Privacy-Preserving Low-Rank Adaptation Against Membership Inference Attacks for Latent Diffusion Models

Zihao Luo[1*], Xilie Xu[2*], Feng Liu[3], Yun Sing Koh[1], Di Wang[4], Jingfeng Zhang[1 4†]

[1]The University of Auckland
[2]The National University of Singapore
[3]The University of Melbourne
[4]King Abdullah University of Science and Technology

Abstract

Low-rank adaptation (LoRA) is an efficient strategy for adapting latent diffusion models (LDMs) on a private dataset to generate specific images by minimizing the adaptation loss. However, the LoRA-adapted LDMs are vulnerable to membership inference (MI) attacks that can judge whether a particular data point belongs to the private dataset, thus leading to the privacy leakage. To defend against MI attacks, we first propose a straightforward solution: Membership-Privacy-preserving LoRA (MP-LoRA). MP-LoRA is formulated as a min-max optimization problem where a proxy attack model is trained by maximizing its MI gain while the LDM is adapted by minimizing the sum of the adaptation loss and the MI gain of the proxy attack model. However, we empirically find that MP-LoRA has the issue of unstable optimization, and theoretically analyze that the potential reason is the unconstrained local smoothness, which impedes the privacy-preserving adaptation. To mitigate this issue, we further propose a Stable Membership-Privacy-preserving LoRA (SMP-LoRA) that adapts the LDM by minimizing the ratio of the adaptation loss to the MI gain. Besides, we theoretically prove that the local smoothness of SMP-LoRA can be constrained by the gradient norm, leading to improved convergence. Our experimental results corroborate that SMP-LoRA can indeed defend against MI attacks and generate high-quality images.

Code —
https://github.com/WilliamLUO0/StablePrivateLoRA
Extended version — https://arxiv.org/abs/2402.11989

1 Introduction

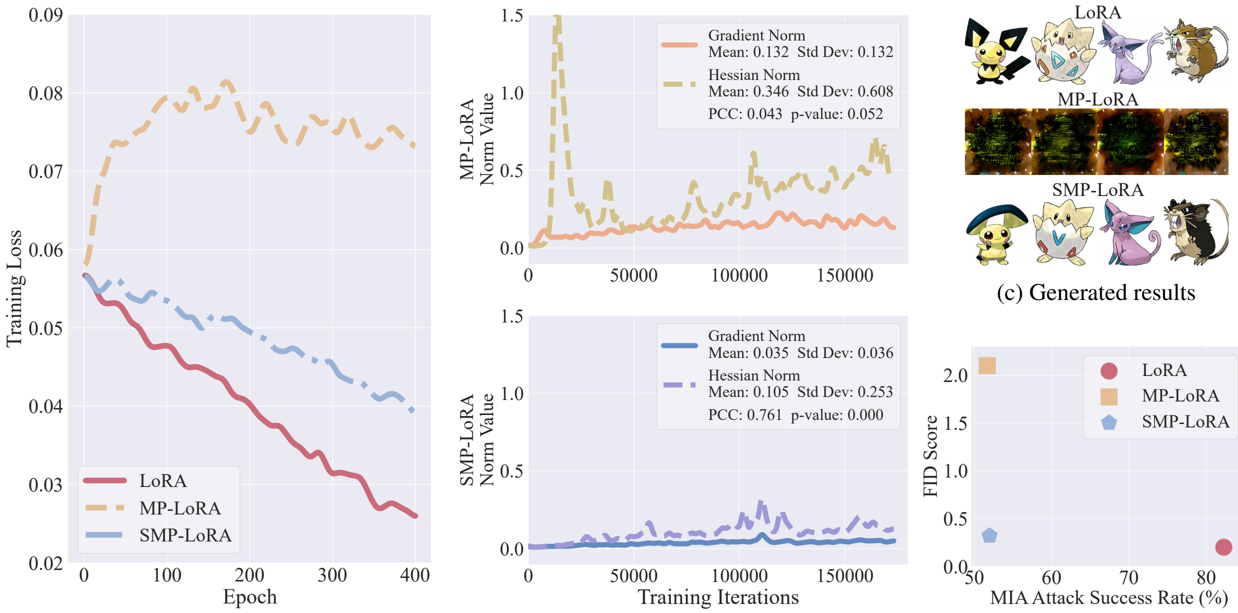
Generative diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021) are leading a revolution in AI-generated content, renowned for their unique generation process and fine-grained image synthesis capabilities. Notably, the Latent Diffusion Model (LDM) (Rombach et al. 2022; Podell et al. 2024) stands out by executing the diffusion process in latent space, enhancing computational efficiency without compromising image quality. Thus, LDMs can be efficiently adapted to generate previously unseen contents or styles (Meng et al. 2022; Gal et al. 2023; Ruiz et al. 2023; Zhang, Rao, and Agrawala 2023), thereby catalyzing a surge across multiple fields, such as facial generation (Huang et al. 2023; Xu et al. 2024) and medicine (Kazerouni et al. 2022; Shavlokhova et al. 2023).

