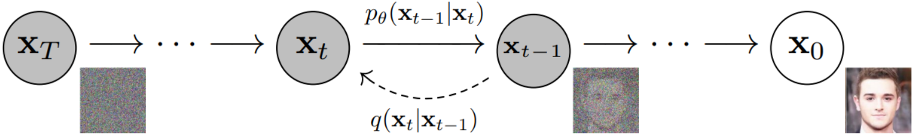
Understanding Memorization in Generative Models via Sharpness in Probability Landscapes

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, $\nabla_{\mathbf{x}} \log p(\mathbf{x})$

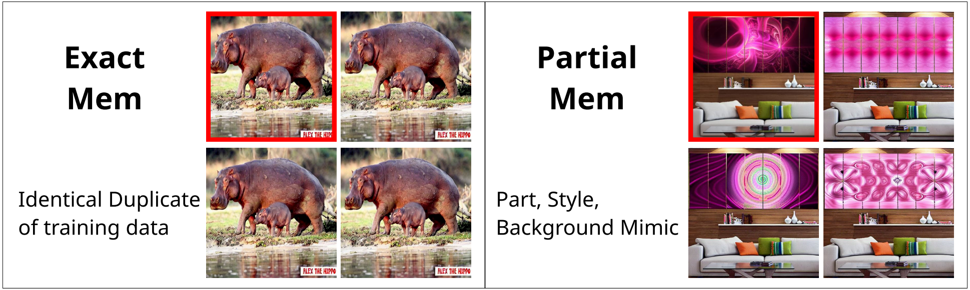
- **Denoising Diffusion (DDPM):**
  - **Denoising Process:**
    - Gradually correct $p(\mathbf{x}_t)$ from Gaussian noise using $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$



[0] Ho et al, "Denoising diffusion probabilistic models.", *NeurIPS*, 2020.

What is Memorization in Diffusion Models?

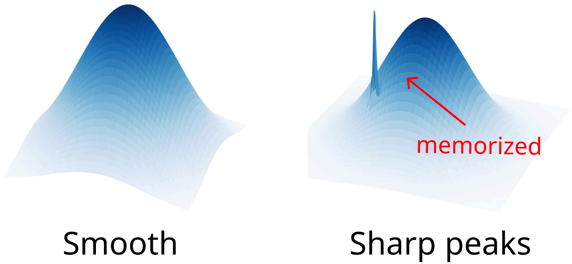
- **Definition:** A phenomenon in which a model nearly replicates training data.
- **Risks:**
  - Copyright & Privacy issues
  - Degradation in utility
- **Memorization Categories:**
  - Training data in Red outline



Memorization in Probability Density Perspective

Geometric view of Memorization

- **Local Intrinsic Dimensionality (LID)**: Exact Memorization $\rightarrow$ 0 dimensionality[1]
- **Probability Density (Ours)**: Sharp peaks in distribution
  *Enable analyzing entire denosing timesteps

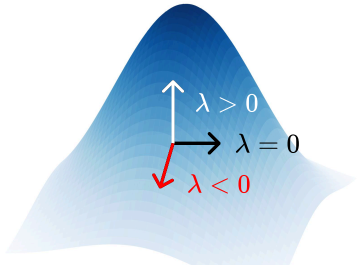


[1] Ross et al. "A geometric framework for understanding memorization in generative models." *ICLR*. 2025.

Sharpness interpreted via Hessian Eigenvalues

- **Score Function:** $s_\theta(\mathbf{x}_t) \approx \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$
- **Jacobian of Score Function (Hessian):** $H_\theta(\mathbf{x}_t) \approx \nabla^2_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$
- **Conditional case:** $s_\theta(\mathbf{x}_t, c), H_\theta(\mathbf{x}_t, c)$

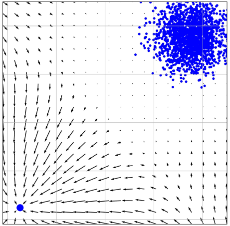
- Hessian Eigenvalues tell Curvature:
  - $\lambda \geq 0$: Concave downward or Flat
  - $\lambda < 0$: Concave upward (Key for finding peaks)



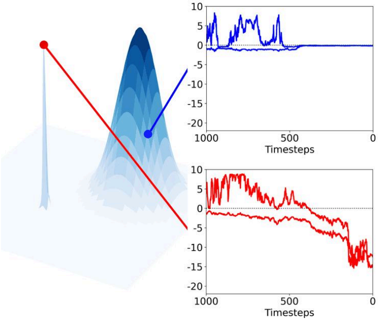
Eigenvalue Analysis in Toy Data

2D Gaussian:

- Duplicated single data point for a sharp peak
- Sharp peak shows large negative λ over timesteps



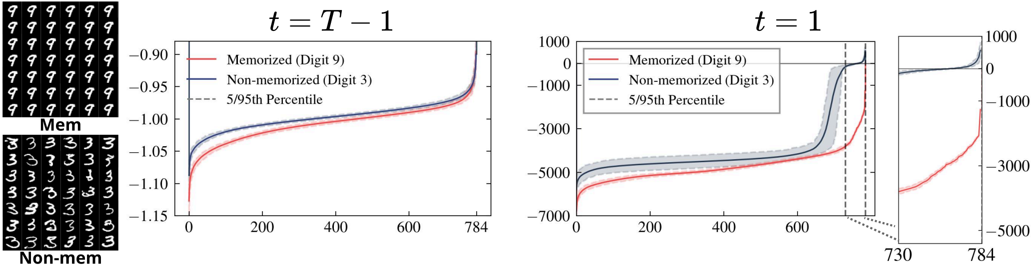
(a) Learned Scores



(b) Eigenvalues over Timestep

Eigenvalue Analysis in Toy Data (Cont'd)

