A modular approach to enhancing safety in text-to-image diffusion models via targeted LoRA fine-tuning

Maksym Kizitskyi*† and Oleksii Turuta*†

Kharkiv National University of Radio Electronics, 14 Nauky Ave., Kharkiv, 61166, Ukraine

Abstract

This paper discusses recent advances in text-to-image generative models, which improve image quality and prompt alignment but also introduce vulnerabilities to harmful content from adversarial prompts. Existing safety measures often fail against complex attacks. We propose a novel safety approach using targeted Low-Rank Adaptation (LoRA) fine-tuning combined with Metric learning. Our approach aimed for modification of latent vectors of harmful prompts, aligning them with safe content to reduce unsafe outputs. Experiments demonstrate safety levels comparable to industry benchmarks, particularly when adapters are trained on specific harm categories. Our study provides a reusable framework to protect against harmful outputs that can be scaled to protect against multiple prompt categories.

Keywords

Harmful Content Mitigation, Low-Rank Adaptation, Metric learning, Generative Models

1. Introduction

Text-to-image (T2I) generative models serve as advanced tools that generate images from text descriptions. A prominent example is Stable Diffusion 3.5, which has made considerable progress in various aspects, such as the overall quality of generated images, precise font rendering, and the capacity to understand complex and detailed prompts.

Despite their remarkable abilities, T2I models present significant safety and ethical issues. They can produce harmful, offensive, or unsuitable content, which can pose risks to users. As a result, it is crucial to utilize these technologies responsibly and to establish safeguards and ethical guidelines to reduce potential harm [1].

While some safety protocols for these models were developed, many existing methods still demonstrate notable problems. They often compromise the quality of content generation, remain susceptible to sophisticated adversarial attacks, and fail to adequately address harmful or hateful material. These challenges are particularly pronounced in open-source models, where safety measures should not impede the model's ability to generate content effectively [2].

This study presents a method for enhancing safety in text-to-image models. We propose a fine-tuning technique that utilizes Low-Rank Adaptation (LoRA) combined with a metric learning approach. The primary objective of this research is to establish a framework aimed at minimizing the generation of harmful content by text-to-image (T2I) models. This is achieved by systematically refining the latent vector representations of unsafe inputs to steer them toward safer alternatives.

The distinctiveness of our research lies in the utilization of specialized LoRA adapters for fine-tuning, which can enhance safety performance to levels comparable to established industry standards, all while preserving image quality. Notably, our findings demonstrate that domain-

ISW-2025: Intelligent Systems Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine

* Corresponding author.

† These authors contributed equally.

maksym.kizitskyi@nure.ua (M. Kizitskyi); oleksii.turuta@nure.ua (O.P. Turuta).

0000-0001-9771-5771 (M. Kizitskyi); 0000-0002-0970-8617 (O.P. Turuta).

specific adapters, which focus on particular categories of harm such as self-harm content, outperform traditional general filtering methods.

This research contributes to ethical AI practices by providing a flexible safety framework that effectively addresses complex content safety challenges while retaining the generative capabilities of state-of-the-art models.

2. Related works

Recent advantages in AI [3-9], and especially Text-to-image (T2I) generative models, have significantly transformed digital content creation by generating highly realistic imagery from textual prompts. However, despite these advancements, these models are still susceptible to producing harmful or unsafe outputs when confronted with adversarial or inappropriate prompts. This review synthesizes recent research on the safety of T2I models, focusing on robustness evaluation, mitigation strategies, and emerging alignment methodologies. It underscores the importance of specialized datasets, such as Adversarial Nibbler and Inappropriate Image Prompts (I2P), for benchmarking vulnerabilities, and discusses the effectiveness of Low-Rank Adaptation (LoRA) in fine-tuning for safety. Additionally, it highlights the limitations of current NSFW detection frameworks in addressing generative outputs.

Contemporary T2I architectures, including Stable Diffusion 3.5 [10] with Multimodal Diffusion Transformers (MMDiT-X), utilize dual attention blocks and mixed-resolution training to enhance both image fidelity and textual alignment. Nonetheless, their open-ended generative capabilities present potential attack vectors, particularly through adversarial prompts that can bypass content filters [11]. Furthermore, diffusion processes may amplify subtle perturbations within latent spaces, leading to outputs that veer toward harmful or inappropriate content [12].

Safety risks primarily arise from three key factors: the generation of violent, explicit, or hateful imagery; the reinforcement of biased associations present in training data; and the creation of plausible yet misleading imagery. The I2P dataset systematically categorizes these risks into seven harm classes—harassment, hate, violence, self-harm, sexual content, shocking imagery, and illegal activities—based on crowdsourced red-teaming efforts. Its hierarchical structure includes over 12,000 annotated prompts, enabling multi-layered analyses of model failures, ranging from prompt misinterpretation to latent exploitation. While the dataset facilitates thorough evaluations, it shows a notable skew toward sexual content (38%), revealing an underrepresentation of harm categories such as microaggressions [13].