Among various adaptation methods, Low-Rank Adaptation (LoRA) (Hu et al. 2022) is the superior strategy for adapting LDMs by significantly reducing computational resources while ensuring commendable performance with great flexibility. Compared to the full fine-tuning method which fine-tunes all parameters, LoRA optimizes the much smaller low-rank matrices, making the training more efficient and lowering the hardware requirements for adapting LDMs (Hu et al. 2022). By performing the low-rank decomposition of the transformer structure within the LDM, LoRA offers performance comparable to fine-tuning all LDM parameters (Cuenca and Paul 2023). Moreover, LoRA allows flexible sharing of a pre-trained LDM to build numerous small LoRA modules for various tasks.

However, recent studies (Pang and Wang 2023; Pang et al. 2023; Dubiński et al. 2024) have pointed out that adapted LDMs are facing the severe risk of privacy leakage. The leakage primarily manifests in the vulnerability to Membership Inference (MI) attacks (Shokri et al. 2017), which utilize the model's loss of a data point to differentiate whether it is a member of the training dataset or not. As shown in Figure 1d, the LoRA-adapted LDM (red circle marker) exhibits an incredibly high Attack Success Rate (ASR) of 82.27%.

To mitigate the issue of privacy leakage, we make the first effort to propose a Membership-Privacy-preserving LoRA (MP-LoRA) method, which is formulated as a min-max optimization problem. Specifically, in the inner maximization step, a proxy attack model is trained to maximize its effectiveness in inferring membership privacy which is quantitatively referred to as MI gain. In the outer minimization step, the LDM is adapted by minimizing the sum of the adaptation loss and the MI gain of the proxy attack model to enhance the preservation of membership privacy.

---
[*]These authors contributed equally.
[†]Correspondence to: Jingfeng Zhang <jingfeng.zhang@auckland.ac.nz>.



(a) Training loss (b) Gradient scale (c) Generated results (d) Performance

Figure 1: Figure 1a shows the trajectory of the training loss during the adaptation process via LoRA, MP-LoRA, and SMP-LoRA on the Pokemon dataset. Figure 1b displays the mean and standard deviation of the gradient norms and Hessian norms for MP-LoRA and SMP-LoRA throughout the training iterations. It also presents the Pearson correlation coefficients (PCC) and p-values assessing their correlation. Note that each epoch contains 433 training iterations. Figures 1c and 1d demonstrate the generated images and a comparison of evaluation metrics including FID Score and MI attack success rate (ASR). MP-LoRA preserves membership privacy but compromises image generation capability. In contrast, SMP-LoRA effectively preserves membership privacy while maintaining the quality of the generated image, demonstrating its effectiveness in defending against MI attacks without significant loss of functionality. Extensive generated images are visualized in Appendix H of the extended version (Luo et al. 2024).

However, the vanilla MP-LoRA encounters an issue of effective optimization of the training loss, as evidenced in the orange dashed line of Figure 1a. We theoretically find that during MP-LoRA, the local smoothness, quantified by the Hessian norm (the norm of the Hessian matrix) (Bubeck et al. 2015), is independent of and not bounded by the gradient norm (see Proposition 1 for details). This independence hinders the privacy-preserving adaptation of the MP-LoRA (Zhang et al. 2019), thus impeding optimizing the training loss. Besides, we empirically show that the correlation between the Hessian norm and the gradient norm during MP-LoRA is insignificant. This is manifested by the Pearson Correlation Coefficient (PCC) of 0.043 and the p-value above 0.05, as shown in the upper panel of Figure 1b, which corroborates our theoretical analyses.

