



**ICLR**  
International Conference On  
Learning Representations

# Exploring Local Memorization in Diffusion Models via Bright Ending Attention

Chen Chen, Daochang Liu, Mubarak Shah, Chang Xu

ICLR 2025 Spotlight



THE UNIVERSITY OF  
**SYDNEY**



THE UNIVERSITY OF  
**WESTERN  
AUSTRALIA**



# Memorization in Diffusion Models

- Pretrained diffusion models can memorize and repeat training data during inference without informing data owners and model users.

Training Image

Generated Image



# Memorization in Diffusion Models

- Pretrained diffusion models can memorize and repeat training data during inference without informing data owners and model users.
- This exposes potential violations of copyright laws and the introduction of ethical dilemmas.
- Two factors have heightened such litigation risks:
  - The widespread use and deployment of open-source state-of-the-art diffusion models.
  - The extensive size of training sets impedes detailed human review.



# Local Memorization in Diffusion Models

- Diffusion models can exhibit both global and local memorization.
  - Global memorization: the entire training image is memorized.
  - Local memorization: only parts of the training image are memorized.

Training Image

Generated Image



Global Memorization

Local Memorization

# Performance Gap in Local Memorization

- Existing methods underperform in local memorization's evaluation, detection, and mitigation.

# Performance Gap in Local Memorization – Evaluation Strategy

- Existing evaluation strategies rely on Self-Supervised Copy Detection (SSCD) embeddings.

$$Sim = cosine\_similarity(\Phi_{SSCD}(\hat{x}), \Phi_{SSCD}(x))$$

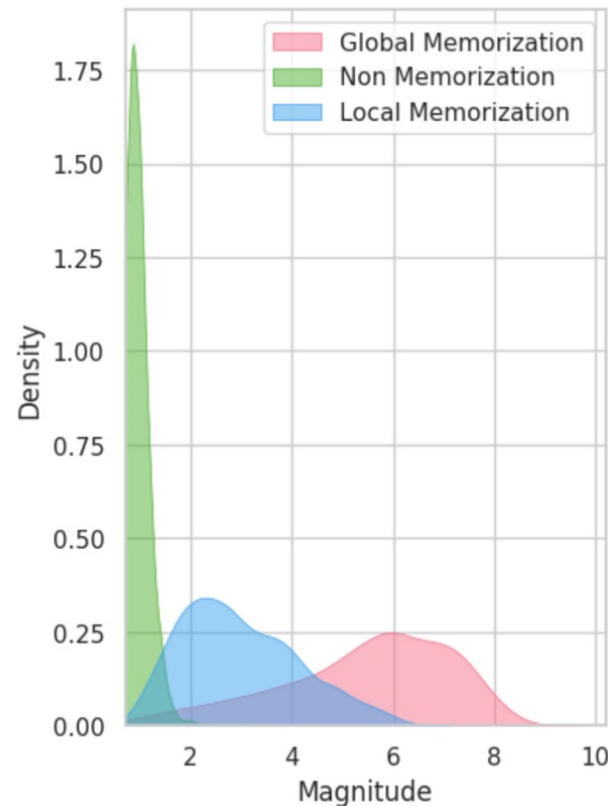
$$Memorization_{SSCD}(\hat{x}, x) = 1_{Sim > 0.5}$$



# Performance Gap in Local Memorization – Detection Strategy

- A popular detection strategy has developed “Magnitude” as a strong signal of memorization.

$$\text{Magnitude} = \|\varepsilon_{\theta}(x_t, e_p) - \varepsilon_{\theta}(x_t, e_{\phi})\|_2$$