Studies typically assess robustness through three metrics: the percentage of adversarial prompts that produce harmful outputs; statistical distributions that indicate the likelihood of inappropriate content across multiple generations; and CLIP-based fidelity scores that evaluate the alignment between benign inputs and safety-filtered results. Current NSFW detection systems, predominantly based on ViT and trained on extensive photographic datasets, achieve high accuracy (~98%) on real images but see a decline to ~86% on synthetic outputs. Techniques like SafeText, which fine-tunes CLIP text encoders to adjust unsafe prompt embeddings, can reduce harmful generations by 72% without significantly degrading benign outputs. Conversely, methods focused on diffusion modules often introduce image artifacts[14].

LoRA (Low-Rank Adaptation) [15] incorporates trainable matrices into cross-attention layers for efficient parameter fine-tuning. When combined with Subcenter ArcFace loss, it can reduce NSFW output by 41% in Stable Diffusion 3.5 while preserving overall generative quality. However, adaptations driven by privacy concerns (e.g., SMP-LoRA) highlight a trade-off between privacy and fidelity; for instance, membership inference attacks decrease by 15%, yet FID scores increase by 0.22, indicating a decline in visual quality [16].

Iterative stress testing using frameworks such as SEAS—with the implementation of self-evolving prompts—can achieve GPT-4-level robustness following multiple optimization cycles. The ACE (Adversarial Concept Erasure) [17] technique effectively prevents unauthorized fine-tuning through targeted noise perturbations, outperforming untargeted methods by 29%. However, many defenses

still treat text and image modalities in isolation, leaving models vulnerable to multimodal jailbreaks. Hybrid architectures that combine ViT-based NSFW detection with language-based contextual analysis enhance safety, albeit at the expense of approximately 37% slower inference speeds. Moreover, culturally adaptive strategies remain insufficiently explored, as Western-centric annotations fail to effectively identify region-specific harms, resulting in a drop in F1 scores to 0.61[18].

There is an emerging consensus that treats text-to-image (T2I) content moderation as a form of hypothesis testing—distinguishing between safe (H_0) and unsafe (H_1) states through sensitivity analysis and loss landscape profiling. Techniques such as Gradient Cuff successfully detect 89% of adversarial jailbreaks but remain vulnerable to sophisticated obfuscation methods. Therefore, it is essential for the field to develop broader harm taxonomies, enhance cultural coverage, and adopt transparent, explainable frameworks to align generative AI with societal standards [19].

In summary, while T2I models have made significant advancements in generating high-quality images, they still face the risk of producing harmful outputs under malicious prompts. Innovations such as LoRA-based fine-tuning and new datasets (I2P, Adversarial Nibbler) present promising avenues for safety improvements. [20] However, they also underscore the necessity for standardization, culturally nuanced approaches, and more robust multimodal defenses. These insights inform our choice of Subcenter ArcFace loss in LoRA training and our implementation of ViT-based NSFW classification methods, highlighting the urgent need for community-driven benchmarks to evaluate safety in synthetic media [21, 22].

3. Methods and materials

Consider the data that will be used in further experiments and some other materials and methods proposed to solve the problem under consideration.

3.1. Datasets description

We utilize two complementary text-to-image (T2I) resources to examine model robustness, potential harms, and mitigation strategies in generative AI:

Adversarial Nibbler Dataset [23] is an innovative collection of T2I prompts and their corresponding images, designed to evaluate how effectively generative AI models manage implicitly adversarial attacks. Originating from the Adversarial Nibbler Challenge—a red-teaming methodology that crowdsources prompts intended to reveal how seemingly innocuous text inputs can generate harmful or unsafe content—the dataset comprises two primary components:
- Attempted Prompts: This includes all submitted prompts, generated images, and relevant metadata such as timestamps, model identifiers, and automated safety annotations for text and images.

Submitted Prompts: This is a validated subset of prompts recognized for containing safety violations. Alongside the original prompt and generated image, these entries feature participant rewrites detailing the nature of the harm, demographic targets, and potential failure modes, as well as responses from crowdsourced validation. Multiple annotators provide their assessments concerning safety, uncertainty, and risks. In total, there are 3748 unique prompts. Picture 1 shows the attack types distribution distribution of this dataset

Inappropriate Image Prompts (I2P) [24] is a dataset that offers real-world text-to-image (T2I) prompts which are disproportionately likely to generate inappropriate outputs when processed by diffusion-based generative models. Introduced in the 2023 CVPR paper "Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models," the I2P dataset is designed to assess and compare methods aimed at preventing harmful or problematic image generation. Grounded in the definition of inappropriate content—where data may be offensive, threatening, or anxiety-inducing—the benchmark focuses on seven primary categories of harmful imagery: hate, harassment, violence, self-harm, sexual content, shocking imagery, and illegal activity. The distribution of this classes shown on picture 2

