This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Chen Chen[1] Daochang Liu[2] Mubarak Shah[3] Chang Xu[1]
[1]School of Computer Science, Faculty of Engineering, The University of Sydney, Australia
[2]School of Physics, Mathematics and Computing, The University of Western Australia, Australia
[3]Center for Research in Computer Vision, University of Central Florida, USA
{cche0711@uni., c.xu@}sydney.edu.au daochang.liu@uwa.edu.au shah@crcv.ucf.edu

Abstract

Text-to-image diffusion models have demonstrated remarkable capabilities in creating images highly aligned with user prompts, yet their proclivity for memorizing training set images has sparked concerns about the originality of the generated images and privacy issues, potentially leading to legal complications for both model owners and users, particularly when the memorized images contain proprietary content. Although methods to mitigate these issues have been suggested, enhancing privacy often results in a significant decrease in the utility of the outputs, as indicated by text-alignment scores. To bridge the research gap, we introduce a novel method, PRSS, which refines the classifier-free guidance approach in diffusion models by integrating prompt re-anchoring (PR) to improve privacy and incorporating semantic prompt search (SS) to enhance utility. Extensive experiments across various privacy levels demonstrate that our approach consistently improves the privacy-utility trade-off, establishing a new state-of-the-art.

1. Introduction

Leveraging classifier-free guidance (CFG) [12], text-to-image diffusion models like Stable Diffusion [28] and Midjourney [1] are now capable of generating highly realistic images that closely align with user-provided text prompts. This capability has propelled their popularity, leading to widespread use and distribution of their generated images. However, recent research [6, 8, 31, 34] has revealed a critical issue: these models can memorize training data, leading them to reproduce parts of images, such as foregrounds or backgrounds (local memorization, see Fig. 1), or even entire images (global memorization, see Fig. 2) during inference, instead of generating genuinely novel content. When the training data includes sensitive or copyrighted material, these memorization issues can infringe on copyright laws without notifying either the model's owners or users or the copyright holders of the replicated content. The risk is especially substantial given the extensive use of these models and their reliance on massive datasets, such as LAION [30], which contain billions of web-scale images and are impractical to thoroughly review or filter manually. This risk is further highlighted by real-world cases, where several artists have filed lawsuits, arguing that models like Stable Diffusion act as "21st-century collage tools" that remix their copyrighted works, implicating Stability AI, DeviantArt, Midjourney, and Runway AI. Recognizing the potential for unauthorized reproductions, Midjourney has even banned prompts containing the term "Afghan" to prevent the generation of the copyrighted Afghan Girl photograph. Yet, as [35] demonstrates, such restrictions alone are insufficient to fully prevent the reproduction of copyrighted images. This underscores the urgent need for timely effective mitigation strategies to address these concerns.

In response to these legal challenges, recent initiatives [7, 8, 27, 32, 36] have focused on developing strategies to minimize memorization, achieving notable success. These approaches vary in scope, with some targeting the model's training phase and others making adjustments during inference. While training-phase strategies are theoretically viable, they lack practical significance as they predominantly experiment on a small 10k subset of the entire LAION5B dataset due to the impracticality of fine-tuning on the full dataset — a constraint stemming from computational limitations. In contrast, inference-phase strategies, which involve modifications only during inference, offer computational efficiency and adaptability, as they require no additional training or fine-tuning. This makes them highly applicable across pre-trained diffusion models, regardless of specific training configurations. Despite their promise, inference-time strategies can also suffer from computational constraints. For example, the approach proposed by [7] involves searching the entire LAION dataset for the nearest neighbor to optimize the generated latent representation during each denoising step. Other inference-time strategies [8, 27, 32, 36] that avoid nearest neighbor search



Figure 1. Examples of *Local Memorization*: Stable Diffusion generations (columns 2–4) can replicate local regions of training images (column 1). Our PRSS method effectively mitigates memorization while maintaining strong alignment with the input prompts (columns 5-7).



Figure 2. Examples of *Global Memorization*: Stable Diffusion generations (columns 2–4) can replicate the entire training images (column 1). Our PRSS method effectively mitigates memorization while maintaining strong alignment with the input prompts (columns 5-7).

demonstrate high efficiency but only moderate effectiveness, leaving a substantial gap in optimizing the privacy-utility trade-off. In particular, these methods often require significant compromises in text-image alignment to reduce memorization and conversely, struggle to sufficiently suppress memorization when trying to preserve alignment.

To address these limitations, we carefully analyze why existing strategies struggle to achieve an optimal privacy-utility trade-off, uncovering critical insights that the two inputs (text-conditional and unconditional predictions) used in the CFG framework are both suboptimal for mitigating memorization, even though they are standard for most other tasks in Stable Diffusion models. Specifically, we find that the text-conditional term in CFG, when optimized for privacy by prompt engineering, can significantly reduce text alignment, indicating the need for alternative methods to find privacy-preserving yet intent-aligned text-conditioning. Additionally, we observe that the unconditional term in CFG is less effective as an anchor for privacy preservation, suggesting that alternative anchoring points that enhance privacy are necessary. Building on these insights, we introduce *PRSS*, a novel method that incorporates two complementary strategies to refine the widely used CFG approach for improved memorization mitigation. Our method includes (1) *prompt re-anchoring*, designed to preserve high privacy with a smaller utility cost, which re-anchors the unconditional component in CFG using the memorized prompt to consistently guide generations away from memorization during the inference process; (2) *semantic prompt search*, designed to preserve high utility with minimal privacy compromise, which optimizes the prompt away from the memorized one in the language space while keeping similar semantics. Our experimental results reveal that employing both strategies synergistically can consistently outperform baseline methods in achieving superior privacy-utility trade-offs at different privacy levels. In particular, this results in higher text alignment (utility) and substantially fewer memorized generations (improved privacy). Furthermore, ablation studies validate the efficacy of each component within our strategy. From an implementation standpoint, our strategies are simple, efficient, and effective, necessitating only adjustments to the existing CFG equation for updating the noise prediction during inference. This integration is seamless, avoiding disruption to the inference workflow and eliminating the need for additional re-training, fine-tuning, or searching over the training data.

