ZIUM: Zero-Shot Intent-Aware Adversarial Attack on Unlearned Models

Supplementary Material

A. Appendix

A1. Further Visual Comparison of generated images

For further visual comparison of ZIUM's attack performance with existing adversarial methods, we present various examples of generated images.

Fig. 1 illustrates examples of images generated by vanilla Stable Diffusion [3] without any MU applied, and images generated by each of the adversarial attack methods against the ESD [2] model under Van Gogh unlearned concept scenario.

For the Van Gogh unlearned concept scenario, the images generated by UnlearnDiffAtk [7] and P4D [1] both included a flower figure, resembling the image generated by vanilla Stable Diffusion. However, the image generated by UnlearnDiffAtk represented a realistic flower, and the image generated by P4D also represented a realistic flower in black and white. Ring-A-Bell [5], in particular, failed to depict the flower figure at all. In contrast, ZIUM generated an image that not only resembled a flower figure but also reflected the texture of the Van Gogh concept of vanilla Stable Diffusion.

Fig. 2 illustrates examples of images generated by vanilla Stable Diffusion without any MU applied, and images generated by each of the adversarial attack methods against the ESD [2] model under church and parachute unlearned concept scenarios.

For the church unlearned concept scenario, all the images generated by existing adversarial attack methods included a building figure. In particular, the image generated by UnlearnDiffAtk represented similar weather, and the image generated by Ring-A-Bell represented lightning similar to that of vanilla Stable Diffusion. However, they all failed to fully depict a church. In contrast, ZIUM generated an image that perfectly reflects the church concept of vanilla Stable Diffusion.

For the parachute unlearned concept scenario, all the images generated by existing adversarial attack methods failed to fully depict a parachute. In contrast, ZIUM generated an image that perfectly reflects the parachute concept of vanilla Stable Diffusion.

To evaluate the superior attack performance of ZIUM, we further assessed the generated images by each of the adversarial attack methods against the FMN [6] and SLD [4]. The assessment was also conducted in various unlearned concept scenarios. Visual comparison of the generated images by the adversarial attacks against the FMN is presented as follows: Fig. 3, Fig. 4, and Fig. 5. Also, visual comparison of the generated images by the adversarial attacks against the SLD is presented in Fig. 6.

A2. Further Evaluation of ZIUM's Customization effectiveness using user-intent prompts

To evaluate ZIUM's customization effectiveness using user-intent prompts, we analyzed the change in generated images based on ZIUM's user-intent prompt. Fig. 7 and Fig. 8 present the generated images without a user-intent prompt (Prompt: None) and the images reflecting the attacker's intent through three different user-intent prompts for three unlearned concepts (nudity, church, and violence).

Fig. 7(a) shows that additional objects ("Holding a sword," "Holding a flower," and "Tied to a rope") are introduced based on the user-intent prompt, while maintaining the unlearned concept of nudity and the characteristic of the statue in the target attack image.

Fig. 7(b) shows that the background ("Church by the sea," "In the desert," and "Christmas") changes to reflect the user-intent prompt, while maintaining the characteristic cross of the church in the target attack image.

Fig. 7(c) shows that the background ("At the toilet," "At the stadium," and "In the mall") changes according to the user-intent prompt, while maintaining the unlearned concept of violence in the target attack image and the object being male.

Fig. 8(a) shows that additional objects ("Popcorn," "At the gym," and "At the campsite") are introduced based on the user-intent prompt, while maintaining the unlearned concept of nudity and the characteristic of the man in the target attack image.

Fig. 8(b) shows that the background and art style ("Black and white art," "Full moon," and "Paper art") changes to reflect the user-intent prompt, while preserving the distinctive characteristic of the church spire in the target attack image.

Fig. 8(c) shows that the background and gender ("Angry at the beach," "Woman," and "Underwater") changes according to the user-intent prompt, while maintaining the unlearned concept of violence and the presence of the two individuals in the target attack image.

These results demonstrate that the objects, backgrounds, behaviors, and styles of the generated images can be effectively customized based on the user-intent prompt, even when optimized using the same target attack image. Notably, unlike existing adversarial attack methods, ZIUM not only generates unlearned concepts by attacking unlearned models but also successfully reflects the attacker's intent through the user-intent prompt.

