Diffusion Unlearning Optimization for Robust and Safe Text-to-Image Models

Yong-Hyun Park [1] Sangdoo Yun [2 3] Jin-Hwa Kim [2 3] Junho Kim [2] Geonhui Jang [4] Yonghyun Jeong [5 6]
Junghyo Jo [1] Gayoung Lee [2]

1. Introduction

Recently, as the performance of text-to-image models (Rombach et al., 2022; Ho et al., 2020) has significantly improved, there have been many concerns about their negative social impact. For example, these models can be used to generate explicit or violent images, and often create copyrighted images. Blocking inappropriate content with classifiers is one available approach, but attackers can bypass this by using the publicly available model weights. This poses a risk for many services and companies that publish their model weights, ultimately hindering the advancement of T2I model research.

To solve this problem, recent studies (Gandikota et al., 2023; Kumari et al., 2023; Zhang et al., 2023a; Heng & Soh, 2024; Gandikota et al., 2024) have aimed to remove unwanted concepts from the models. While removing target concepts, it is desired that the performance on non-target concepts remains as close to the original model as possible. Existing studies mainly adopted methods to block the flow of prompts containing unsafe keywords within the model. Since the blocking is only applied to specific prompts, it has the advantage of preserving non-target concepts. However, they have the disadvantage of being vulnerable to adversarial prompt attacks as shown in Figure 1. Recent studies (Tsai et al., 2023; Pham et al., 2023; Yang et al., 2024) have shown that adversarial attacks using prompts are possible even in black-box scenarios. This demonstrates that visual features themselves need to be unlearned to prevent vulnerability to such prompt attacks.

We propose a method to prevent the model from creating unsafe visual features regardless of the prompt. This differs from existing methods that shallowly block information conveyed from the prompt. The biggest challenge is that unlearning visual features may reduce the image generation quality for unrelated topics. To prevent this, we propose a method to precisely guide the model to forget only the unwanted concepts. First, we used SDEdit to generate paired ground truth images that remove the unsafe concepts from images containing these concepts. Using paired data for supervised learning is common in prompt-based unlearning methods. However, since we aim for the model to not generate the unsafe visual features regardless of the prompt, supervised learning is not suitable. Therefore, we used the Direct Preference Optimization (DPO) method to guide the model to prefer generating the paired ground truth images over the images containing unsafe concepts.

We demonstrate that using paired data for preference optimization is effective in selectively unlearning visual features. To this end, we show that our method is robust against adversarial prompt attacks, which existing prompt-based unlearning methods are vulnerable to.

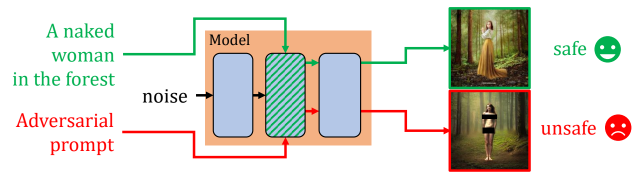

Figure 1. Task Visualization: Existing prompt-based unlearning methods primarily focused on unlearning the prompt's embedding or the dependent cross-attention layers. While these methods do not compromise the quality of unrelated topic images, they have the drawback of being vulnerable to adversarial prompts.

2. Related Work

Recently, there has been active research on safety mechanisms to prevent text-to-image models from generating images with unwanted concepts. One prominent approach is fine-tuning-based unlearning, which is advantageous as it avoids the need for training from scratch. Notable works like ESD (Gandikota et al., 2023), CA (Kumari et al., 2023), UCE (Gandikota et al., 2024), Forget-me-not (Zhang et al., 2023a), and SA (Heng & Soh, 2024) have developed methods to handle unsafe prompts during training.

[1]Department of Physics Education, Seoul National University [2]NAVER AI Lab [3]AI Institute of Seoul National University or SNU AIIS [4]School of Industrial and Management Engineering, Korea University [5]NAVER Cloud [6]Korea Institute for Advanced Study (KIAS). Correspondence to: Junghyo Jo <jojunghyo@snu.ac.kr>, Gayoung Lee <gayoung.lee@navercorp.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Figure 2. To guide the model to unlearn only the desired concept without losing the ability to generate unrelated concepts, we generate synthetic paired data using SDEdit. We use for publication purposes.

However, these strategies are often susceptible to adversarial prompt attacks (Tsai et al., 2023; Pham et al., 2023; Yang et al., 2024; Chin et al., 2023; Zhang et al., 2023b; Han et al., 2024). Our research aims to develop a robust safety mechanism that can withstand red teaming efforts.

3. Method

We aim to remove visual features associated with unsafe concepts from the model. Although we experimented with diffusion models, we believe that this approach can be broadly applied to other image synthesis methods as well. The loss function of the diffusion model (Ho et al., 2020) is commonly expressed as follows:

$$L_{\text{DSM}} = \mathbb{E}_{x_0 \sim q(x_0), x_t \sim q(x_t|x_0)}[||\epsilon - \epsilon_\theta(x_t)||_2^2] \quad (1)$$

Where x_0 is an image and x_t is a noisy image sampled from $q(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, \sqrt{1-\alpha_t}I)$. $\epsilon_\theta(\cdot)$ is the model that predicts the added noise ϵ.