2. Preliminaries

2.1. Classifier-Free Guidance in Diffusion Models

Diffusion models [13, 23] are popular generative models that consist of a forward diffusion process that consistently adds Gaussian noise to real images x_0 for T steps based on the following conditional function, resulting in $x_T \sim \mathcal{N}(0, \mathbf{I})$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where x_t represents the noised version of x_0 after adding t noise, and $\{\beta_t\}_{t=1}^T$ is the schedule that controls the magnitude of added noise at time t.

This is followed by the reverse process, in which a denoiser model is trained to predict the distribution (i.e., mean $\mu_\theta(x_t)$ and variance $\sigma_\theta^2(x_t)\mathbf{I}$) of a cleaner version of x_t, denoted as x_{t-1}, at any given timestep t:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t), \sigma_\theta^2(x_t)\mathbf{I}). \tag{2}$$

During training, the objective function is equivalent to have the denoiser $\epsilon_\theta(x_t, y, t)$ predict the noise ϵ_t to be removed, rather than the image x_{t-1} at timestep t, conditioned on y, which may represent a user-provided text prompt:

$$\mathcal{L} = \mathbb{E}_{t\in[1,T],\epsilon\sim\mathcal{N}(0,\mathbf{I})}[\|\epsilon_t - \epsilon_\theta(x_t, y, t)\|_2^2]. \tag{3}$$

Classifier-free guidance (CFG) [12] is a widely used technique for balancing fidelity and diversity in the generated outputs of diffusion models, and it can be easily applied during inference. CFG operates by linearly guiding the noise prediction from the unconditional prediction, $\epsilon_\theta(x_t, t)$ towards the conditional prediction, $\epsilon_\theta(x_t, y, t)$, with a scaling factor $s > 1$ to control the degree of guidance. For simplicity, removing the timestep t from the notation, the adjusted noise prediction can be defined as:

$$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t) + s \cdot (\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t)). \tag{4}$$

2.2. Memorization in Diffusion Models

Definition and Metrics. A widely adopted practical definition of memorization refers to generated images exhibiting high object-level similarity to training data [31]. This includes the concept of reconstructive memory in diffusion models, where foreground and background objects from training images are combined, potentially with modifications such as shifting, scaling, or cropping. These reconstructions may not match any training image pixel-for-pixel but still closely resemble specific training images at the object level. The standard metric for evaluating memorization [7, 8, 27, 31, 32, 36] is designed base on a pre-trained Self-Supervised Copy Detection (SSCD) model [25]:

$$\text{SSCD}(\hat{x}) = \max_{x_i \in \mathcal{D}} \text{Cosine}(\mathbf{E}_{SSCD}(\hat{x}), \mathbf{E}_{SSCD}(x_i)), \tag{5}$$

where for each generated image $\hat{x}$, cosine similarity is computed between its SSCD embedding $\mathbf{E}_{SSCD}(\hat{x})$ and the embeddings $\mathbf{E}_{SSCD}(x_i)$ of all training images x_i in the dataset $\mathcal{D}$. The maximum similarity serves as the final SSCD score.

Most recently, [8] proposed a method to localize memorized regions in generated images, arguing that only memorized regions pose privacy risks, making local memorization a more meaningful concept. They introduced an efficient approach to synthesize a mask that highlights these regions without requiring access to training data. Using this mask, they proposed a new localized metric based on both object-level SSCD and pixel-level Euclidean distance:

$$LS(\hat{x}, x) = -\mathbb{1}_{\text{SSCD}>0.5} \cdot \|(\hat{x} - x) \circ \mathbf{m}\|_2, \tag{6}$$

where $\mathbf{m}$ is the local memorization mask and $\mathbb{1}_{\text{SSCD}>0.5}$ is an indicator function that signifies whether the SSCD score exceeds the 0.5 threshold, which is commonly used to distinguish memorized from non-memorized content.

Memorization Detection. A recent work [36] introduces an innovative approach for detecting potential memorized generations during inference by leveraging the magnitude of the difference between text-conditional $\epsilon_\theta(x_t, e_p)$ and unconditional $\epsilon_\theta(x_t, e_\phi)$ noise predictions:

$$m_t = \|\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi)\|_2, \tag{7}$$

where memorized generations consistently exhibit larger magnitudes than non-memorized ones, making this magnitude a strong signal for detection. Furthermore, accurate detection can be achieved efficiently by evaluating the magnitude m_{T-1} at the first denoising step (timestep $T-1$, highlighting the method's computational efficiency.



Figure 3. Illustration of the logic behind how λ in Eq. (10) governs the privacy-utility trade-off. Currently, the privacy-utility trade-off is governed solely by how extensively the prompt is modified.