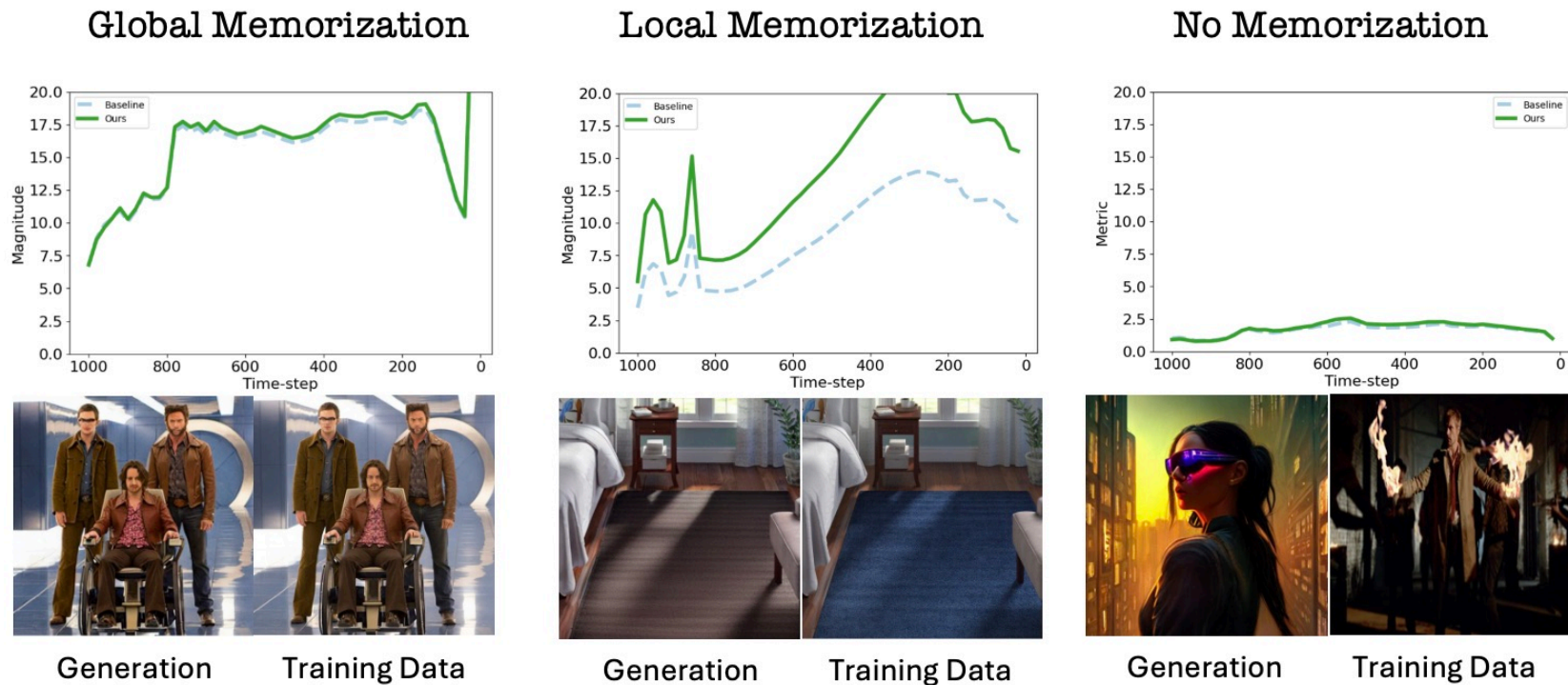




# Performance Gap in Local Memorization – Detection Strategy

- A popular detection strategy has developed “Magnitude” as a strong signal of memorization.

$$\text{Magnitude} = \|\varepsilon_{\theta}(x_t, e_p) - \varepsilon_{\theta}(x_t, e_{\phi})\|_2$$





# Performance Gap in Local Memorization – Mitigation Strategy

- The corresponding mitigation strategy relies on the magnitude-based detection strategy:

$$Magnitude = \|\varepsilon_{\theta}(x_t, e_p) - \varepsilon_{\theta}(x_t, e_{\phi})\|_2$$

- Prompt Engineering – Optimize the embedding of the user input prompt using the detection signal as the loss for computing the gradient.
- Triggered only when the detection signal is greater than a specified threshold.

