Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models

Chen Chen, Daochang Liu, Mubarak Shah, Chang Xu

CVPR 2025

Memorization in Diffusion Models

- Stable Diffusion can memorize training images, leading them to reproduce
  - Entire images (global memorization)
  - Parts of images (local memorization)

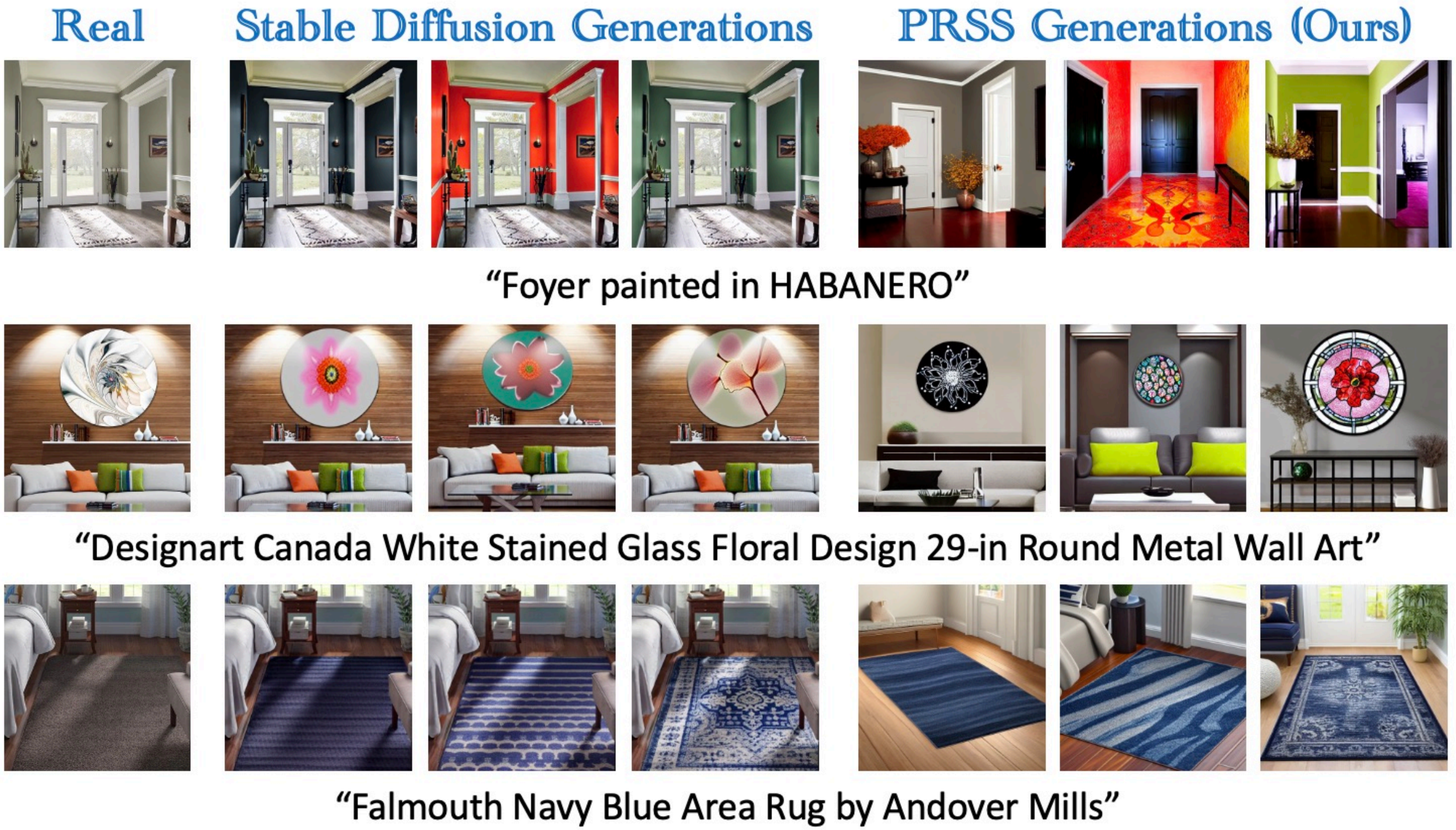
- This has sparked concerns about the
  - Originality of the generated images
  - Privacy issues

Chen et al., “Exploring Local Memorization in Diffusion Models via Bright Ending Attention.” ICLR, 2025.
Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Global Memorization

Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Local Memorization



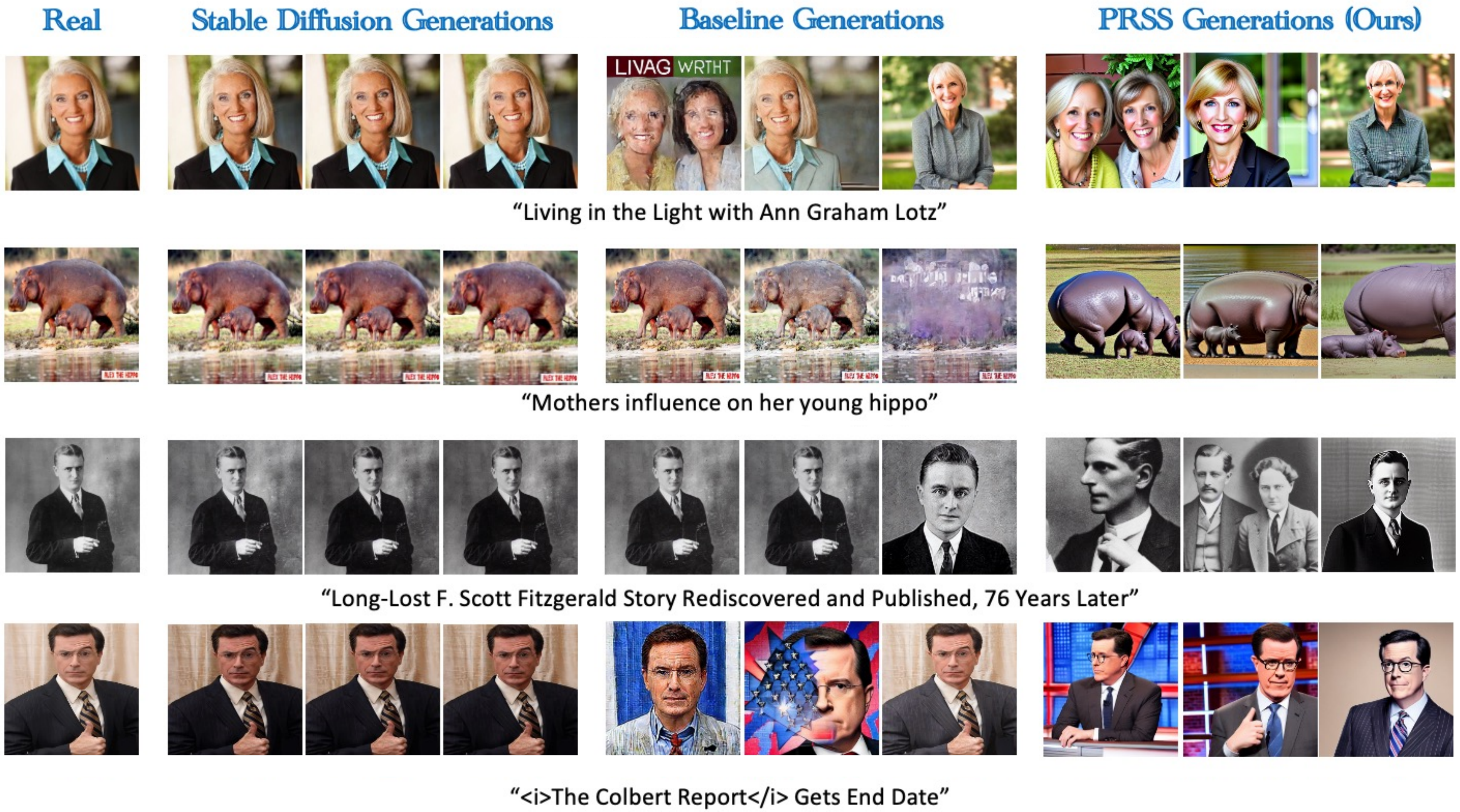
Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Research Gaps