The contemporary work [8] extends this detection method by incorporating their proposed local memorization mask $\mathbf{m}$ into Eq. (7) to compute a masked magnitude, achieving state-of-the-art performance:

$$m'_t = \|(\varepsilon_\theta(x_t, e_p) - \varepsilon_\theta(x_t, e_\phi)) \circ \mathbf{m}\|_2 \Big/ \left(\frac{1}{N}\sum_{i=1}^{N} m_i\right), \tag{8}$$

where $\mathbf{m}$ is the local memorization mask and $\frac{1}{N}\sum_{i=1}^{N} m_i$ is used to normalize the result by the mean of $\mathbf{m}$.

3. Method

We begin with a detailed analysis in Sec. 3.1 to uncover the key challenges that hinder existing strategies from achieving an optimal privacy-utility trade-off. Building on these insights, we evolve the baseline step by step with our proposed prompt re-anchoring (PR) strategy in Sec 3.2, which focuses on improving privacy, and semantic prompt search (SS) in Sec 3.3, which enhances utility. Finally, in Sec. 3.4, we analyze how the combination of PR and SS achieves a synergistic effect, offering the best privacy-utility trade-off.

3.1. Trade-off Analysis of Existing Strategies

Our method is built upon the widely-used classifier-free guidance (CFG) by treating the input text prompt from user e_p as the condition y in Eq. (4) and rewriting the unconditional noise as a null condition e_ϕ,

$$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi)). \tag{9}$$

The one-step prompt engineering (PE) based memorization mitigation method proposed by [36] then counteract memorization by reformulating Eq. (9) as:

$$\begin{aligned}\hat{\epsilon} \leftarrow\ & [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1}<\lambda\}} \\ & + [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e^*) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1}>\lambda\}},\end{aligned} \tag{10}$$

where $m_{T-1} = \|\epsilon_\theta(x_{T-1}, e_p) - \epsilon_\theta(x_{T-1}, e_\phi)\|_2$ represents the magnitude at the initial step $T-1$ as per Eq. (7), $\epsilon_\theta(x_t, e_p)$ and $\epsilon_\theta(x_t, e^*)$ are the noise predictions conditioned on the user prompt embedding e_p and engineered prompt embedding e^* (optimized based on the gradient computed from the first-step magnitude m_{T-1}), and

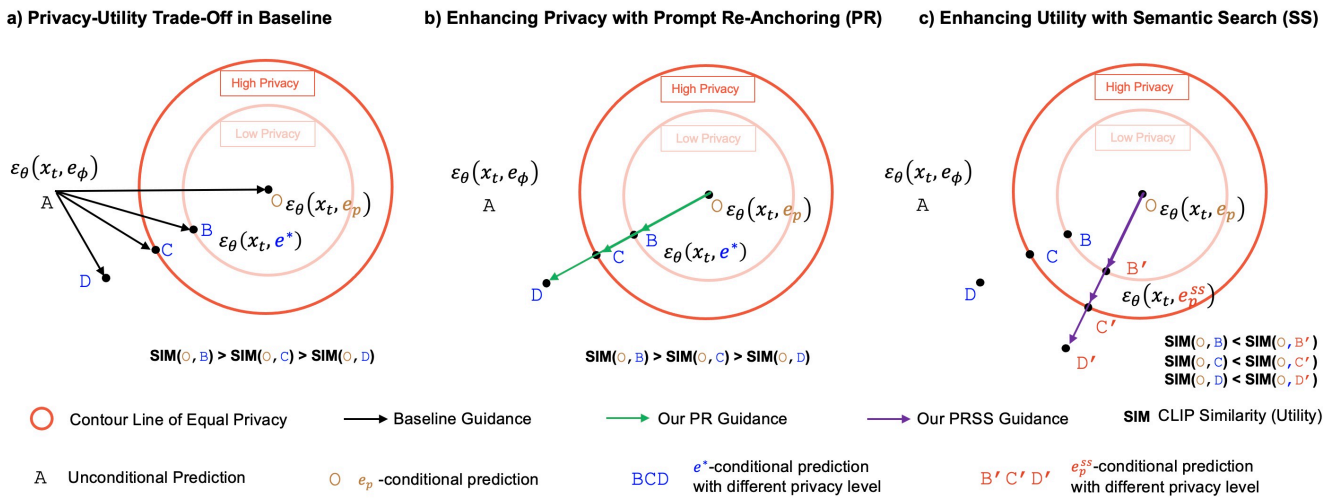


Figure 4. Illustration of our proposed strategy (PRSS) compared to the baseline prompt engineering (PE) method. **a)** The privacy-utility trade-off in the baseline method, where the privacy increases at the cost of utility from B to C to D. **b)** Our proposed PR guidance, i.e, the green arrow, drives the image out of the low-privacy region more efficiently than the baseline guidance, i.e, the black arrows. **c)** Our proposed SS can enhance utility, by finding points B′C′D′ that are at the same level of privacy as BCD but with better utility. In summary, using the proposed PR and SS jointly can enhance the privacy-utility trade-off to mitigate memorization in diffusion models.