# Takeaway: The Localization Insight

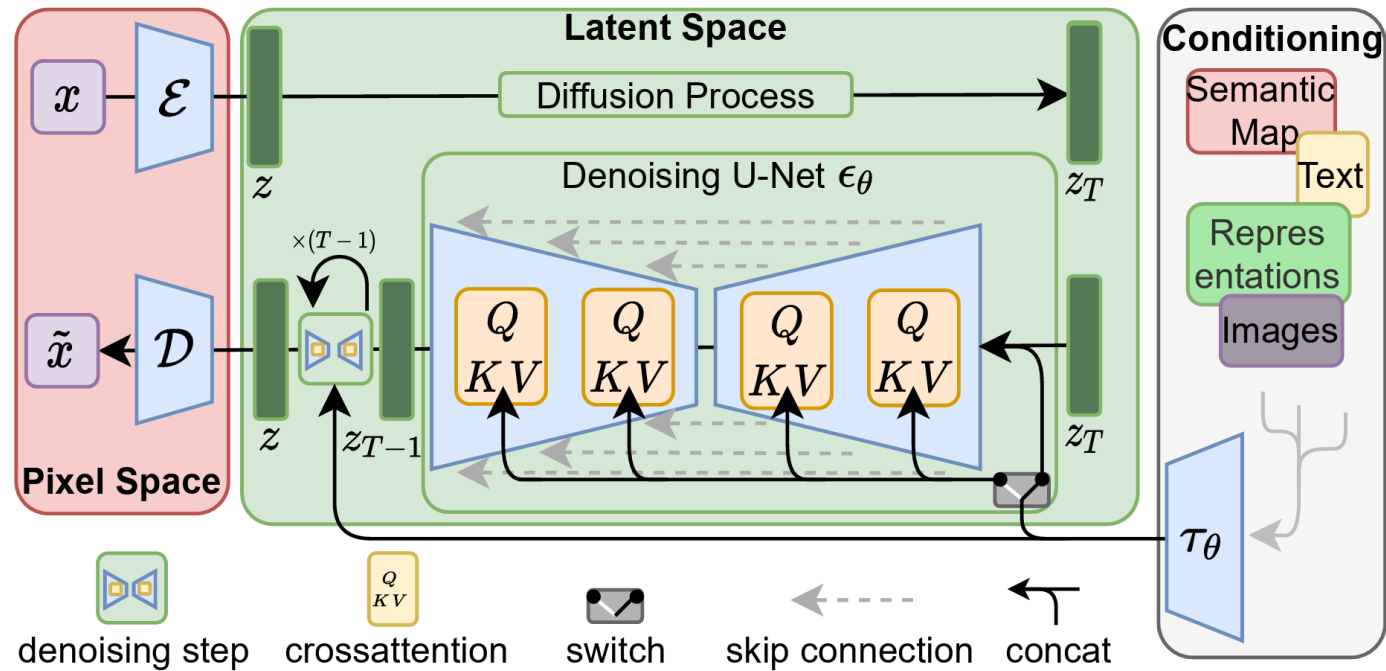
- Local memorization is a more generalized and practical notion of memorization.
  - Unmemorized regions pose no litigation risk and can be disregarded, while even a tiny locally memorized area can present significant legal concerns.
  - Local memorization is a more encompassing definition, with global memorization being a specific instance.
- Therefore, improved metrics and strategies should concentrate exclusively on the locally memorized regions while ignoring the unmemorized parts of the image.

# Next Steps: How to Investigate Locally?

- Extracting a local memorization mask can help revise existing strategies to take a local perspective.
- How to extract such a mask?
  - Directly comparing generated images with training images.
    - Compromises privacy.
    - Computationally heavy.
  - Automatic mask extraction using the pre-trained model's memory.
    - Leverage text-conditioning.
    - Explore cross-attention maps.

# The “Bright Ending” (BE) Phenomenon

- Cross-attention is used in 16 layers of Stable Diffusion’s U-Net to integrate text and other conditioning signals.