Direct Preference Optimization (DPO) (Rafailov et al., 2024), which is a type of preference optimization, has been studied for application to diffusion models in previous research (Wallace et al., 2023). The final equation in this paper is summarized as follows. For detailed derivation, please refer to the referenced paper.

$$L_{\text{Diffusion-DPO}} \leq -\mathbb{E}[\log \sigma(-\beta T \omega(\lambda_t)(d_{pref} - d_{dispref}))] \quad (2)$$

$$d_{pref} = \|\epsilon - \epsilon_\theta(x_t^+, t)\|_2^2 - \|\epsilon - \epsilon_\phi(x_t^+, t)\|_2^2 \quad (3)$$

$$d_{dispref} = \left(\|\epsilon - \epsilon_\theta(x_t^-, t)\|_2^2 + \|\epsilon - \epsilon_\phi(x_t^-, t)\|_2^2\right) \quad (4)$$

Where x_t^+ is a prefered noisy image and x_t^- is an unprefered noisy image. ϵ_ϕ denotes the pretrained model and ϵ_θ is the fine-tuned model.

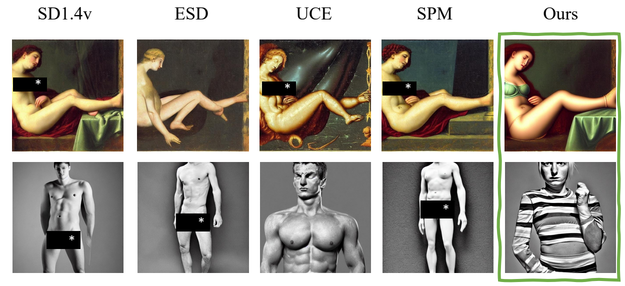

Figure 3. Qualitative results on nudity attacked by Ring-A-Bell method. We use for publication purposes.

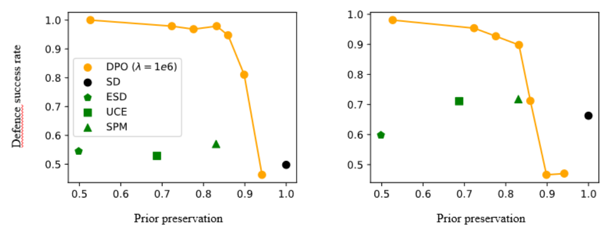

Figure 4. Quantitative results on nudity. With same prior preservation score, our method is more robust on (left) Ring-A-Bell and (right) Concept Inversion.

By using this method, we can increase the probability that the diffusion model generates preferred images while gradually decreasing the probability of generating dispreferred images. We apply this to the unlearning task by replacing dispreferred images with unsafe images and preferred images with safe images. To generate these safe and unsafe pairs, we used SDEdit (Meng et al., 2021) to create synthetic paired data, as shown in Figure 2. Additionally, we discovered a trick to help the model more reliably maintain its prior. This involves ensuring that the noise predicted by the two models is similar for complete noise. We added the following term to the loss function for optimization.

$$L_{\text{prior}} = ||\epsilon_\phi(x_T) - \epsilon_\theta(x_T)||_2^2 \quad (5)$$

4. Experiments

We conducted unlearning for nudity and applied two types of red-teaming attacks: Ring-A-Bell (Tsai et al., 2023) and Concept Inversion (Pham et al., 2023). To generate the dataset, we used prompts containing ‘naked’ as unsafe prompts and replaced them with prompts containing ‘dressed’ to create paired sets. Using these, we generated 64 pairs of images for the unlearning training.

Compared to ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), and SPM (Lyu et al., 2023) unlearning methods, our approach demonstrated significantly more robustness against prompt attacks. We presented qualitative results in

Figure 3. For quantitative evaluation, we used NudeNet (Bedapudi, 2019) to determine the defense success rate and calculated LPIPS score (Zhang et al., 2018) for unrelated topics as a prior preservation score. The results are presented in Figure 4.

References

Bedapudi, P. Nudenet: lightweight nudity detection. https://github.com/notAI-tech/NudeNet/, 2019.

Chin, Z.-Y., Jiang, C.-M., Huang, C.-C., Chen, P.-Y., and Chiu, W.-C. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, October 2023.

Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., and Bau, D. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024.

Han, X., Yang, S., Wang, W., Li, Y., and Dong, J. Probing unlearned diffusion models: A transferable adversarial attack perspective. *arXiv preprint arXiv:2404.19382*, 2024.

Heng, A. and Soh, H. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., and Ding, G. One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications. *arXiv preprint arXiv:2312.16145*, 2023.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Pham, M., Marshall, K. O., Cohen, N., Mittal, G., and Hegde, C. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Tsai, Y.-L., Hsu, C.-Y., Xie, C., Lin, C.-H., Chen, J.-Y., Li, B., Chen, P.-Y., Yu, C.-M., and Huang, C.-Y. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023.

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., and Naik, N. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

Yang, Y., Hui, B., Yuan, H., Gong, N., and Cao, Y. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 123–123. IEEE Computer Society, 2024.

Zhang, E., Wang, K., Xu, X., Wang, Z., and Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023a.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., and Liu, S. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv preprint arXiv:2310.11868*, 2023b.