$\epsilon_\theta(x_t, e_\phi)$ is the unconditional noise prediction. This strategy detects potential memorization for a given input prompt based on the magnitude signal m_{T-1} with a threshold λ, and undertakes soft prompt engineering (*i.e.*, optimizes the prompt embeddings) if potential memorization is detected. In this context, if m_{T-1} is under λ, the prompt is considered safe from memorization, and the generation proceeds with the original user prompt e_p, mirroring Stable Diffusion's output. Conversely, if m_{T-1} is above λ, the prompt is then identified as likely to produce memorized images, thus e_p is optimized to and replaced with an engineered prompt e^* in the conditional prediction while keeping the unconditional prediction $\epsilon_\theta(x_t, e_\phi)$ as is following the classical CFG equation. Specifically, e^* is optimized to reduce the first-step magnitude m_{T-1} using gradient descent and the optimization halts once the magnitude m_{T-1} falls below λ, marking the engineered prompt as less prone to memorization. Both detection and prompt optimization occur only during the initial inference step $T-1$, rather than throughout the entire inference process, enhancing efficiency.

Privacy-utility trade-off. Eq. (10) suffers from a severe privacy-utility trade-off as in Fig. 3. Opting for a lower λ improves the privacy (risk of memorization) of outputs at the expense of utility (text-alignment). It requires more substantial optimization of the input prompt, aiming to achieve a magnitude m_{T-1} lower than λ, thereby enhancing privacy. However, such extensive modifications to the prompt may lead to a loss of its original meaning, compromising utility significantly. Conversely, a higher λ upholds the utility of generated outputs but at the cost of privacy.

Geometric analysis. Fig. 4 (a) visualizes the privacy-utility trade-off in Eq. (10) as a function of λ, where varying levels of λ define distinct contours corresponding to target magnitudes and privacy levels during prompt engineering (PE). Outer contours represent greater prompt modifications, which enhance privacy. This optimization shifts the prompt embedding from the original memorized prompt, e_p, located at point O, to the engineered prompt, e^*, at points B, C, or D. The number of PE iterations determines the distance of the prompt from point O: more iterations lead to greater divergence, achieving higher privacy levels. The process halts once the desired privacy level is attained.

Overview of our improvements. Our trade-off analysis reveals that existing strategies rely solely on prompt modification methods such as prompt engineering to enhance privacy, which is suboptimal and significantly compromises utility. This highlights an urgent need to propose alternative methods to improve privacy and preserve utility. To address this, we introduce two meticulously crafted mitigation strategies that only require refining the guidance components in CFG to achieve a superior privacy-utility trade-off: (1) *prompt re-anchoring (PR)* and (2) *semantic prompt search (SS)*. Specifically, SS enhances utility with minimal privacy cost, while PR serves as an effective alternative to improve privacy, which has the potential of costing less utility when used in combination with SS.

3.2. Enhancing Privacy Through Prompt Re-Anchoring

One of our objectives is to safeguard privacy while preserving as much utility as possible. Recall that in Eq. (10) enhancing privacy requires reducing the magnitude, which correlates with a significant decrease in utility. It is important to note that the utility loss stems not from the reduction in magnitude itself, but from the baseline's inefficient

approach to achieving that reduction. Visually, in Fig. 4, predictions that lie on the same magnitude contour may exhibit notably different levels of utility. This observation motivates us to ponder: Is it possible to reduce magnitude through a more efficient approach (incurring a lower utility cost) compared to the baseline method of prompt engineering? This insight has spurred the creation of our prompt re-anchoring (PR) strategy, which aims to reduce magnitude by steering predictions away from memorized images.

We argue that e_p is discarded too hastily in Eq. (10) when the memorization is detected ($m_{T-1} > \lambda$). Instead, recognizing e_p as a valuable anchor point — its conditional generation can trace the memorization path — we use e_p to re-anchor the generation process, by replacing the unconditional prediction in CFG with the e_p-conditioned prediction to divert the generation away from memorization. Thereby, Eq. (10) is re-written to the following form,

$$\hat{\epsilon} \leftarrow [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1}<\lambda\}} + [\epsilon_\theta(x_t, e_p) + s(\epsilon_\theta(x_t, e^*) - \epsilon_\theta(x_t, e_p))]\mathbb{1}_{\{m_{T-1}>\lambda\}}. \tag{11}$$

Analysis: Enhancing Privacy with Less Utility Cost. The baseline approach in Eq. (10) to enhancing privacy necessitates significant utility sacrifices, primarily due to the need for more extensive gradient updates during prompt optimization to obtain e^*. This leads to substantial modifications of the original user prompt and a large gap between e^* and e_p, causing the prompt to diverge significantly from the user's initial intent and thus undermine the text alignment. In contrast, our prompt re-anchoring strategy employs the intuition from the original CFG approach to guide generations toward a preferred output while distancing them from an undesirable one, effectively achieved through linear interpolation between two generations. Uniquely, our strategy designates the undesired generation as the memorized one that enforces conditioning on the original user prompt e_p, rather than relying on unconditional conditioning e_ϕ. This approach steers generations away from memorized content more efficiently with the contrastive direction, which therefore requires fewer gradient updates to obtain e^* and thus less alternation on the prompt for the same privacy enhancement level. Our previous question is hence answered.

Fig. 4 represents a geometric interpretation of our prompt re-anchoring strategy. Classical CFG directs $\hat{\epsilon}$ away from unconditional (point A) towards conditional predictions (O for e_p, and B, C, D for e^*) as in the left figure. PR, however, guides $\hat{\epsilon}$ from the memorized prediction (O) towards less or unmemorized predictions with lower magnitude (B, C, D), enhancing privacy, as in the middle figure.