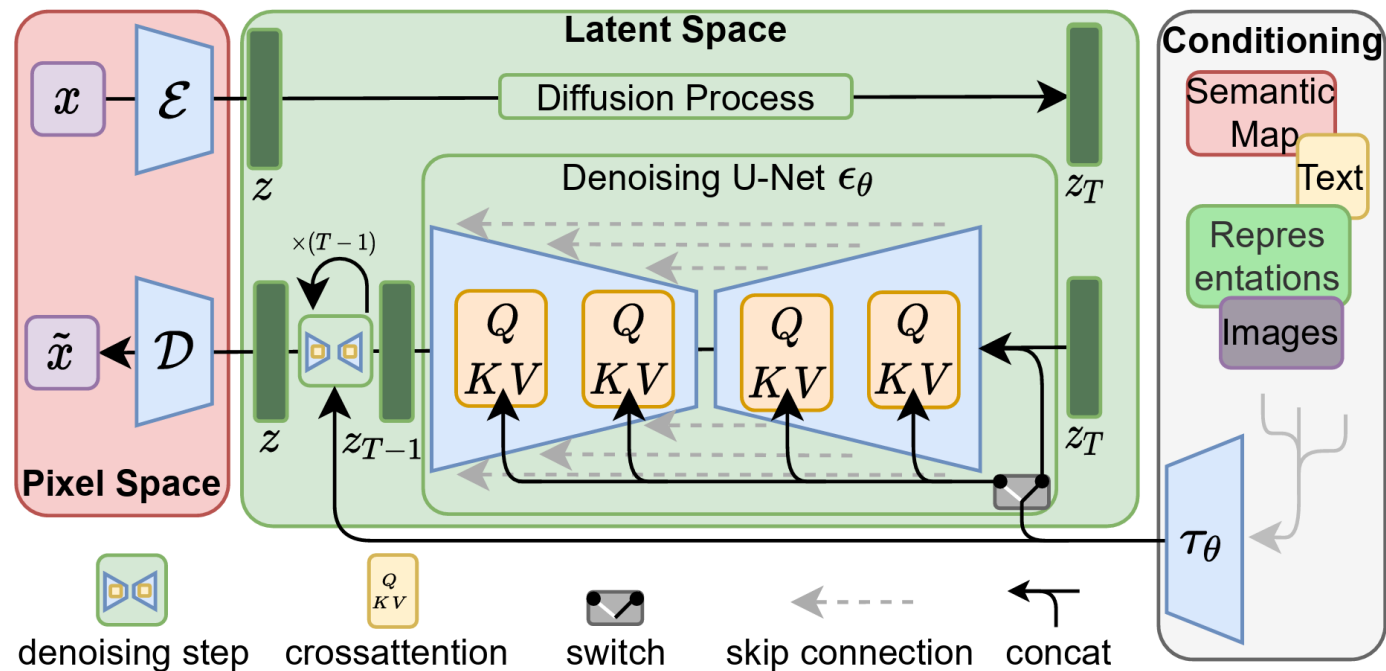




# The “Bright Ending” (BE) Phenomenon

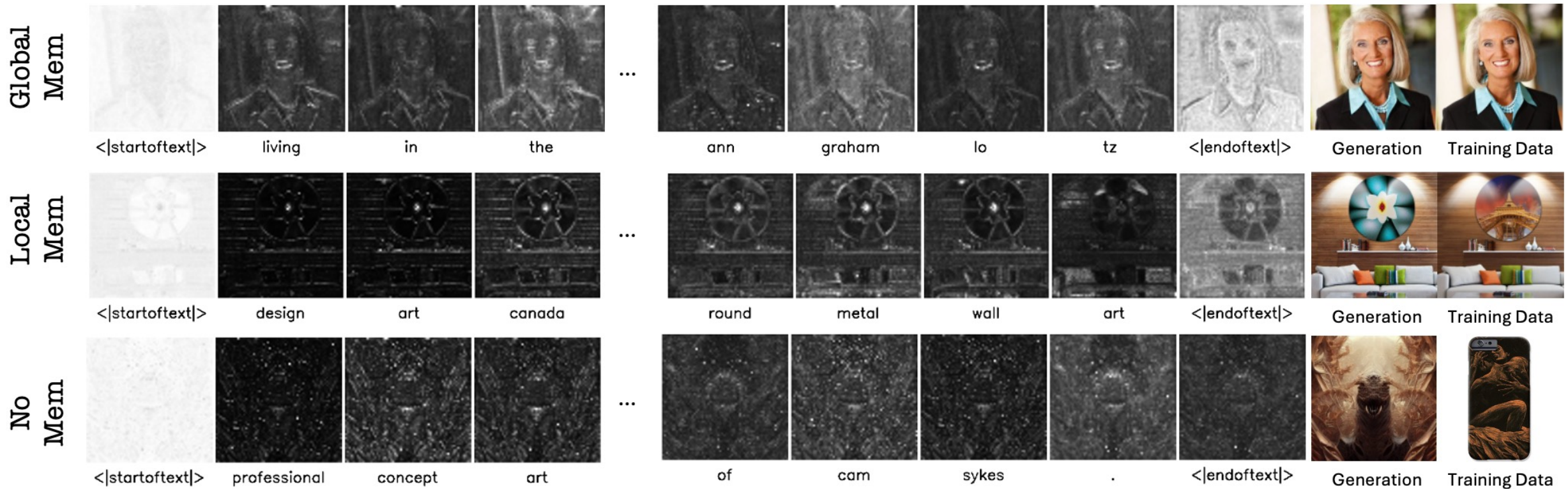
- We derive the cross-attention maps for text-to-image generations:

$$\text{AttentionMap}(Q, K) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}}\right)$$