To stabilize the optimization procedure of MP-LoRA, we further propose a Stable Membership-Privacy-preserving LoRA (SMP-LoRA) method, which incorporates the MI gain into the denominator of the adaptation loss instead of directly summing it. We theoretically demonstrate that this modification ensures a positive correlation (see Proposition 2 for details). Specifically, the local smoothness (that is quantified by the Hessian norm) is positively correlated with and upper bounded by the gradient norm during adaptation, which can improve convergence. Furthermore, we empirically corroborate that during SMP-LoRA, the Hessian norm is positively correlated with the gradient norm, as evidenced by the higher PCC (0.761) and the p-value of less than 0.001 in the lower panel of Figure 1b. The constrained local smoothness allows the SMP-LoRA to achieve better optimization, as shown in the blue dash-dot line of Figure 1a.

To evaluate the performance of the SMP-LoRA, we conducted adapting experiments using the Stable Diffusion v1.5 (CompVis 2022) on the Pokemon (Pinkney 2022) and CelebA (Liu et al. 2015) datasets, respectively. Figure 1d shows that, although MP-LoRA (orange square marker) lowers the ASR to near-random levels, it significantly degrades the image generation capability of LoRA, as evidenced by a high FID score of 2.10 and the poor visual quality in Figure 1c. In contrast, the SMP-LoRA (blue pentagon marker) effectively preserves membership privacy without sacrificing generated image quality significantly, as evidenced by its FID score of 0.32 and ASR of 51.97%.

2 Membership-Privacy-Preserving LoRA

In this section, we first use the min-max optimization to formulate the learning objective of MP-LoRA. Then, we disclose the issue of unstable optimization of MP-LoRA. Finally, we propose the stable SMP-LoRA and its implementation.

A Vanilla Solution: MP-LoRA

Objective function. In MI attack, the conflicting objectives of defenders and adversaries can be modelled as a privacy game (Shokri et al. 2012; Manshaei et al. 2013; Alvim et al. 2017). Adversaries can adjust their attack models to maximize MI gain against the target model, which requires that the defense can anticipate and withstand the strongest inference attacks. Consequently, the defender's goal is to enhance the preservation of membership privacy in worst-case scenarios where the adversary achieves the maximum MI gain while maintaining the model performance. Inspired by Nasr, Shokri, and Houmansadr (2018), we propose MP-LoRA to defend against MI attacks which is formulated as a min-max optimization problem as follows:

$$\min_{\{\mathbf{B},\mathbf{A}\}} \left(\underbrace{\mathcal{L}_{\mathrm{ada}}(f_{\bar{\theta}+\mathbf{BA}}, \mathcal{D}_{\mathrm{tr}})}_{\text{Adaptation loss}} + \lambda \underbrace{\max_{\omega} G\left(h_\omega, \mathcal{D}_{\mathrm{aux}}, f_{\bar{\theta}+\mathbf{BA}}\right)}_{\text{Membership inference gain}} \right), \tag{1}$$

where $\mathcal{L}_{\mathrm{ada}}(f_{\bar{\theta}+\mathbf{BA}}, \mathcal{D}_{\mathrm{tr}})$ refers to the adaptation loss for the LDM with LoRA module $f_{\bar{\theta}+\mathbf{BA}}$ on the training dataset $\mathcal{D}_{\mathrm{tr}}$, h_ω is the proxy attack model parameterized by ω, $G\left(h_\omega, \mathcal{D}_{\mathrm{aux}}, f_{\bar{\theta}+\mathbf{BA}}\right)$ represents the MI gain of the proxy attack model h_ω on the auxiliary dataset $\mathcal{D}_{\mathrm{aux}}$.

Therein, the inner maximization aims to search for the most effective proxy attack model h_ω for a given adapted LDM $f_{\bar{\theta}+\mathbf{BA}}$ via maximizing the MI gain. The outer minimization, conversely, searches for the LDM $f_{\bar{\theta}+\mathbf{BA}}$ that can best preserve membership privacy under the strong proxy attack model h_ω while being able to adapt on the training dataset.