Analysis: Continuous Memorization Diversion without Loss of Efficiency. Although [36] found that minimizing the initial magnitude m_{T-1} can indirectly reduce magnitudes in later steps, we observe that the magnitude can still spike and memorization may still occur later in the inference process. For example, as illustrated in Fig. 8 (in Supplementary Material), while the engineered prompt achieves a low magnitude initially, it fails to ensure that this low level is maintained throughout subsequent inference steps, potentially leading back to high magnitudes indicative of Stable Diffusion's memorized outputs. To counteract this, the PR strategy can continuously divert generations away from initially identified memorized conditions (point O) across the entirety of the inference process as shown in Fig. 8 (in Supplementary Material). The effectiveness is also evident from the qualitative appearance of the final generated image, and its quantitative similarity score as in Fig. 8 (in Supplementary Material).

3.3. Enhancing Utility Through Semantic Prompt Search

In this section, we aim to propose a strategy to enhance utility with minimal privacy cost. Recall that low utility is primarily due to the increased difference between the user prompt and the engineered prompt. We suggest conducting semantic search to find a prompt e_p^{ss} semantically similar to the original user prompt e_p but with a lower risk of memorization, aiming to replace the engineered prompt e^* that results in the unsatisfactory trade-off. Eq. (11) is thus further re-formulated to the following form after incorporating semantic prompt search strategy in addition to the prompt re-anchoring strategy,

$$\hat{\epsilon} \leftarrow [\epsilon_\theta(x_t, e_\phi) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\phi))]\mathbb{1}_{\{m_{T-1}<\lambda\}} + [\epsilon_\theta(x_t, e_p) + s(\epsilon_\theta(x_t, e_p^{ss}) - \epsilon_\theta(x_t, e_p))]\mathbb{1}_{\{m_{T-1}>\lambda\}}. \tag{12}$$

In Eq. (12), the engineered prompt e^* from Eq. (10) and Eq. (11) is replaced by e_p^{ss}. The search is conducted in the linguistic space to inherently preserve semantics. We use a large language model (LLM) to find up to n_s alternatives to the user prompt. This search early stops once an alternative prompt e_p^{ss}'s magnitude falls below the λ threshold. Visually in Fig. 4 (c), this is equivalent to reaching the target privacy contour at points B′, C′, or D′. Notably, although B and B′ lie on the same privacy contour, the utility of B′ is often higher than that of B. Implementationally, if all n_s alternatives result in magnitudes above λ, the one with the lowest magnitude will be chosen.

Analysis: Enhancing Utility with No Privacy Cost. This semantic search differs from soft prompt engineering in that the optimization happens in the language space instead of the embedding space. The baseline strategy seeks to boost utility by minimizing the number of prompt optimization steps, which diminishes the extent of prompt modifications, significantly undermining privacy as ostensibly secure prompts may inadvertently produce memorized outputs. Conversely, our approach employs a semantic prompt

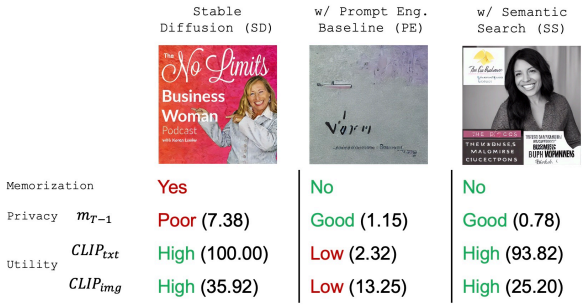


Figure 5. Both the baseline approach and our method effectively inhibit Stable Diffusion from producing memorized images by minimizing the magnitude in the initial inference step for prompt "The No Limits Business Woman Podcast". Nonetheless, our *semantic prompt search (SS)* strategy, which results in "The Empowered Business Woman's Podcast", retains considerably more utility, as demonstrated by the substantially higher CLIP similarities between the original user text prompt and both our (1) modified prompt, and (2) ultimately generated image. Qualitative results further corroborate this finding. m_{T-1} is the first-step magnitude, $CLIP_{txt}$ and $CLIP_{img}$ refer to the original prompt's CLIP similarity with the modified prompt and the generated image.

search to inherently preserve utility, leveraging the capabilities of Large Language Models (LLM) to navigate the linguistic space for suitable alternatives. By pinpointing a substitute prompt in the language space, we keep the semantic similarity and obviate the need to trade privacy for enhanced utility.

Analysis: Enhancing Privacy with Minimum Utility Cost. This is also guaranteed by design. As discussed in Sec. 3.2, the baseline requires significant utility sacrifices to gain extra privacy. In contrast, the utility impact associated with our semantic prompt search is limited to the semantic discrepancy between the original user prompt and the LLM-derived alternative, which is much smaller. Furthermore, when the search yields an alternative prompt devoid of memorization risks, it effectively represents a significant boost in privacy for a minimal cost of utility. Geometrically as in Fig. 4 (c), this means shifting the prediction from O to D′ with minimum utility cost.