- Existing mitigation strategies
  - Require re-training, searching over billions of images, or
  - Suffer from sub-optimal privacy-utility trade-offs

Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

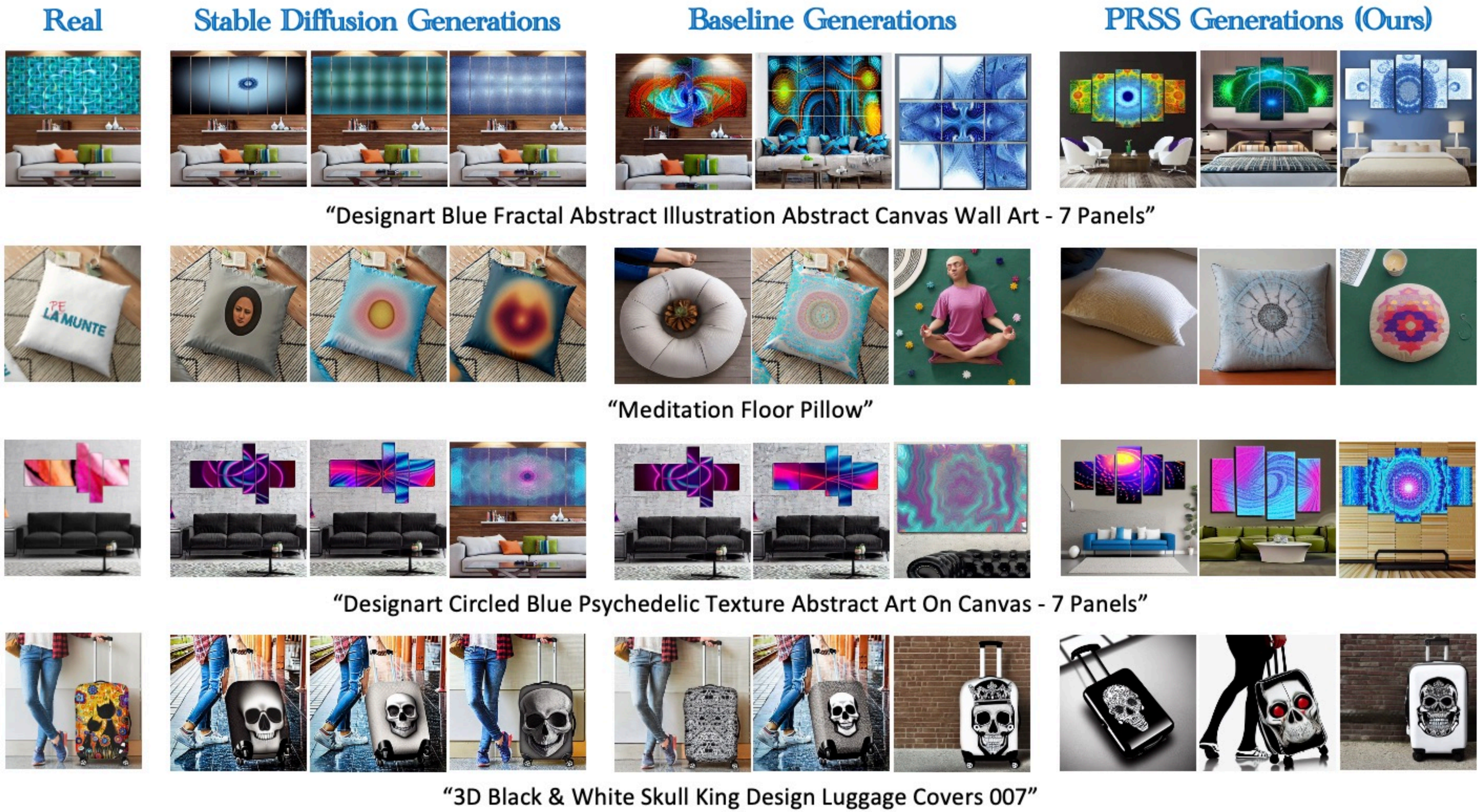
Research Gaps



Global Memorization Examples

Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Research Gaps



Local Memorization Examples

Chen et al., "Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models." CVPR, 2025.