Updating the proxy attack model in inner maximization. The proxy attack model h_ω equipped with white-box access to the target LDM $f_{\bar{\theta}+\mathbf{BA}}$, aims to infer whether a specific image-text pair (x, y) is from the training dataset $\mathcal{D}_{\mathrm{tr}}$ for adapting the target LDM $f_{\bar{\theta}+\mathbf{BA}}$. The model achieves this by constructing an auxiliary dataset $\mathcal{D}_{\mathrm{aux}}$, which consists of half of the member data from $\mathcal{D}_{\mathrm{tr}}$, denoted as $\mathcal{D}_{\mathrm{aux}}^{\mathrm{m}}$, and an equal amount of local non-member data $\mathcal{D}_{\mathrm{aux}}^{\mathrm{nm}}$. Using the auxiliary dataset $\mathcal{D}_{\mathrm{aux}}$, h_ω trains a binary classifier based on the adaptation loss of the target LDM $f_{\bar{\theta}+\mathbf{BA}}$ to predict the probability of (x, y) for being a member of the $\mathcal{D}_{\mathrm{tr}}$. Consequently, the MI gain of h_ω can be quantified based on its performance on the $\mathcal{D}_{\mathrm{aux}}$ as follows:

$$\begin{aligned} &G\left(h_\omega, \mathcal{D}_{\mathrm{aux}}, f_{\bar{\theta}+\mathbf{BA}}\right) = \\ &\frac{1}{2\left|\mathcal{D}_{\mathrm{aux}}^{\mathrm{m}}\right|} \sum_{(x,y)\in\mathcal{D}_{\mathrm{aux}}^{\mathrm{m}}} \log\left(h_\omega\left(\ell_{\mathrm{ada}}\left(x, y; f_{\bar{\theta}+\mathbf{BA}}\right)\right)\right) \\ &+ \frac{1}{2\left|\mathcal{D}_{\mathrm{aux}}^{\mathrm{nm}}\right|} \sum_{(x,y)\in\mathcal{D}_{\mathrm{aux}}^{\mathrm{nm}}} \log\left(1 - h_\omega\left(\ell_{\mathrm{ada}}\left(x, y; f_{\bar{\theta}+\mathbf{BA}}\right)\right)\right). \end{aligned} \tag{2}$$

In the inner maximization, the proxy attack model optimizes the parameters ω by maximizing the MI gain, i.e., $\max_{\omega} G\left(h_\omega, \mathcal{D}_{\mathrm{aux}}, f_{\bar{\theta}+\mathbf{BA}}\right)$.

Adapting the LDM in outer minimization. MP-LoRA optimizes the LDM by directly minimizing a weighted sum of the MI gain for the h_ω and the adaptation loss, which enables it to adapt to the training data and protect the private information of the training dataset simultaneously. To be specific, the training loss of MP-LoRA is formulated as

$$\mathcal{L}_{\mathrm{PL}} = \mathcal{L}_{\mathrm{ada}}(f_{\bar{\theta}+\mathbf{BA}}, \mathcal{D}_{\mathrm{tr}}) + \lambda \cdot G(h_\omega, \mathcal{D}_{\mathrm{tr}}, f_{\bar{\theta}+\mathbf{BA}}), \tag{3}$$

where $\lambda \in \mathbb{R}$ controls the importance of optimizing the adaptation loss versus protecting membership privacy. In the outer minimization of MP-LoRA, the parameters $\mathbf{B}$ and $\mathbf{A}$ is updated by minimizing the $\mathcal{L}_{\mathrm{PL}}$, i.e., $\min_{\{\mathbf{B},\mathbf{A}\}} \mathcal{L}_{\mathrm{PL}}$.

MP-LoRA is realized by one step of inner maximization to obtain a power proxy attack model by maximizing the MI gain in Equation (2) and one step of outer minimization to update $\mathbf{A}$ and $\mathbf{B}$ by minimizing the training loss in Equation (3). The algorithm of MP-LoRA is shown in Algorithm 2 (Appendix A of the extended version (Luo et al. 2024)).