Fig. 5 demonstrate SS's superiority over the baseline using a widely-investigated memorized prompt: "The No Limits Business Woman Podcast". While the baseline's prompt engineering (PE) strategy successfully reduces the magnitude to 1.15 and prevents memorization, it results in a significant semantic deviation from the original prompt, as indicated by low CLIP similarity scores for both the engineered text prompt and the generated image. In contrast, our SS strategy not only reduces the magnitude further to 0.78, effectively preventing memorization, but also significantly improves text alignment. By navigating the linguistic space, our method finds "The Empowered Business Woman's Podcast" as a privacy-preserving yet semantically aligned alternative.

3.4. Synergy Effect of the Two Strategies

Analysis: Tackling the Limitation of Semantic Search (SS). In an extreme scenario, SS might encounter a limitation where all n_s searched alternatives result in magnitudes exceeding the set threshold, failing to meet the required privacy standards. This situation can be geometrically represented by a movement from point O to B′ in Fig. 4 (c). However, it's important to note that even in such cases, the transition to a lower magnitude contour, where B′ is located, still offers a certain degree of privacy enhancement without additional cost. Further advancements in privacy (*e.g.*, moving from B′ to D′) can be achieved through our synergistic application of the PR strategy, which incurs a smaller utility cost compared to the baseline method. The right side of Fig. 8 (in Supplementary Material) showcases the synergistic benefit of utilizing the PR strategy to compensate for the SS strategy's limitations. In this example, although SS may not secure an alternative prompt with a sufficiently reduced magnitude, lowering it from the original 7.48 to 6.02 has proved to be adequate for PR to leverage the original prompt as the new anchor point for steering the generation away from memorized content effectively.

Analysis: Addressing the Privacy-Utility Dilemma. For the baseline strategy [36], the dilemma in balancing the privacy-utility trade-off through adjusting the λ hyperparameter manifests in two ways: opting for a higher λ to bolster utility inadvertently raises the False Negative (FN) rate, leading to potentially unsafe prompts being misclassified as safe and thus compromising privacy; conversely, selecting a lower λ to enhance privacy increases the False Positive (FP) rate, resulting in unnecessary modifications to already safe prompts, thereby reducing utility. *Prompt re-anchoring* is tailored to mitigate the privacy repercussions of False Negatives (FN) encountered when users aim to augment utility with a higher λ. It achieves this by actively directing generation away from memorized content, as elucidated in Sec. 3.2 and depicted in Fig. 4 as a divergence from O. This approach thus optimizes the privacy-utility trade-off under such conditions. *Semantic prompt search*, on the other hand, aims to alleviate the utility decline tied to False Positives (FP) when enhancing privacy with a lower λ. It conducts a linguistically sensitive search that avoids semantic degradation, consequently refining the privacy-utility balance in these scenarios. Employing both strategies synergistically ensures an enhanced privacy-utility trade-off, regardless of whether the user's priority lies with maximizing privacy or utility.

4. Experiments

4.1. Evaluation settings

Although our method is designed to enhance the approach in [36], it can be easily adapted to any recent advance-

ment of the detection strategy by simply replacing m_{T-1} in Eq. (12) with alternative detection signals, making PRSS highly practical and adaptable for future improvements. For example, a contemporary work [8] introduced the masked magnitude m'_t (see Eq. (8)), achieving new state-of-the-art detection performance. The mitigation strategy proposed by [8] also uses prompt engineering based on the upgraded detection signal m'_t, offering two key advantages: (1) m'_t provides a more accurate detection trigger for initiating the mitigation strategy, (2) it serves as a more effective loss function than unmasked magnitude, yielding better gradients for optimizing prompts. These improvements make it the current state-of-the-art in mitigation. PRSS can seamlessly integrate this enhanced detection signal by replacing m_{T-1} with m'_{T-1} in Eq. (12), producing our best results to date. For a fair comparison, we re-implement the mitigation strategy from [8] and evaluate its performance alongside PRSS. This comparison not only highlights PRSS's capability to surpass the most recent state-of-the-art but also serves as an ablation study, assessing PRSS's impact since both methods use the same detection strategy to trigger mitigation. Additionally, we produce a version of PRSS that uses m_t as the detection signal, aligning with [36], to further demonstrate PRSS's effectiveness in comparison to [36].

We follow the baselines [8, 36] by experimenting on Stable Diffusion v1-4, using its corresponding prompt dataset from [34], which contains 500 prompts, over 300 of which are classified as memorized and organized by different memorization types. To provide a comprehensive view, we adopt the settings from [8] and report results for both global and local memorization prompts separately. We use the default hyperparameters from [36] to re-implement the detection and prompt engineering mitigation strategies from [36] and [8]. For semantic prompt search with LLMs, we use the GPT-4 model via the OpenAI API, instructing it to generate up to $n_s = 25$ semantically similar prompts for each user prompt. The associated inference time and cost are minimal, with each alternative prompt generated in approximately 0.9 seconds, costing around $0.02.