Trade-off Analysis of Existing Strategies

- Classifier-free guidance (CFG):

$$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi)),$$

- The one-step memorization detection strategy:

$$m_t = \|\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi)\|_2,$$

- The corresponding one-step memorization mitigation strategy:

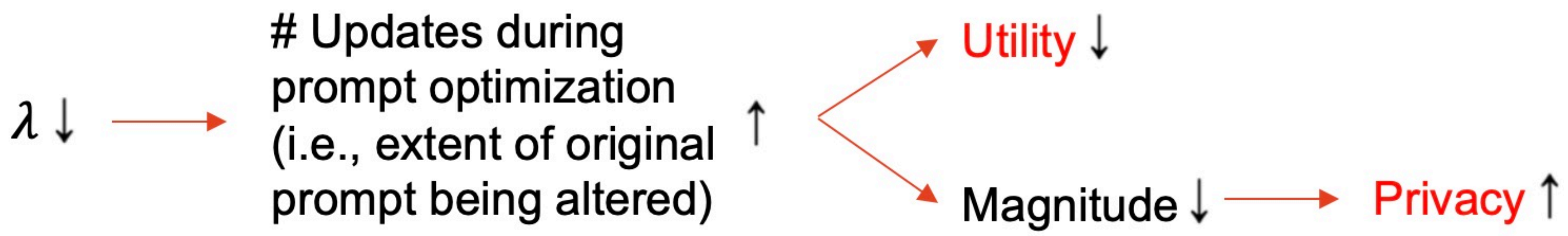
$$\begin{aligned}\hat{\epsilon} \leftarrow & [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1} < \lambda\}} \\ & + [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e^*) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1} > \lambda\}},\end{aligned}$$

Wen et al., ”Detecting, Explaining, and Mitigating Memorization in Diffusion Models.” ICLR, 2024.
Chen et al., “Exploring Local Memorization in Diffusion Models via Bright Ending Attention.” ICLR, 2025.

Trade-off Analysis of Existing Strategies

$$\hat{\epsilon} \leftarrow [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1} < \lambda\}}$$
$$+[\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e^*) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1} > \lambda\}},$$

- Currently, the privacy-utility trade-off is governed solely by how extensively the prompt is modified:

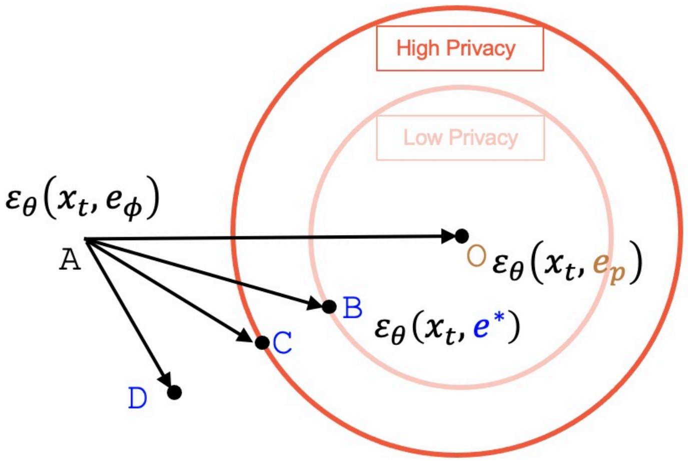


- Is it possible to reduce memorization more effectively using a method that achieves a better privacy-utility trade-off?

Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Trade-off Analysis of Existing Strategies

a) Privacy-Utility Trade-Off in Baseline



SIM(O, B) > SIM(O, C) > SIM(O, D)

$$\hat{\epsilon} \leftarrow [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1} < \lambda\}}$$
$$+[\epsilon_\theta(x_t, e_\phi) + s(\underbrace{\epsilon_\theta(x_t, e^*)}_{\text{Text-conditional prediction}} - \underbrace{\epsilon_\theta(x_t, e_\phi)}_{\text{Unconditional prediction}})]\mathbb{1}_{\{m_{T-1} > \lambda\}},$$

Contour Line of Equal Privacy — Baseline Guidance — Our PR Guidance — Our PRSS Guidance — **SIM** CLIP Similarity (Utility)

A Unconditional Prediction — O e_p -conditional prediction — BCD e^*-conditional prediction with different privacy level — B′C′D′ e_p^{ss}-conditional prediction with different privacy level

Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Insights

- We observe that both inputs in the standard CFG framework (the text-conditional and unconditional predictions) are suboptimal for mitigating memorization.
    - Text-conditional prediction
    - Unconditional prediction

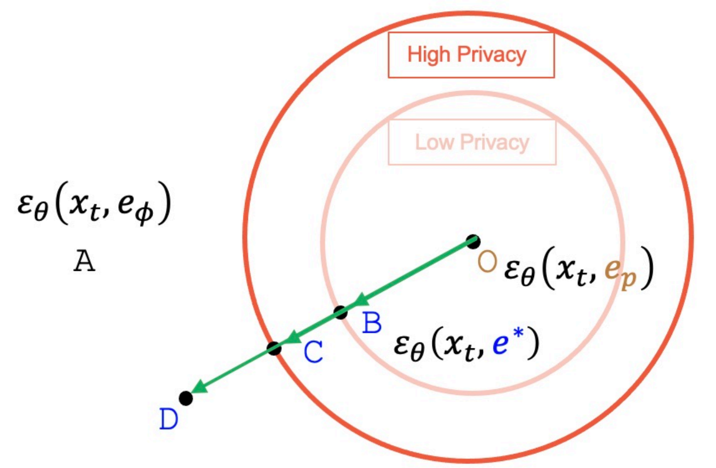
- Thus, we propose two meticulously crafted strategies to refine each of them.

Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Method: Prompt Re-anchoring (PR)

- PR recognizes e_p as a valuable anchor point for replacing the $e_\emptyset$ in CFG.

b) Enhancing Privacy with Prompt Re-Anchoring (PR)



SIM(O, B) > SIM(O, C) > SIM(O, D)

$$\hat{\epsilon} \leftarrow [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1}<\lambda\}} + [\epsilon_\theta(x_t, e_p) + s(\underbrace{\epsilon_\theta(x_t, e^*)}_{\text{Text-conditional prediction}} - \underbrace{\epsilon_\theta(x_t, e_p)}_{\textbf{Unconditional prediction}})]\mathbb{1}_{\{m_{T-1}>\lambda\}}.$$