Unstable Issue of MP-LoRA

In this subsection, we theoretically demonstrate that the convergence for MP-LoRA cannot be guaranteed due to unconstrained local smoothness. Then we validate the theoretical analyses with empirical evidence.

Definition 1 (Relaxed Smoothness Condition from Zhang et al. (2019))**.** *A second order differentiable function f is $(L_0, L1)$-smooth if*

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|. \tag{4}$$

Lemma 1 (Zhang et al. (2019))**.** *Let f be a second-order differentiable function and $(L_0, L1)$-smooth. If the local smoothness, quantified by the Hessian norm (the norm of the Hessian matrix), is positively correlated with the gradient norm (i.e., $L_1 > 0$), then the gradient norm upper bounds the local smoothness, facilitating faster convergence and increasing the likelihood of converging to an optimal solution.*

Proposition 1. *MP-LoRA does not satisfy the positive correlation as described in Lemma 1, therefore the convergence cannot be guaranteed and the model may settle at a suboptimal solution.*

Proof. We establish the Relaxed Smoothness Condition for MP-LoRA as follows:

$$\begin{aligned} &\|\frac{\partial^2 \mathcal{L}_{\mathrm{PL}}}{\partial \mathbf{BA}^2}\| \leq L_0 + L_1 \|\frac{\partial \mathcal{L}_{\mathrm{PL}}}{\partial \mathbf{BA}}\|, \\ &\text{where } L_0 = \|\frac{\partial^2 \mathcal{L}_{\mathrm{ada}}}{\partial \mathbf{BA}^2}\| + \lambda \|\frac{\partial^2 G}{\partial \mathbf{BA}^2}\|,\ L_1 = 0, \end{aligned} \tag{5}$$

in which $\mathcal{L}_{\mathrm{ada}}$ represents the adaptation loss and G represents the MI gain. The detailed derivation is presented in Appendix B of the extended version (Luo et al. 2024). The value of L_1 being zero indicates that the Hessian norm is independent of and not bounded by the gradient norm, suggesting that the local smoothness is unconstrained. □

Next, we provide empirical evidence to support our theoretical analyses. We tracked the gradient norm and the Hessian norm of the training loss at each training iteration, and calculated their Pearson Correlation coefficient (PCC) and p-value as shown in Figure 1b. The details for calculating the gradient norm and the Hessian norm can be found in Appendix C. In Figure 1b, the low PPC of 0.043 for MP-LoRA suggests a very weak correlation between the Hessian norm and the gradient norm. Additionally, with the p-value of 0.052, there is insufficient evidence to reject the hypothesis of no correlation. This indicates that the Hessian norm is unbounded, implying that the local smoothness, quantified by the Hessian norm (Bubeck et al. 2015), is unconstrained. Such unconstrained local smoothness leads to the unstable optimization issue in MP-LoRA, and even to the failure of adaptation, as evidenced in the orange dashed line of Figure 1a and the poor visual quality of the generated images in Figure 1c.

Stabilizing MP-LoRA

To mitigate the aforementioned optimization issue of MP-LoRA, we propose SMP-LoRA by incorporating the MI gain into the denominator of the adaptation loss. The objective function of SMP-LoRA is formulated as follows:

$$\min_{\{\mathbf{B},\mathbf{A}\}} \left(\frac{\mathcal{L}_{\text{ada}}(f_{\bar{\theta}+\mathbf{BA}}, \mathcal{D}_{\text{tr}})}{1 - \lambda \max_{\omega} G\left(h_\omega, \mathcal{D}_{\text{aux}}, f_{\bar{\theta}+\mathbf{BA}}\right)} \right). \tag{6}$$

To optimize Equation (6), SMP-LoRA targets to minimize the following training loss function, i.e.,

$$\mathcal{L}_{\text{SPL}} = \frac{\mathcal{L}_{\text{ada}}(f_{\bar{\theta}+\mathbf{BA}}, \mathcal{D}_{\text{tr}})}{1 - \lambda \cdot G\left(h_\omega, \mathcal{D}_{\text{tr}}, f_{\bar{\theta}+\mathbf{BA}}\right) + \delta}, \tag{7}$$

where δ is a stabilizer with a small value such as $1e-5$. This prevents the denominator from approaching zero and ensures stable calculation.

