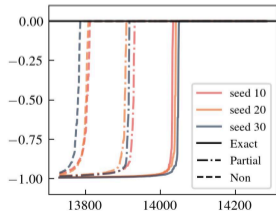
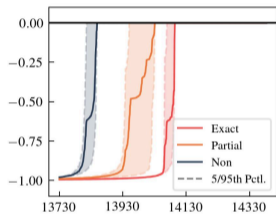
(c) Eigenvalues on initial ( $t=T-1$ ) and last ( $t=1$ ) sampling step.

# Eigenvalue Analysis in Stable Diffusion (SD)

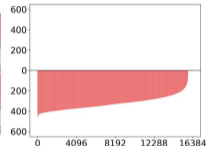
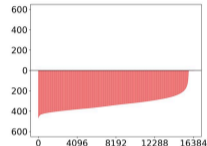
- Similar phenomenon in Stable Diffusion with 16,384 dimension

$t = T - 1$

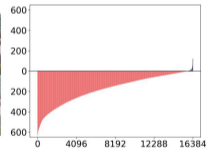
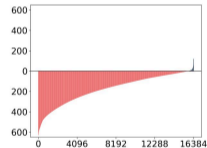
$t = 1$



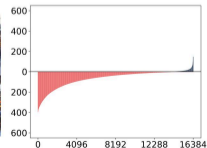
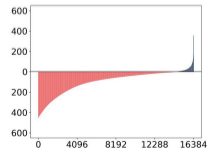
**Exact Mem**



**Partial Mem**



**Non Mem**



# Eigenvalue Statistics for Efficient Detection

- Under Gaussian,

$$E[ks(\mathbf{x})k^2] = \text{tr}(H(\mathbf{x})) = \text{Negative Sum of Eigenvalues}$$

$$E[kH(\mathbf{x})s(\mathbf{x})k^2] = \text{tr}(H(\mathbf{x})^3) = \text{Negative Sum of **Cubic** Eigenvalues}$$

Memorized ! large negative sum (magnitude " )

- **Explain Wen's SOTA Detection Metric<sup>2</sup>:**

$$kS_{\theta}(\mathbf{x}_t)k_{\text{avg}} = \frac{1}{T} \sum_{t=T} kS_{\theta}(\mathbf{x}_t; c) - S_{\theta}(\mathbf{x}_t)k$$

Sharpness difference between  $\log p_t(\mathbf{x}_t; c)$  and  $\log p_t(\mathbf{x}_t)$

- **Our enhanced metric:**

$$kH_{\theta}(\mathbf{x}_t) - S_{\theta}(\mathbf{x}_t)k$$

---

<sup>2</sup>Wen, Yuxin, et al. "Detecting, explaining, and mitigating memorization in diffusion models.", *ICLR*. 2024.

# Our Mitigation Strategy: SAIL

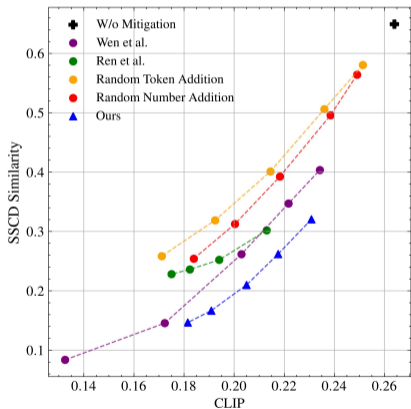
- Existing mitigation strategies modify prompts (or text embeddings)  
Degraded generation quality / user purpose
- Our strategy **SAIL** (Sharpness-Aware Initialization for Latent diffusion)  
Idea: Optimize the initial noise  $\mathbf{x}_T$  to lie on smoother regions.  
Objective function:

$$L_{\text{SAIL}}(\mathbf{x}_T) = \underbrace{\|H_\theta(\mathbf{x}_T)\|_{S_\theta(\mathbf{x}_T)}^2}_{\text{Sharpness measure}} + \underbrace{\alpha \|\mathbf{x}_T\|^2}_{\text{Gaussian regularization}}$$

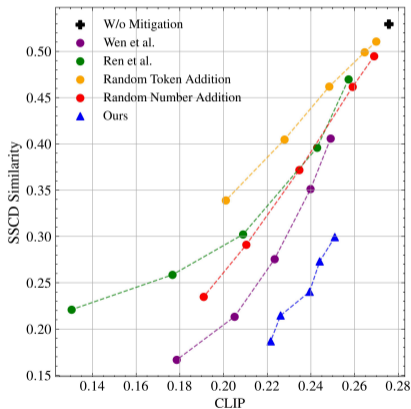
No extra modifications on model, only the initial noise  $\mathbf{x}_T$  changed.

# Quantitative Result of SAIL

- SAIL achieves superior performance
- Low similarity scores & High CLIP scores (better prompt-img alignment)



**SD v1.4**



**SD v2.0**

# Qualitative Result of SAIL

- SAIL protects key details in prompts while others fail