Following standard practice in memorization literature, we use the CLIP score to measure the alignment between the generated output and the input user prompt, serving as our utility metric. For privacy metrics, which quantify the degree of memorization, most studies rely on SSCD-based scores (see Eq. (5)), reporting metrics such as the average score, the 95th percentile score, and the percentage of memorized generations, with a threshold of 0.5 commonly indicating memorization. Additionally, [8] proposes a localized similarity metric (LS, see Eq. (6)) that measures similarity between the generated image and corresponding training image only within the localized memorization region. We present comparative results using the widely adopted SSCD metric in Fig. 6 and Fig. 7, while other metrics, including

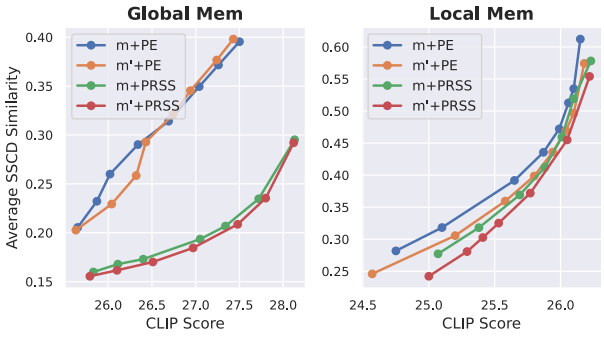


Figure 6. Comparison with baselines. Across different detection signals (m and m'), PRSS consistently achieves superior privacy-utility trade-offs. The improvement is particularly pronounced in global memorization compared to local memorization.

LS, are detailed in the Supplementary Material for a more comprehensive evaluation.

4.2. Comparison with Baselines

As shown in Fig. 6, PRSS consistently outperforms prompt engineering (PE) in achieving better privacy-utility trade-offs, regardless of the detection signal used (whether m or m'). This is visually evident as the red and green lines (PRSS) lie further in the bottom-right direction compared to the yellow and blue lines (PE) for both global and local memorization scenarios. Notably, PRSS demonstrates astonishingly large improvements in global memorization (from the blue line to the green line and from the yellow line to the red line), while [8] achieves only marginal gains over [36] in this case (from the blue line to the yellow line and from the green line to the red line). For local memorization, although the margin of improvement is smaller, PRSS still achieves a significant improvement over PE. Specifically, even using the inferior detection signal m, PRSS (green line) achieves comparable performance with the method proposed by [8] (yellow line), which leverages the superior signal m'. Equipped with m', PRSS can achieve the state-of-the-art results (red line). This highlights PRSS's robust capability to leverage advancements in detection signals.

4.3. Ablation Studies

Fig. 7 highlights the necessity of both components in PRSS, as evidenced by the red line lying further in the bottom-right direction compared to the yellow and green lines for both global and local memorization scenarios. The synergistic effect of combining PR and SS as discussed in Sec. 3.4, is particularly evident in the local memorization scenario, which aligns with [8]'s observation that local memorization poses greater challenges for existing mitigation strategies. Using PR alone effectively enhances privacy but results in suboptimal utility, while SS alone improves utility but achieves limited privacy protection. Together, they

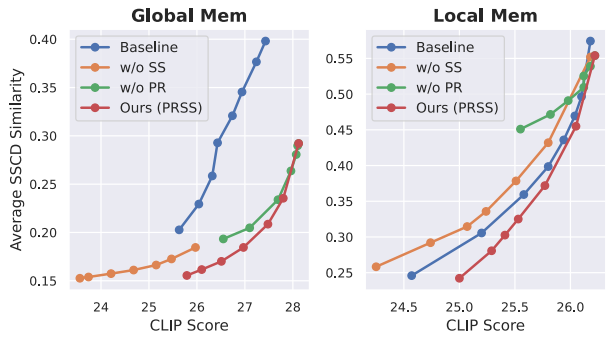


Figure 7. Ablation studies of PRSS components. Using PR alone results in low utility, while SS alone cannot ensure privacy. Combining PR and SS achieves a synergistic effect, delivering the best privacy-utility trade-off.

leverage each other's strengths, achieving the best privacy-utility trade-off. For global memorization, either PR or SS alone significantly outperforms the baseline. However, their combination further enhances the trade-off, underscoring the robustness of PRSS in tackling both global and local memorization challenges.

5. Related Work

Memorization in Generative Models The phenomenon of generative models memorizing training data has been extensively studied across domains, including language models [3, 4, 15, 16, 18, 20], image generation [2, 10, 11, 21, 22, 33], and video generation [9, 26], due to the significant privacy and legal risks it poses [5, 14, 29, 37]. Specifically, in diffusion models, [6, 31, 34] first validated the memorization issue on various models including DDPMs [13, 23] and Stable Diffusion [28] on datasets include Oxford Flowers [24], CelebA [19], CIFAR-10 [17] and LAION [30]. Further analysis by [32] revealed that Stable Diffusion's text-conditioning exacerbates overfitting, making its memorization of LAION particularly problematic. To better understand memorization, [34] categorized the phenomenon as Matching Verbatim, Retrieval Verbatim, and Template Verbatim based on the cause, distinguishing between one-to-one and many-to-many mappings in text-image pairs, while [8] classified memorization into global and local cases, demonstrating the practical significance of this distinction. In this work, we focus on Stable Diffusion due to its widespread use, open-source nature, and real-world legal challenges [29]. We adopt the framework of [8] to analyze both global and local memorization for a more nuanced evaluation of our proposed strategies.

Detection Strategies for Diffusion Memorization. Early detection methods required generating large image sets to estimate density [6] or comparing generations with the entire training dataset [31], which were computationally intensive. Recent approaches [27, 36] introduced more efficient one-step detection strategies. [27] detected memorization by observing less concentrated cross-attention scores for memorized generations than non-memorized ones. [36] proposed using the abnormally high magnitude of text-conditional noise predictions as a detection signal. Building on this, [8] incorporated localized memorization masks into the magnitude computation, offering a more accurate detection signal and achieving state-of-the-art performance. Our proposed PRSS approach leverages these detection mechanisms to trigger mitigation strategies. Notably, PRSS can adapt to any detection method by substituting the indicator variable condition in its framework. We employ the state-of-the-art method from [8] for optimal results.