The implementation of SMP-LoRA is detailed in Algorithm 1. At each training step, SMP-LoRA will first update the proxy attack model by maximizing the MI gain and then update the LDM by minimizing the training loss $\mathcal{L}_{\text{SPL}}$.

Proposition 2. *SMP-LoRA satisfies the positive correlation as described in Lemma 1, thus promoting faster convergence, and the model is more likely to converge to an optimal solution.*

Proof. We establish the Relaxed Smoothness Condition for SMP-LoRA as follows:

$$\begin{aligned}
&\|\frac{\partial^2 \mathcal{L}_{\text{SPL}}}{\partial \mathbf{BA}^2}\| \leq L'_0 + L'_1 \|\frac{\partial \mathcal{L}_{\text{SPL}}}{\partial \mathbf{BA}}\|,\\
&\text{where } \mu = \frac{\partial \mathcal{L}_{\text{ada}}}{\partial \mathbf{BA}},\ \nu = \lambda \frac{\partial G}{\partial \mathbf{BA}},\\
&L'_0 = \frac{1}{1-\lambda G + \delta} \cdot \|\frac{\partial^2 \mathcal{L}_{\text{ada}}}{\partial \mathbf{BA}^2}\| + \frac{\lambda \mathcal{L}_{\text{ada}}}{(1-\lambda G+\delta)^2} \cdot \|\frac{\partial G^2}{\partial \mathbf{BA}^2}\|,\\
&L'_1 = \frac{2\|\nu\|}{1-\lambda G+\delta}.
\end{aligned} \tag{8}$$

Please refer to Appendix B of the extended version (Luo et al. 2024) for detailed derivation. The value of L'_1 being greater than zero indicates that the Hessian norm is positively correlated with and upper bounded by the gradient norm, suggesting that the gradient norm constrains the local smoothness during adaptation. □

Algorithm 1: Stable Membership-Privacy-preserving LoRA

Input: Training dataset $\mathcal{D}_{\text{tr}}$ for adaptation process, Auxiliary dataset $\mathcal{D}_{\text{aux}} = \mathcal{D}^{\text{m}}_{\text{aux}} \cup \mathcal{D}^{\text{nm}}_{\text{aux}}$, a pre-trained LDM f_θ, a proxy attack model h_ω parameterized by ω, learning rate η_1 and η_2
Output: a SMP-LoRA for LDMs

```
Perform low-rank decomposition on f_θ to obtain f_{θ̄+BA}
(B and A are trainable LoRA modules)
for each epoch do
    for each training iteration do
        Sample batches S^m and S^nm from D^m_aux and D^nm_aux
        Calculate the MI gain G* on S^m ∪ S^nm
        Update the parameters ω ← ω + η_1 · ∇_ω G*.
        Sample a fresh batch from D_tr
        Calculate the training loss L* = L_SPL
        Update parameters A ← A − η_2 · ∇_A L* and B ←
        B − η_2 · ∇_B L*, respectively
    end for
end for
```

Subsequently, we further corroborate our theoretical analyses with the following empirical evidence. Compared to MP-LoRA's insignificant correlation, SMP-LoRA demonstrates a strong positive correlation between the Hessian norm and the gradient norm, evidenced by the PCC of 0.761 and the p-value less than 0.001 in the lower panel of Figure 1b. This indicates that the Hessian norm, which represents the local smoothness, is upper bounded by the gradient norm, resulting in lower mean (0.105) and standard deviation (0.253) than MP-LoRA. Consequently, the constrained local smoothness mitigates the issue of unstable optimization and enables the SMP-LoRA to converge to a more optimal solution, as demonstrated by the progressively decreasing training loss shown in the blue dash-dot line of Figure 1a and the superior performance on both FID and ASR metrics illustrated by the blue pentagon marker in Figure 1d.

Notably, SMP-LoRA also exhibits lower mean and standard deviation of the gradient norm compared to MP-LoRA. We provide further empirical analysis in Appendix D of the extended version (Luo et al. 2024).