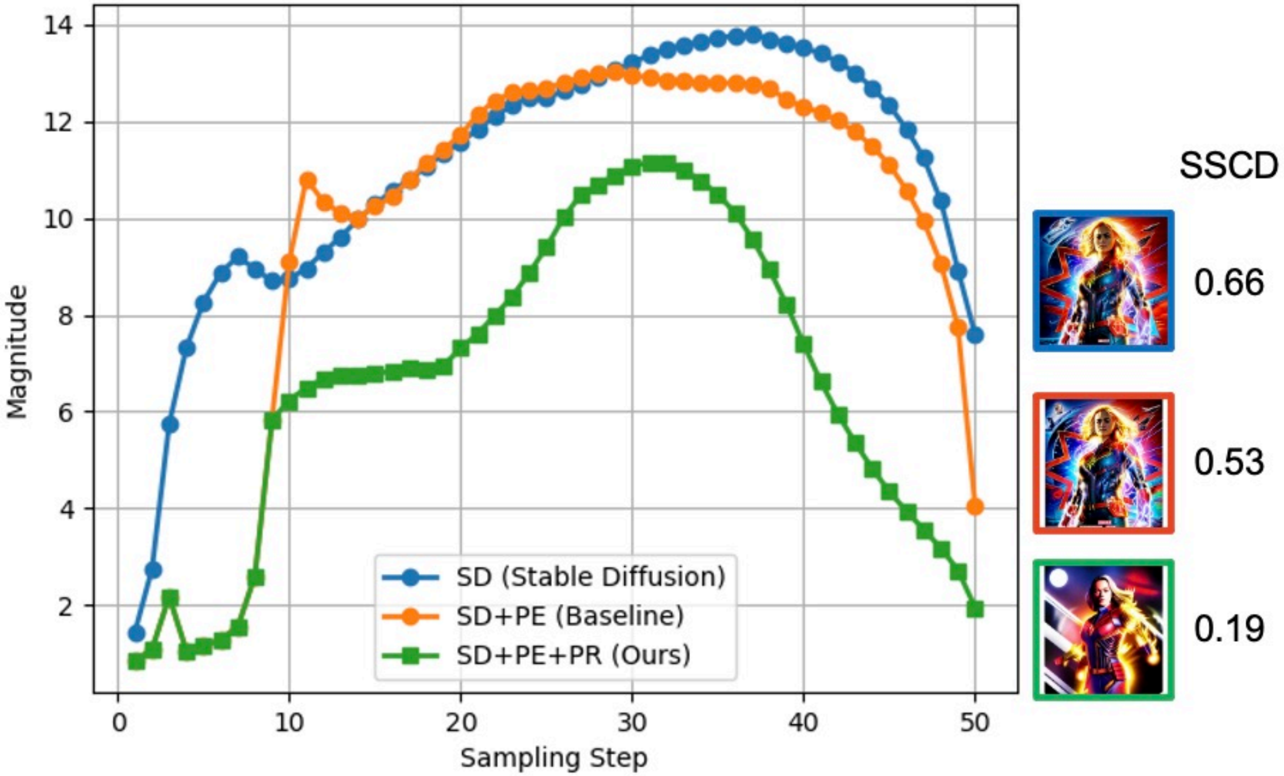
Contour Line of Equal Privacy — Baseline Guidance — Our PR Guidance — Our PRSS Guidance — **SIM** CLIP Similarity (Utility)

A Unconditional Prediction — O e_p -conditional prediction — BCD e^*-conditional prediction with different privacy level — B′C′D′ e_p^{ss}-conditional prediction with different privacy level

Chen et al., "Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models." CVPR, 2025.

Method: Prompt Re-anchoring (PR)

- Analysing PR

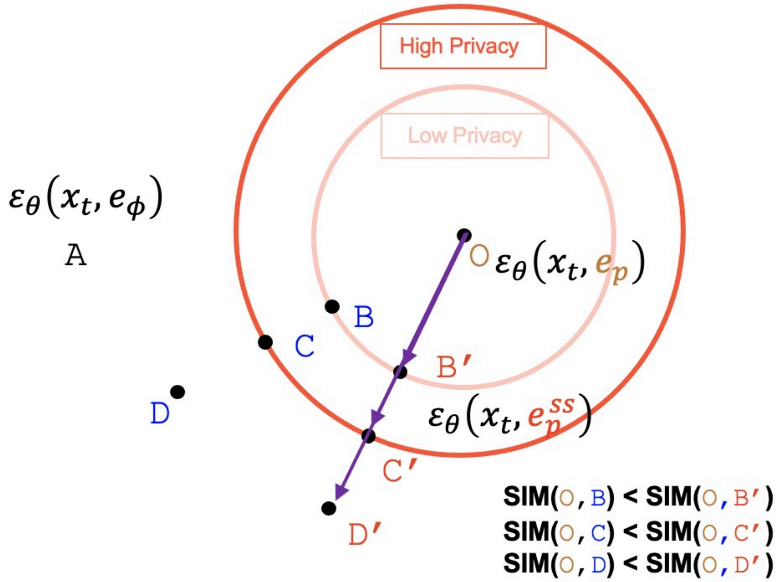


Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Method: Semantic Search (SS)

- Instead of e^*, SS uses an LLM to find less memorized e_p^{ss} that is semantically similar to e_p:

c) Enhancing Utility with Semantic Search (SS)



$$\hat{\epsilon} \leftarrow [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1} < \lambda\}}$$
$$+[\epsilon_\theta(x_t, e_p) + s(\underbrace{\epsilon_\theta(x_t, e_p^{ss})}_{\textbf{Text-conditional prediction}} - \underbrace{\epsilon_\theta(x_t, e_p)}_{\text{Unconditional prediction}})]\mathbb{1}_{\{m_{T-1} > \lambda\}}.$$

Contour Line of Equal Privacy — Baseline Guidance — Our PR Guidance — Our PRSS Guidance — **SIM** CLIP Similarity (Utility)

A Unconditional Prediction

O e_p -conditional prediction

BCD e^*-conditional prediction with different privacy level

B′ C′ D′ e_p^{ss}-conditional prediction with different privacy level

Chen et al., "Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models." CVPR, 2025.