Mitigation Strategies for Diffusion Memorization. Mitigation strategies have evolved rapidly. Early methods, such as retraining models on de-duplicated data [6], were computationally expensive and showed limited effectiveness. [32] introduced randomness into text prompts or embeddings but indiscriminately affected non-memorized prompts, reducing overall utility. To improve on it, [7] proposed to compute the generation's similarity with the closest training image to engineer the image latent using the classifier guidance framework. However, this relies on computationally expensive nearest-neighbor searches across the training dataset. [36] proposed to conduct prompt engineering using the magnitude of text-conditional prediction as the loss, and [27] proposed to re-weight cross-attention scores for all tokens in the prompt. They offer improved utility-privacy trade-offs and achieve comparable performance. Most recently, [8] extended these efforts by leveraging localized magnitudes as loss functions for prompt engineering, achieving the best results to date. Our investigation reveals the limitations in the prevailing method's privacy-utility balance. In response, we introduce new mitigation strategies that achieve more favorable trade-offs.

6. Conclusion and Future Work

This paper introduces a novel *PRSS* method that consists of *prompt re-anchoring* and *semantic prompt search* as deliberately designed strategies to mitigate memorization in diffusion models. Empirical results show that PRSS demonstrates superior privacy-utility trade-offs over existing methods. Beyond its effectiveness, PRSS is simple and efficient to implement. It requires only minor modifications to the CFG equation to update noise predictions during inference, seamlessly integrating into the inference pipeline without additional training, fine-tuning, or reliance on training data. The efficacy of each component is validated through comprehensive analysis and ablation studies. A current limitation is that the performance of PRSS, as well as baseline mitigation strategies, relies heavily on the accuracy of the detection mechanism. When it fails to trigger accurately, they all revert to the standard Stable Diffusion process, compromising privacy preservation. Thus, improving the detection mechanism's accuracy in future works would orthogonally bolster the effectiveness of PRSS.

7. Acknowledgement

This work was supported in part by the Australian Research Council under Projects DP210101859 and FT230100549.

References

[1] Midjourney. https://www.midjourney.com/. 1

[2] Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do gans learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018. 8

[3] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, 2021. 8

[4] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022. 8

[5] Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramèr. The privacy onion effect: Memorization is relative. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 8

[6] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, pages 5253–5270, 2023. 1, 8

[7] Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *CVPR*, 2024. 1, 3, 8

[8] Chen Chen, Daochang Liu, Mubarak Shah, and Chang Xu. Exploring local memorization in diffusion models via bright ending attention. *ICLR*, 2025. 1, 3, 7, 8

[9] Chen Chen, Enhuai Liu, Daochang Liu, Mubarak Shah, and Chang Xu. Investigating memorization in video diffusion models. *ICLR 2nd Workshop on Navigating and Addressing Data Problems for Foundation Models (DATA-FM)*, 2025. 8

[10] Yunhao Chen, Xingjun Ma, Difan Zou, and Yu-Gang Jiang. Towards a theoretical understanding of memorization in diffusion models. *arXiv preprint arXiv:2410.02467*, 2024. 8

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 8

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 1, 3

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 8

[14] H. H. Jiang, L. Brown, J. Cheng, M. Khan, A. Gupta, D. Workman, A. Hanna, J. Flowers, and T. Gebru. AI Art and Its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 363–374. ACM, 2023. 8

[15] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*, 2022. 8

[16] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, 2022. 8

[17] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 8

[18] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the Association for Computational Linguistics*, 2022. 8

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 8

[20] Fatemehsadat Mireshghallah, Anish Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. Memorization in nlp fine-tuning methods. *arXiv preprint arXiv:2205.12506*, 2022. 8

[21] Amir Naseh, Jaemin Roh, and Amir Houmansadr. Memory triggers: Unveiling memorization in text-to-image generative models through word-level duplication. *arXiv preprint arXiv:2312.03692*, 2023. 8

[22] Ali Naseh, Jaechul Roh, and Amir Houmansadr. Understanding (un)intended memorization in text-to-image generative models. *arXiv preprint arXiv:2312.07550*, 2023. 8

[23] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2, 8

[24] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729, 2008. 8

[25] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *CVPR*, pages 14532–14542, 2022. 3

[26] Aimon Rahman, Malsha V. Perera, and Vishal M. Patel. Frame by familiar frame: Understanding replication in video diffusion models. *arXiv preprint arXiv:2403.19593*, 2024. 8

[27] Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. *ECCV*, 2024. 1, 3, 8

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 8

[29] Joseph Saveri and Butterick Matthew. Stable diffusion litigation, 2023. 2023. 8

[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine

Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. 2022. 1, 8

[31] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *CVPR*, pages 6048–6058, 2023. 1, 3, 8

[32] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *NeurIPS*, 2023. 1, 3, 8

[33] Patrick Tinsley, Adam Czajka, and Patrick Flynn. This face does not exist... but it might be yours! identity leakage in generative models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021. 8

[34] Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023. 1, 7, 8

[35] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In *Advances in Neural Information Processing Systems*, pages 51008–51025. Curran Associates, Inc., 2023. 1

[36] Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *ICLR*, 2024. 1, 3, 5, 6, 7, 8

[37] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 8