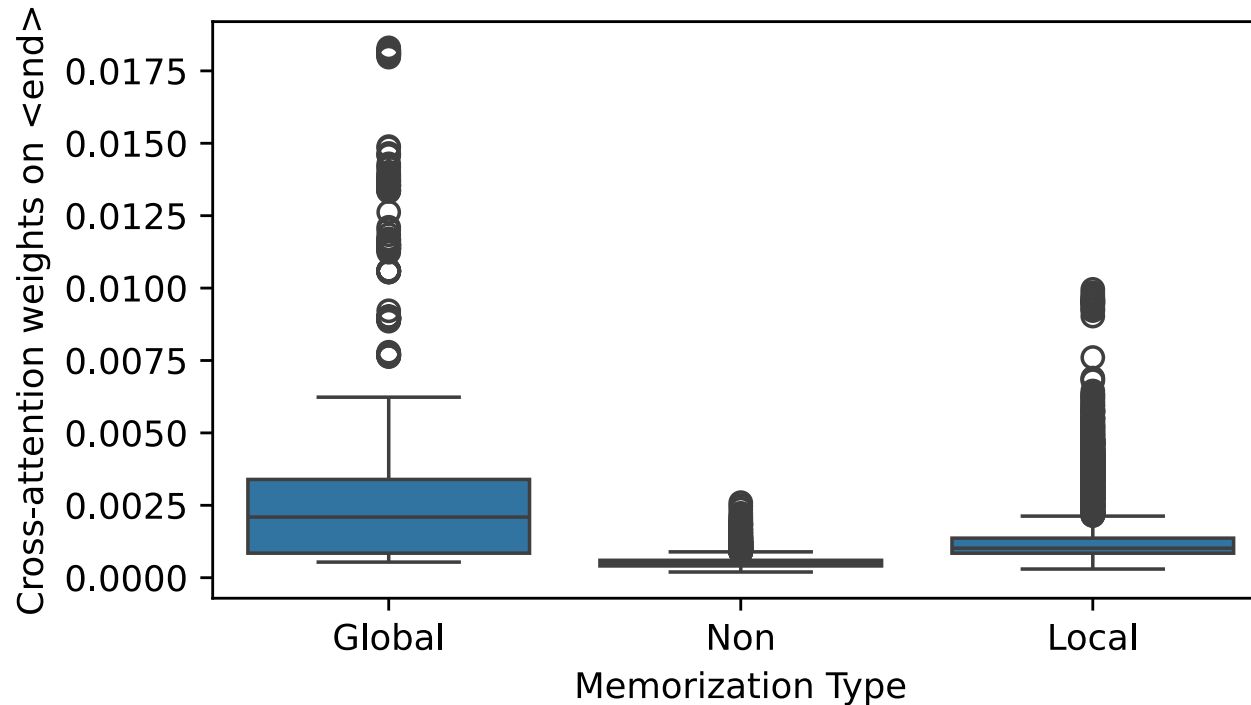
# The “Bright Ending” (BE) Phenomenon

- We then average and visualize such maps from the first two downsampling U-Net layers.
- **Observing the “bright ending” anomaly for memorized Diffusion Models.**