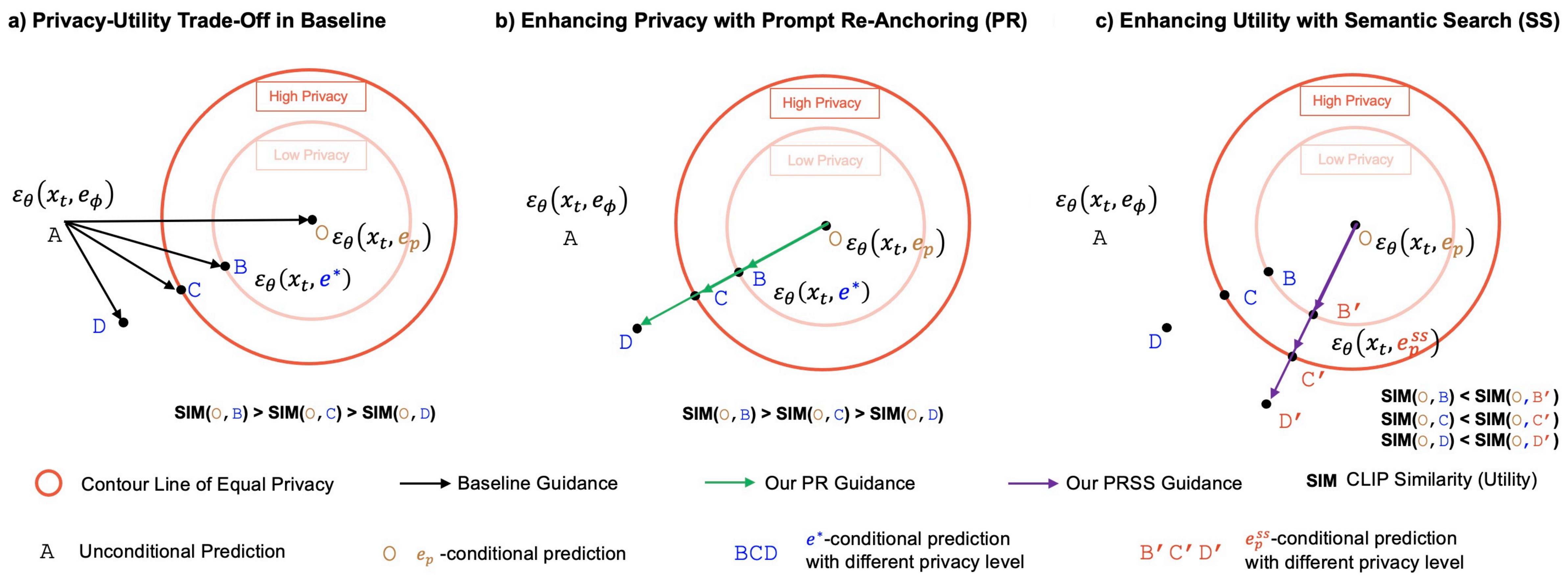
Method: Semantic Search (SS)

- Analysing SS

		Stable Diffusion (SD)	w/ Prompt Eng. Baseline (PE)	w/ Semantic Search (SS)
				
Memorization		Yes	No	No
Privacy	m_{T-1}	Poor (7.38)	Good (1.15)	Good (0.78)
Utility	$CLIP_{txt}$	High (100.00)	Low (2.32)	High (93.82)
	$CLIP_{img}$	High (35.92)	Low (13.25)	High (25.20)

Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

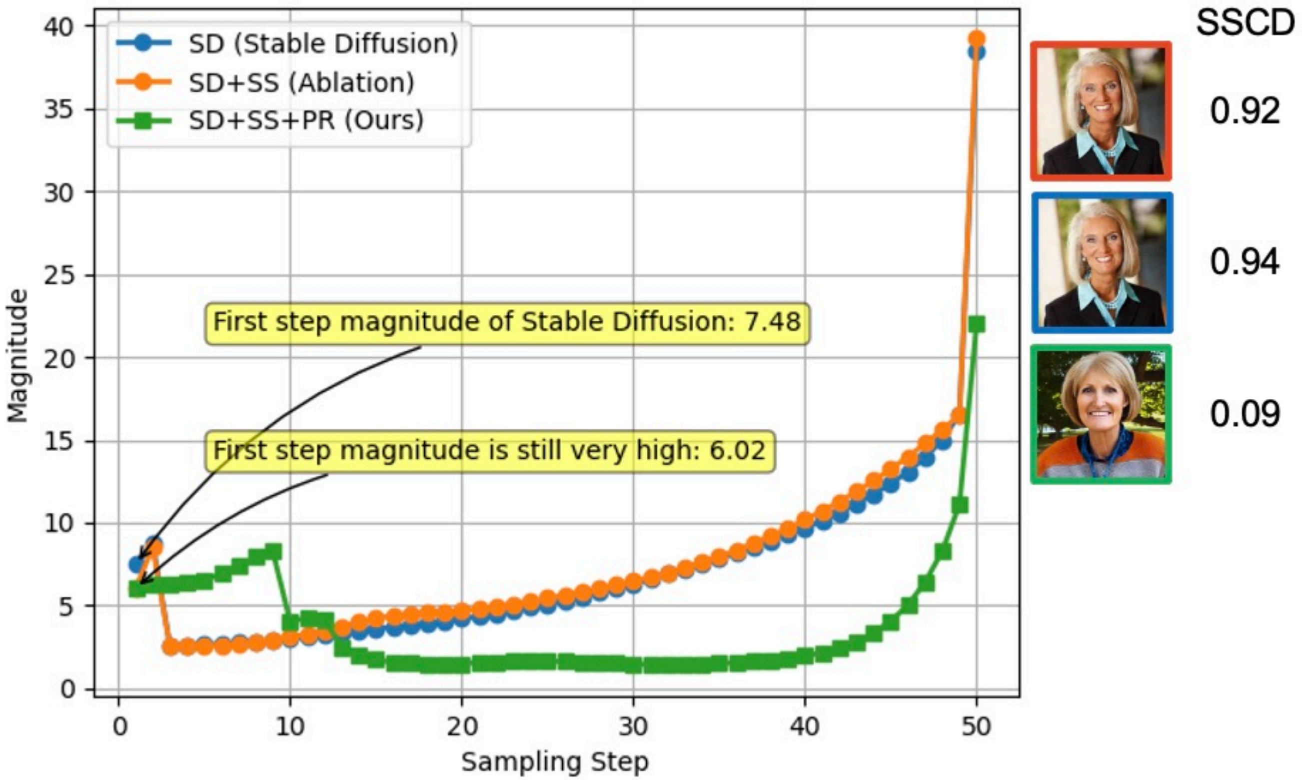
Overview of Our Improvements via PRSS



Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

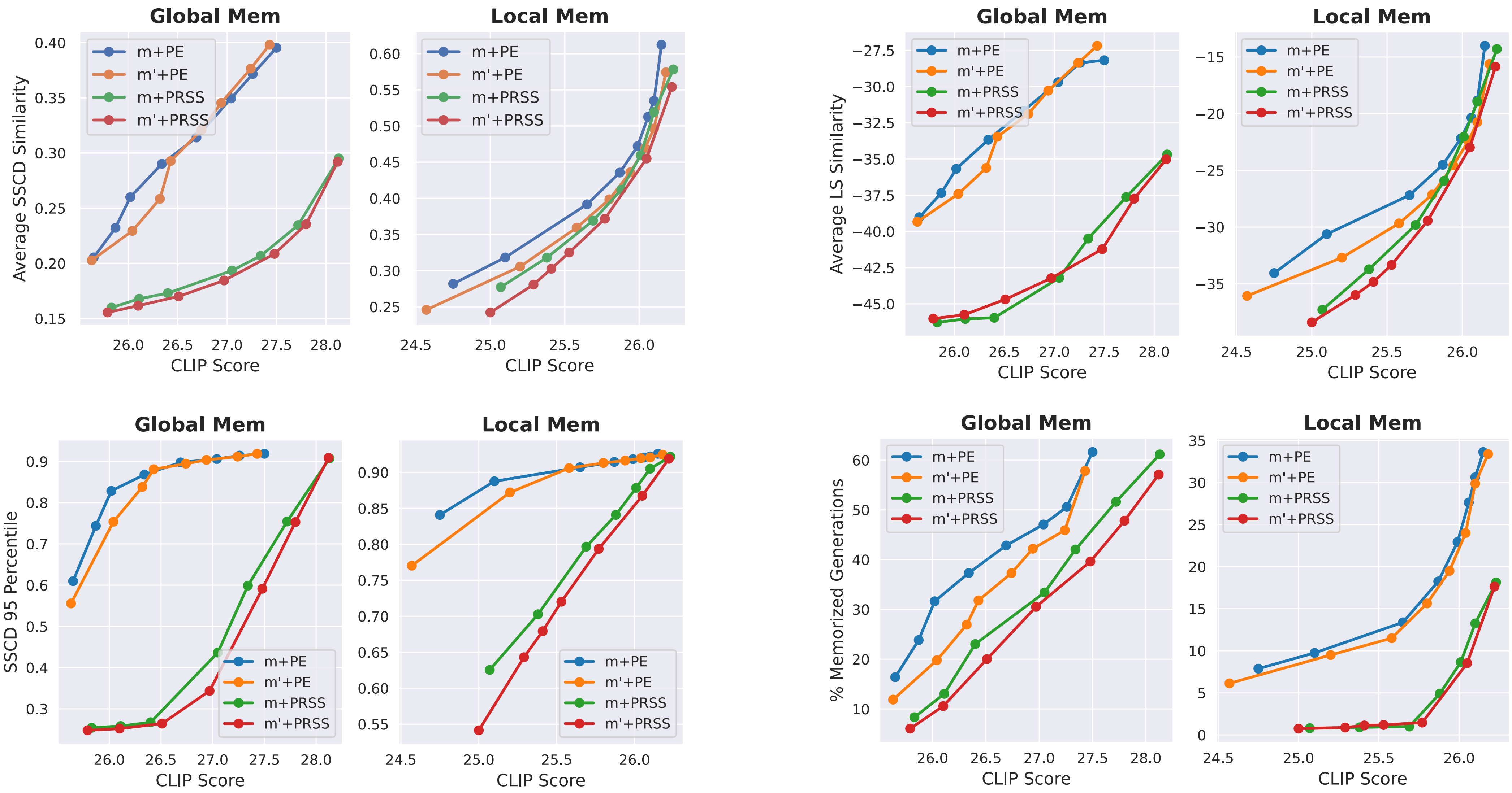
Synergy Effect of the Two Strategies (PR and SS)

- Analysing PRSS



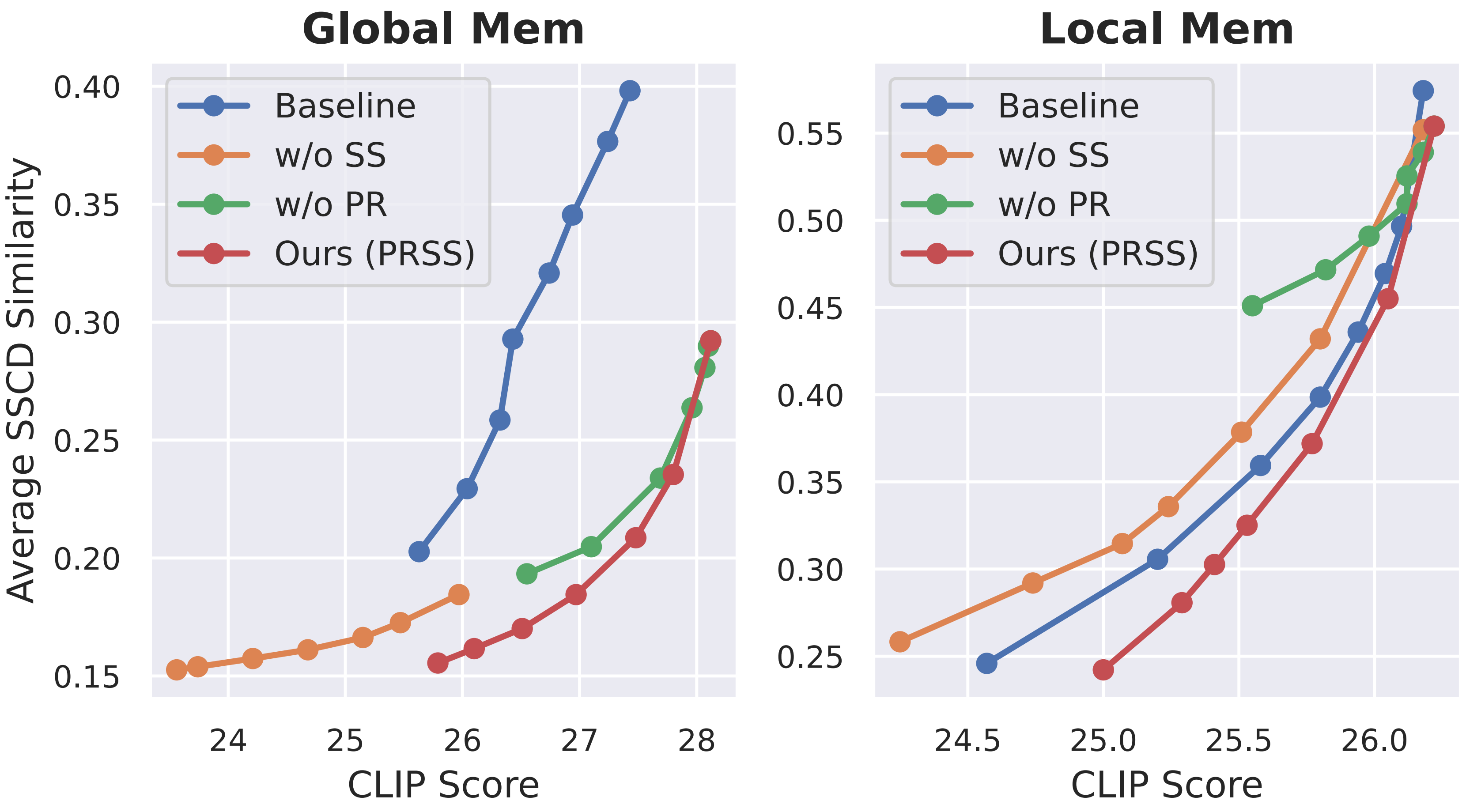
Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Quantitative Results



Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

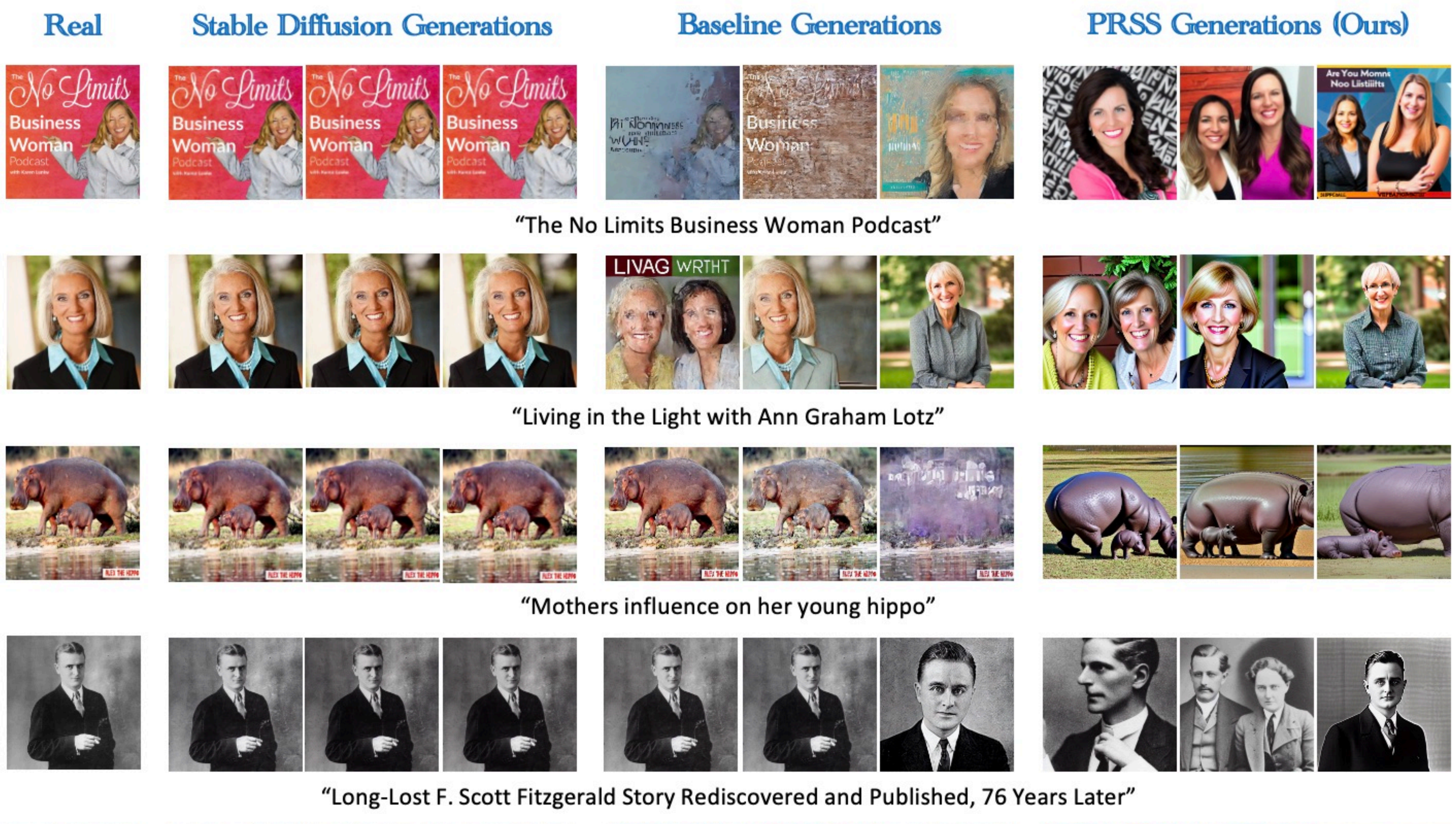
Ablation Studies



Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Qualitative Results

- Global memorization mitigation



Chen et al., "Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models." CVPR, 2025.