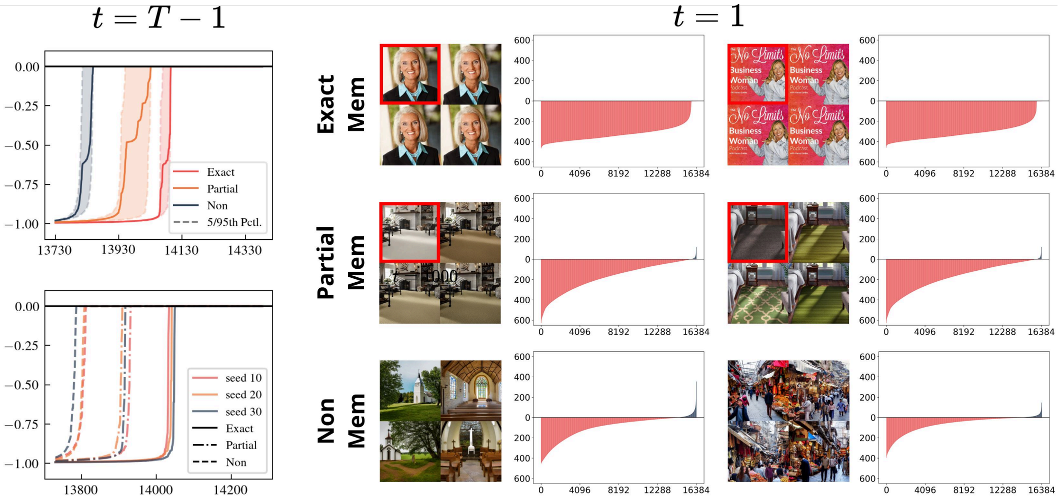
- **MNIST:**
  - Digit 3 for Non-mem, digit 9 for Mem
  - Memorized samples consistently show large negative λ even at the initial step



(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



Eigenvalue Statistics for Efficient Detection

- Under Gaussian,

$$E[\|s(\mathbf{x})\|^2] = -\operatorname{tr}(H(\mathbf{x})) = \text{Negative Sum of Eigenvalues}$$
$$E[\|H(\mathbf{x})s(\mathbf{x})\|^2] = -\operatorname{tr}(H(\mathbf{x})^3) = \text{Negative Sum of } \textbf{Cubic} \text{ Eigenvalues}$$

  - Memorized $\rightarrow$ large negative sum (magnitude $\uparrow$)

- **Explain Wen's SOTA Detection Metric[2]:**

$$\|s_\theta^\Delta(\mathbf{x}_t)\|_{\text{avg}} = \frac{1}{T}\sum_{t=T}^{1}\|s_\theta(\mathbf{x}_t, c) - s_\theta(\mathbf{x}_t)\|$$

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

- **Our enhanced metric:**

$$\|H_\theta^\Delta(\mathbf{x}_t)\, s_\theta^\Delta(\mathbf{x}_t)\|$$

[2] Wen, Yuxin, et al. "Detecting, explaining, and mitigating memorization in diffusion models.", *ICLR*. 2024.

Our Mitigation Strategy: SAIL

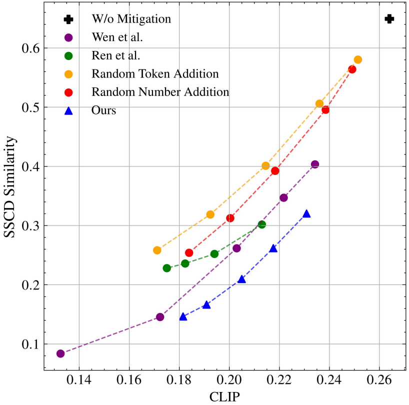
- Existing mitigation strategies modify prompts (or text embeddings)
  - Degraded generation quality / user purpose
- Our strategy **SAIL** (Sharpness-Aware Initialization for Latent diffusion)
  - Idea: Optimize the initial noise $\mathbf{x}_T$ to lie on smoother regions.
  - Objective function:

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

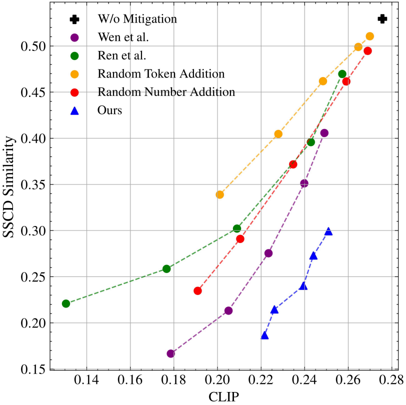
  - No extra modifications on model, only the initial noise $\mathbf{x}_T$ changed.

Quantitative Result of SAIL

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



SD v1.4



SD v2.0

Qualitative Result of SAIL

- SAIL protects key details in prompts while others fail