3 Conclusion

In this paper, we proposed Membership-Privacy-preserving LoRA (MP-LoRA), a method based on low-rank adaptation (LoRA) for adapting latent diffusion models (LDMs), while mitigating the risk of privacy leakage. We first highlighted the unstable issue in MP-LoRA. Directly minimizing the sum of the adaptation loss and MI gain can lead to unconstrained local smoothness, which results in unstable optimization. To mitigate this issue, we further proposed a Stable Membership-Privacy-preserving LoRA (SMP-LoRA) method, which constrains the local smoothness through the gradient norm to improve convergence. Detailed theoretical analyses and comprehensive empirical results demonstrate that the SMP-LoRA can effectively preserve membership privacy against MI attacks and generate high-quality images.

Acknowledgments

Feng Liu is supported by the Australian Research Council (ARC) with grant numbers DP230101540 and DE240101089, and the NSF&CSIRO Responsible AI program with grant number 2303037.

Di Wang is supported in part by the funding BAS/1/1689-01-01, URF/1/4663-01-01, REI/1/5232-01-01, REI/1/5332-01-01, and URF/1/5508-01-01 from KAUST, and funding from KAUST - Center of Excellence for Generative AI, under award number 5940.

References

Alvim, M. S.; Chatzikokolakis, K.; Kawamoto, Y.; and Palamidessi, C. 2017. Information leakage games. In *Decision and Game Theory for Security: 8th International Conference, GameSec 2017, Vienna, Austria, October 23-25, 2017, Proceedings*, 437–457. Springer.

Bubeck, S.; et al. 2015. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4): 231–357.

CompVis. 2022. Stable Diffusion. https://github.com/CompVis/stable-diffusion. Accessed on January 16, 2024.

Cuenca, P.; and Paul, S. 2023. Using LoRA for Efficient Stable Diffusion Fine-Tuning. Accessed on January 16, 2024.

Dubiński, J.; Kowalczuk, A.; Pawlak, S.; Rokita, P.; Trzciński, T.; and Morawiecki, P. 2024. Towards More Realistic Membership Inference Attacks on Large Diffusion Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4860–4869.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, Z.; Chan, K. C.; Jiang, Y.; and Liu, Z. 2023. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6080–6090.

Kazerouni, A.; Aghdam, E. K.; Heidari, M.; Azad, R.; Fayyaz, M.; Hacihaliloglu, I.; and Merhof, D. 2022. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Luo, Z.; Xu, X.; Liu, F.; Koh, Y. S.; Wang, D.; and Zhang, J. 2024. Privacy-Preserving Low-Rank Adaptation for Latent Diffusion Models. *arXiv preprint arXiv:2402.11989*.

Manshaei, M. H.; Zhu, Q.; Alpcan, T.; Bacşar, T.; and Hubaux, J.-P. 2013. Game theory meets network security and privacy. *ACM Computing Surveys (CSUR)*, 45(3): 1–39.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 634–646.

Pang, Y.; and Wang, T. 2023. Black-box Membership Inference Attacks against Fine-tuned Diffusion Models. *arXiv preprint arXiv:2312.08207*.

Pang, Y.; Wang, T.; Kang, X.; Huai, M.; and Zhang, Y. 2023. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*.

Pinkney, J. N. M. 2022. Pokemon BLIP captions. https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/. Accessed on January 16, 2024.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Shavlokhova, V.; Vollmer, A.; Zouboulis, C. C.; Vollmer, M.; Wollborn, J.; Lang, G.; Kübler, A.; Hartmann, S.; Stoll, C.; Roider, E.; et al. 2023. Finetuning of GLIDE stable diffusion model for AI-based text-conditional image synthesis of dermoscopic images. *Frontiers in Medicine*, 10.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Shokri, R.; Theodorakopoulos, G.; Troncoso, C.; Hubaux, J.-P.; and Le Boudec, J.-Y. 2012. Protecting location privacy: optimal strategy against localization attacks. In *Proceedings of the 2012 ACM conference on Computer and communications security*, 617–627.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Xu, J.; Motamed, S.; Vaddamanu, P.; Wu, C. H.; Haene, C.; Bazin, J.-C.; and De la Torre, F. 2024. Personalized Face

Inpainting with Diffusion Models by Parallel Visual Attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5432–5442.

Zhang, J.; He, T.; Sra, S.; and Jadbabaie, A. 2019. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In *International Conference on Learning Representations*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.