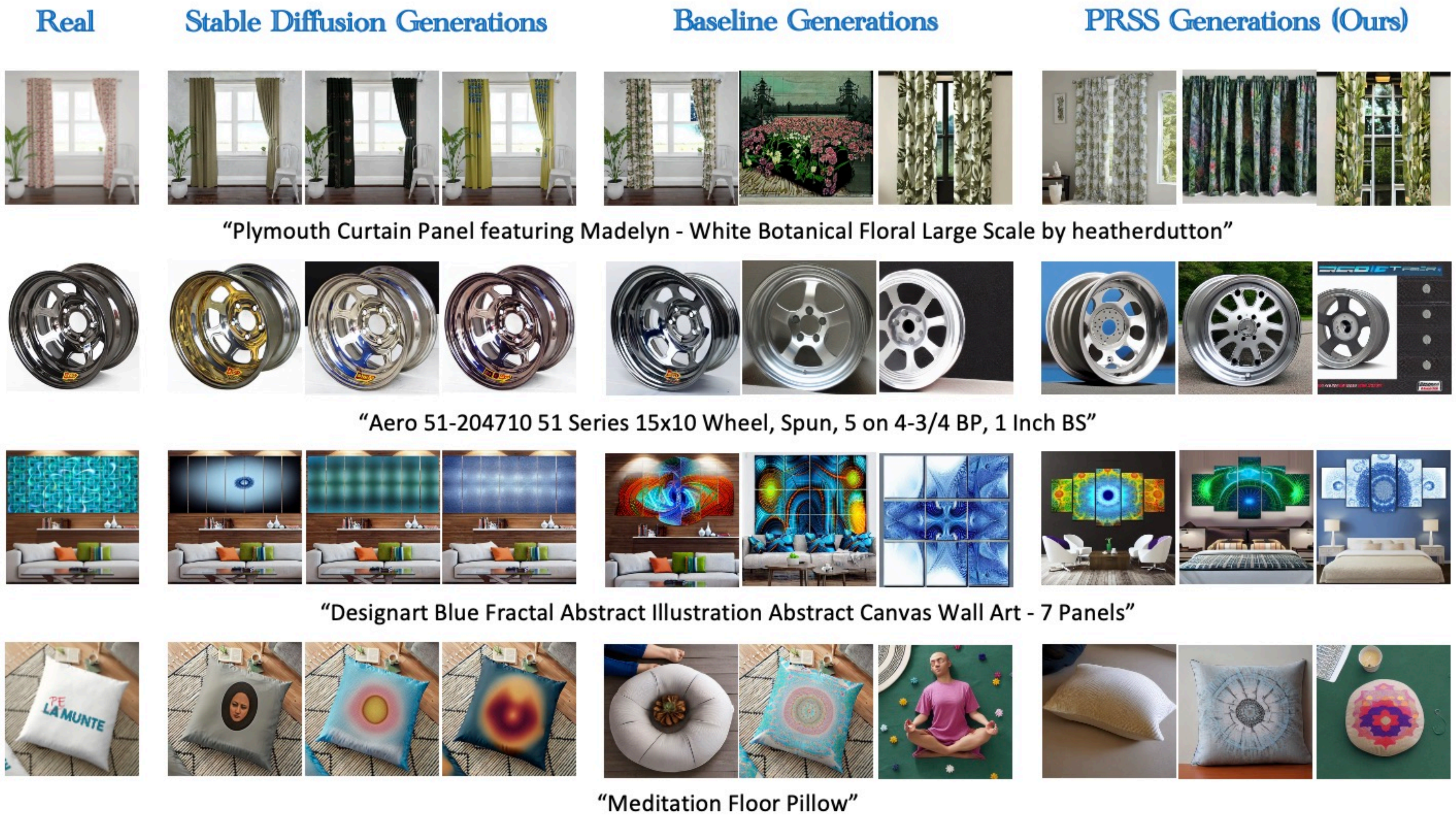
Qualitative Results

- Global memorization mitigation

Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Qualitative Results

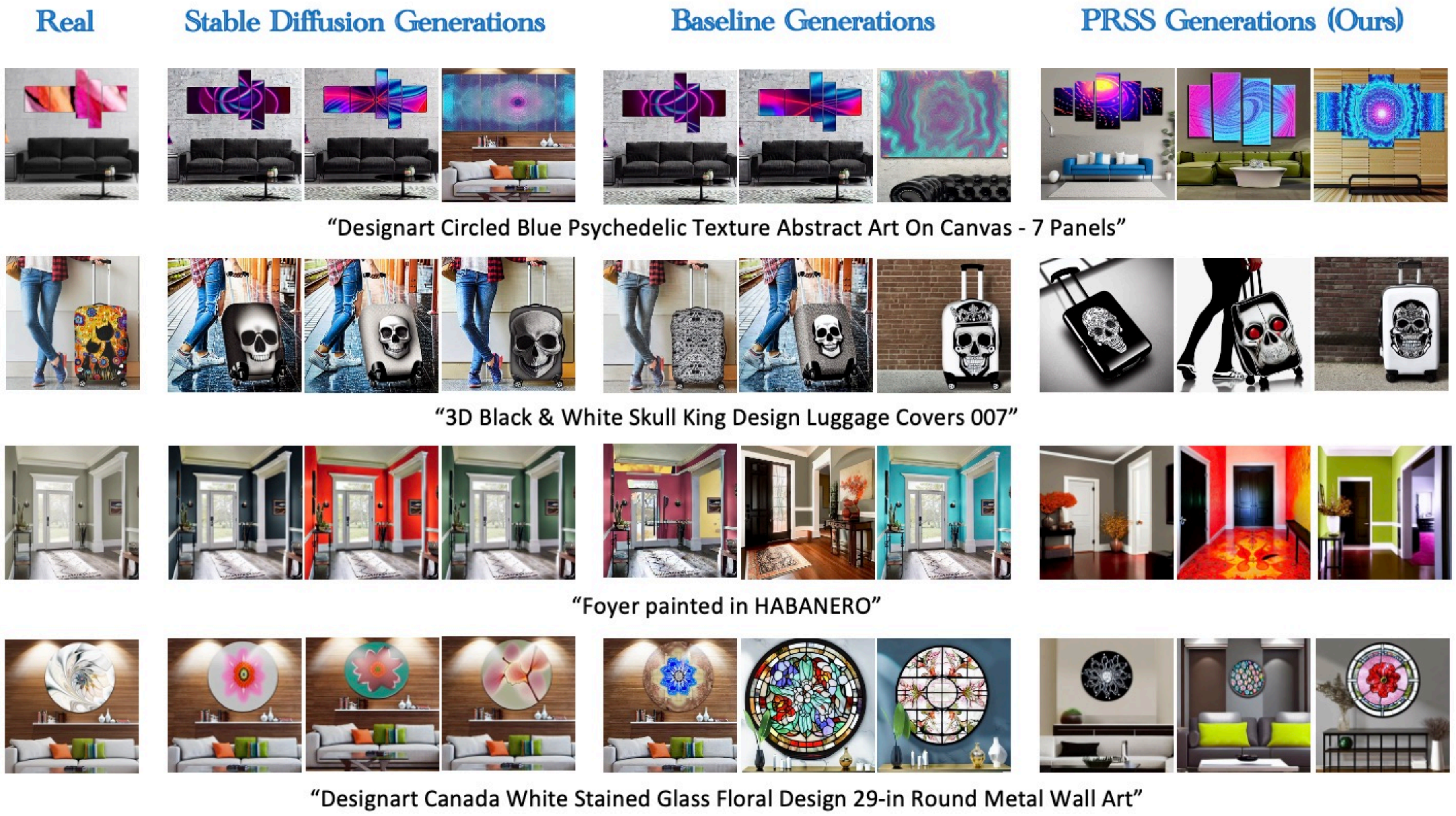
- Local memorization mitigation



Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.

Qualitative Results

- Local memorization mitigation



Chen et al., "Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models." CVPR, 2025.

Conclusion

- This paper introduces *PRSS*, which consists of *Prompt Re-anchoring (PR)* & *Semantic Search (SS)* as deliberately designed strategies to mitigate memorization in diffusion models.

- Results show that PRSS enhances existing methods' privacy-utility trade-off at different privacy levels.

- Implementationally, PRSS is simple and efficient, necessitating only adjustments to the existing CFG equation without re-training, fine-tuning, or searching over the training data.

Chen et al., "Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models." CVPR, 2025.

Limitation and Future Work

- PRSS relies on the accuracy of a pre-defined detection strategy to trigger the mitigation mechanism.

- Thus, improving the detection mechanism's accuracy in future works would orthogonally bolster the effectiveness of PRSS.

Chen et al., “Enhancing Privacy-Utility Trade-offs to Mitigate Memorization in Diffusion Models.” CVPR, 2025.