# The “Bright Ending” (BE) Phenomenon

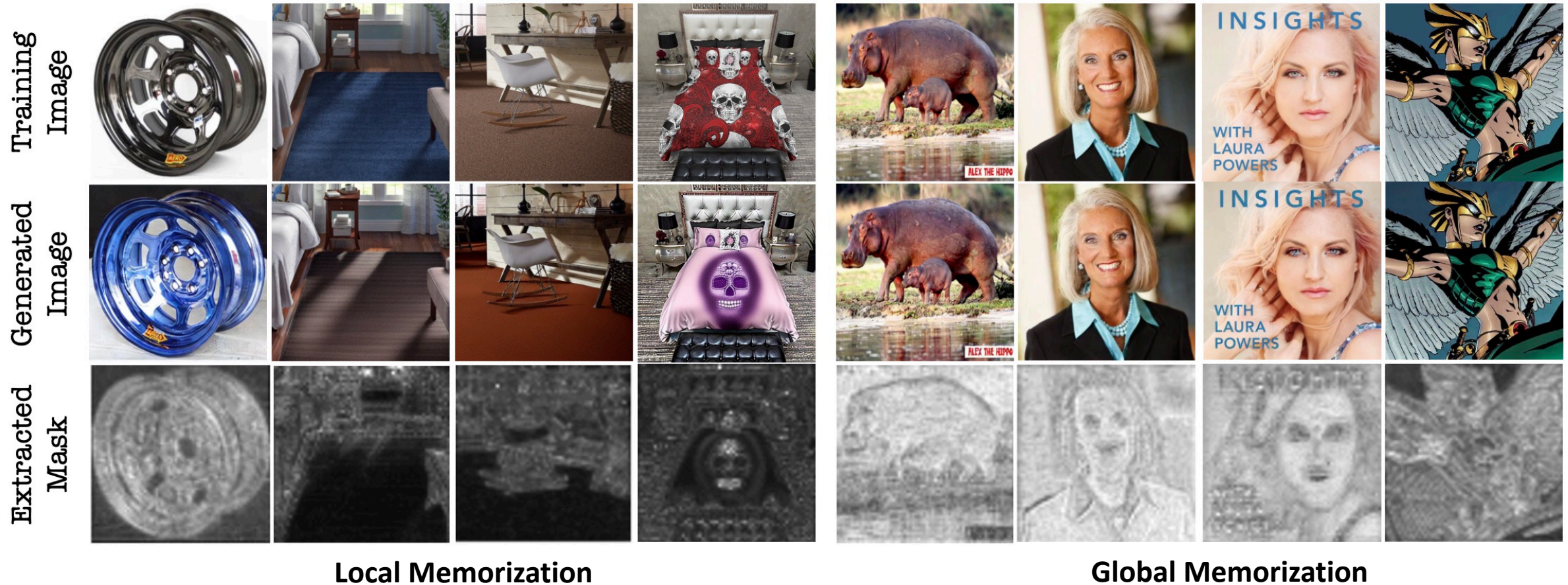
- Distributions of the attention scores of the  $\langle \text{EOS} \rangle$  token at the final inference step for global, local, and non-memorization scenarios can further validate our observation:





# The “Bright Ending” (BE) Phenomenon

- More examples:





# Localized Memorization Detection Strategy

- The cost function of the denoiser network (without text-conditioning):

$$L = \mathbb{E}_{t \in [1, T], \varepsilon \sim \mathcal{N}(0, I)} \left[ \left\| \varepsilon_t - \varepsilon_\theta(x_t, e_\phi) \right\|_2^2 \right]$$

- The cost function of the denoiser network (with text-conditioning):

$$L = \mathbb{E}_{t \in [1, T], \varepsilon \sim \mathcal{N}(0, I)} \left[ \left\| \varepsilon_t - \varepsilon_\theta(x_t, e_p) \right\|_2^2 \right]$$

- Computing the Bright Ending (BE) mask:

$$\mathbf{m} = \text{SoftMax} \left( \frac{QK^T}{\sqrt{d}} \right)$$

- Localized detection strategy by incorporating the BE mask:

$$LD = \frac{1}{T} \sum_{t=1}^T \left\| \left( \varepsilon_\theta(x_t, e_p) - \varepsilon_\theta(x_t, e_\phi) \right) \circ \mathbf{m} \right\|_2 / \left( \frac{1}{N} \sum_{i=1}^N m_i \right)$$

# Localized Memorization Mitigation Strategy

- The corresponding mitigation strategy relies on the magnitude-based detection strategy:

$$LD = \frac{1}{T} \sum_{t=1}^T \left\| \left( \varepsilon_{\theta}(x_t, e_p) - \varepsilon_{\theta}(x_t, e_{\phi}) \right) \circ \mathbf{m} \right\|_2 / \left( \frac{1}{N} \sum_{i=1}^N m_i \right)$$

- Localized mitigation strategy by incorporating the BE mask:
  - Improved loss function ( $LD$ ) that allows more effective prompt optimization.
  - More accurate trigger signal ( $LD$ ) based on the localized detection strategy.

# Localized Memorization Evaluation Strategy

- Global evaluation strategies:

$$SSCD(\hat{x}, x) = 1_{SSCD > 0.5}$$

$$S(\hat{x}, x) = -1_{SSCD < 0.5} \cdot \|\hat{x} - x\|_2$$

- Computing the Bright Ending (BE) mask:

$$\mathbf{m} = SoftMax\left(\frac{QK^T}{\sqrt{d}}\right)$$

- Localized evaluation strategy by incorporating the BE mask :

$$LS(\hat{x}, x) = -1_{SSCD < 0.5} \cdot \|(\hat{x} - x) \circ \mathbf{m}\|_2$$

# Results – Detecting Memorization

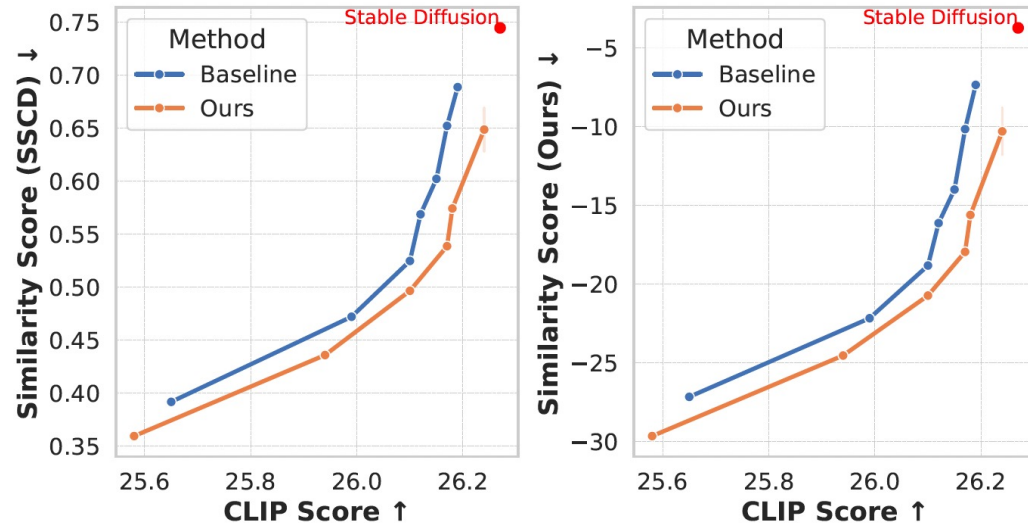
- Results show the incorporation of localization, and “bright ending” insights can improve existing state-of-the-art **detection** strategy’s performance on local memorization cases:

	First Step			First 10 Steps			All Steps		
	AUC	F1	T@1%F	AUC	F1	T@1%F	AUC	F1	T@1%F
Baseline - Local	0.918	0.864	0.629	0.989	0.982	0.953	0.990	0.983	0.560
Ours - Local	<b>0.943</b>	<b>0.893</b>	<b>0.731</b>	<b>0.995</b>	<b>0.987</b>	<b>0.985</b>	<b>0.996</b>	<b>0.988</b>	<b>0.926</b>
Baseline - Global	0.979	0.944	0.934	<b>1.000</b>	<b>0.987</b>	<b>1.000</b>	<b>0.999</b>	0.976	<b>1.000</b>
Ours - Global	<b>0.981</b>	<b>0.948</b>	<b>0.940</b>	<b>1.000</b>	<b>0.987</b>	<b>1.000</b>	<b>0.999</b>	<b>0.977</b>	<b>1.000</b>

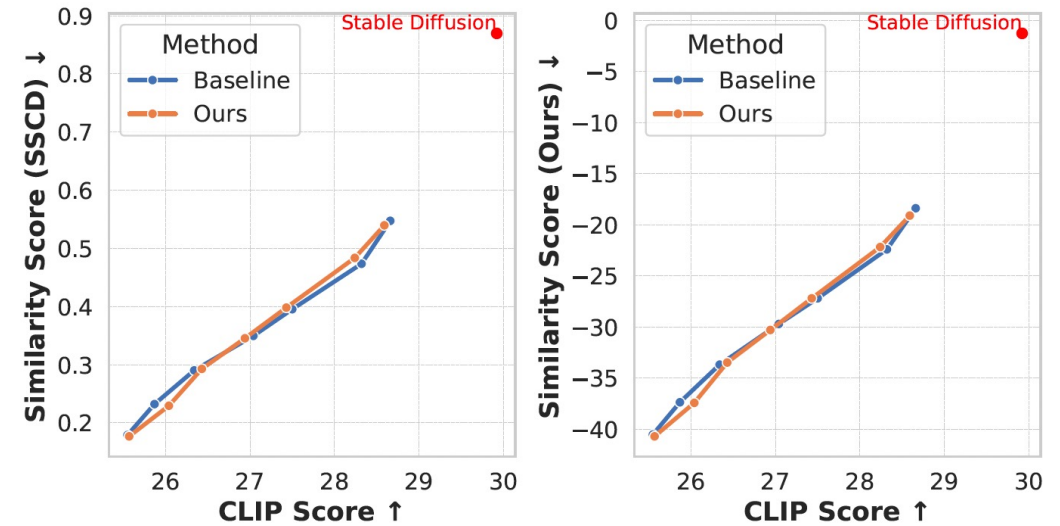


# Results – Mitigating Memorization

- Results show the incorporation of localization, and “bright ending” insights can improve existing state-of-the-art **mitigation** strategy’s performance on local memorization cases:



Local memorization's mitigation



Global memorization's mitigation

# Results – Evaluating Memorization

- Results show the incorporation of localization, and “bright ending” insights can improve existing state-of-the-art **evaluation** strategy’s performance on local memorization cases:

	Local	Global
SSCD	0.940	1.000
S	0.991	1.000
LS (Ours)	<b>0.995</b>	1.000

# Results – Summary

- Our localization insight can help improve the existing memorization detection, mitigation, and evaluation strategies.
- Bright Ending (BE) is effective in automatically extracting local memorization masks.