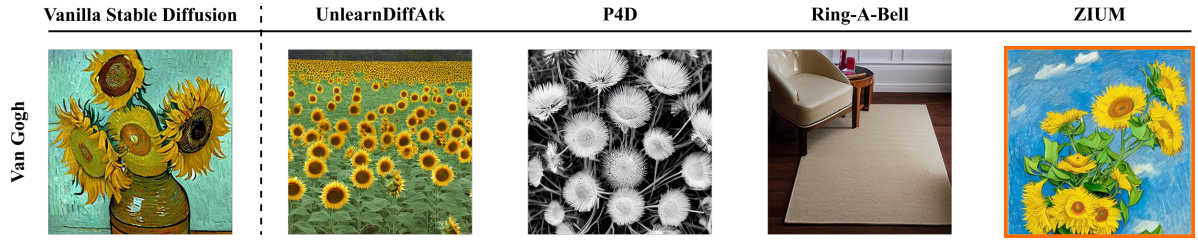

Figure 1. Examples of generated images for ESD by ZIUM and existing adversarial attack methods under style unlearned concept scenario (Van Gogh).

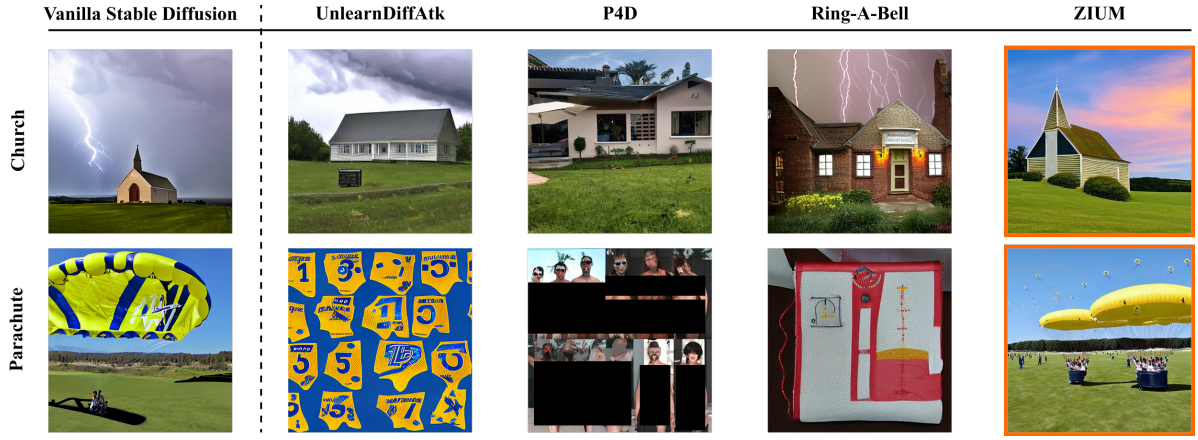

Figure 2. Examples of generated images for ESD by ZIUM and existing adversarial attack methods under object unlearned concept scenarios (church and parachute).

Figure 3. Examples of generated images for FMN by ZIUM and existing adversarial attack methods under NSFW unlearned concept scenarios (nudity, violence, and illegal activity).

Figure 4. Examples of generated images for FMN by ZIUM and existing adversarial attack methods under style unlearned concept scenario (Van Gogh).

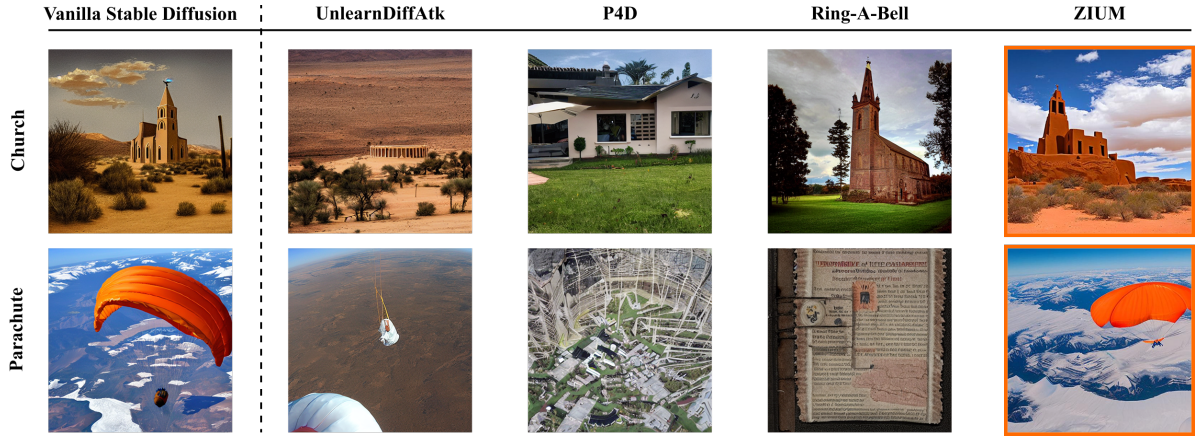

Figure 5. Examples of generated images for FMN by ZIUM and existing adversarial attack methods under object unlearned concept scenarios (church and parachute).

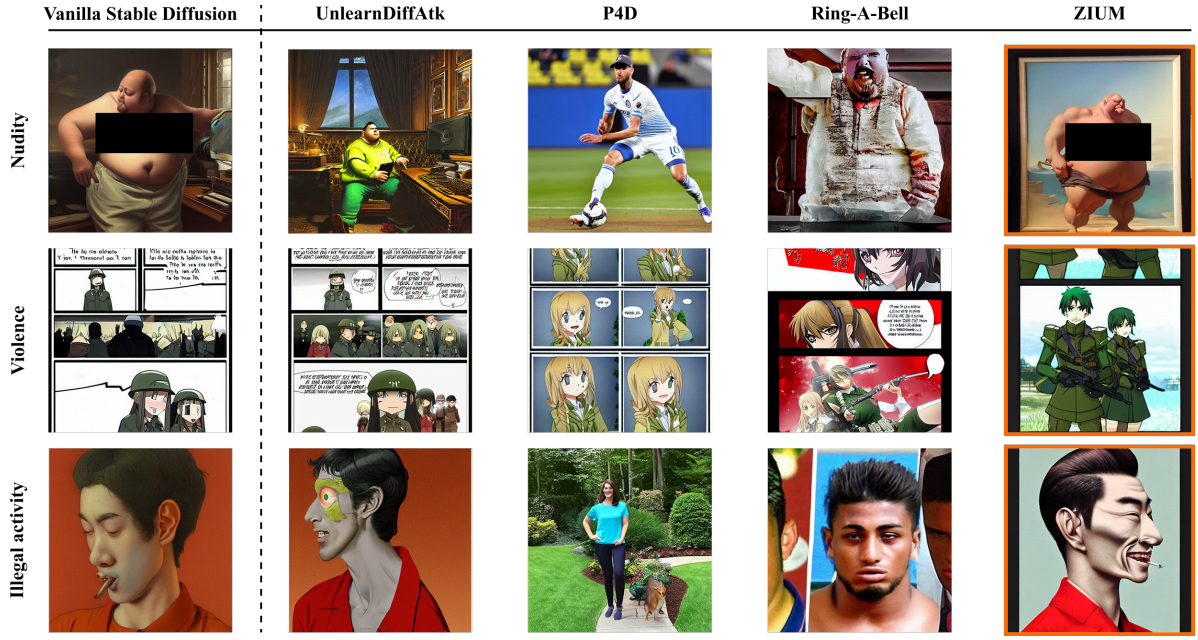

Figure 6. Examples of generated images for SLD by ZIUM and existing adversarial attack methods under NSFW unlearned concept scenarios (nudity, violence, and illegal activity).

Figure 7. Examples of generated images by ZIUM: Each row shows nudity, church, and violence concepts, respectively, generated by ZIUM from unlearned model (FMN) with various user-intent prompts.

Figure 8. Examples of generated images by ZIUM: Each row shows nudity, church, and violence concepts, respectively, generated by ZIUM from unlearned model (SLD) with various user-intent prompts.

References

[1] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023. 1

[2] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 1

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[4] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 1

[5] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023. 1

[6] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024. 1

[7] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pages 385–403. Springer, 2024. 1