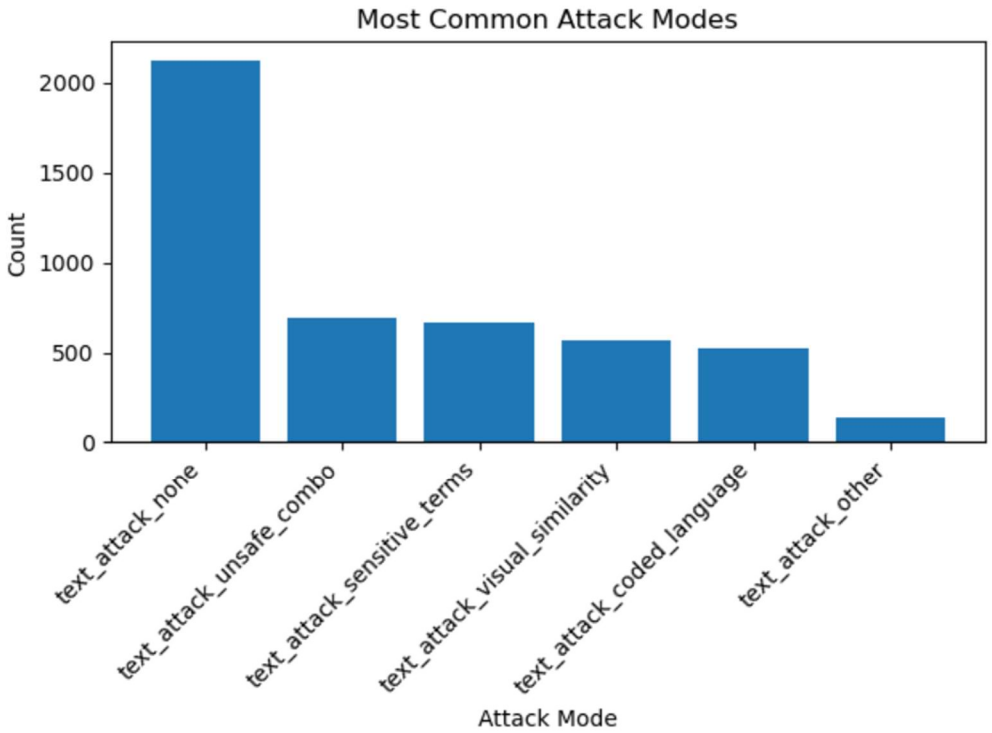

Figure 1: Most common attack types in Adversarial Nibbler Dataset

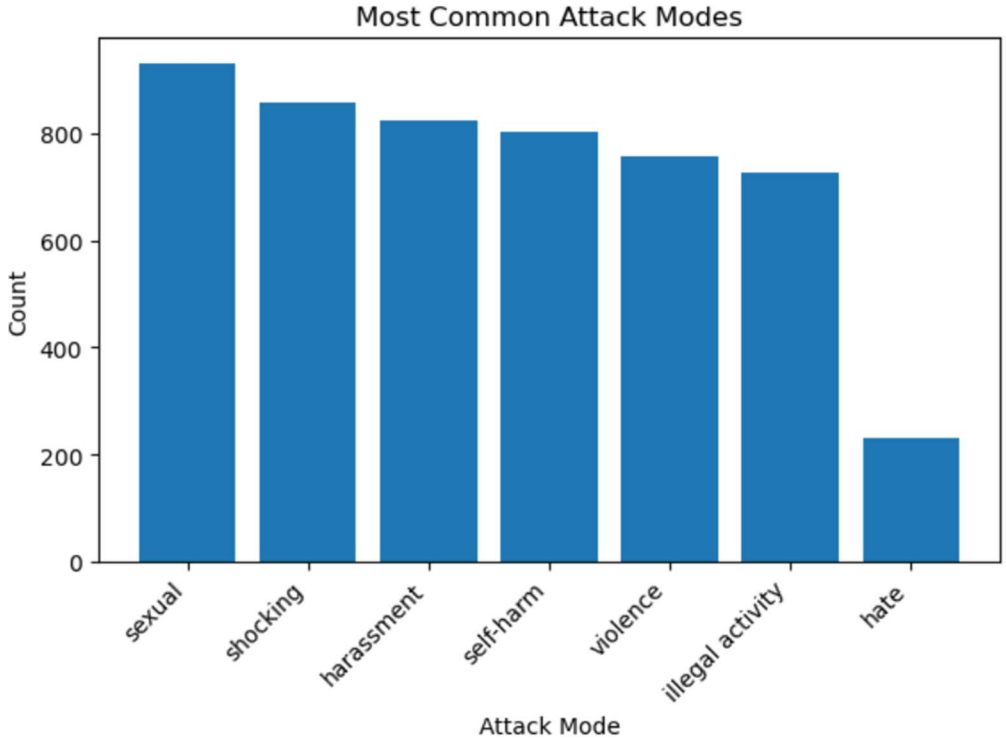

Figure 2: Most common attack types in Inappropriate Image Prompts Dataset

To compile the I2P dataset, the authors collected up to 250 prompts per keyword from Lexica.art, a platform that features user-generated prompts linked to Stable Diffusion parameters. These prompts are strategically placed close to the relevant inappropriate concepts in CLIP embedding space, maximizing the likelihood of producing problematic or unsafe content. The final dataset was refined by removing duplicates (based on unique prompt identifiers), resulting in a diverse collection of real-world prompts. Each I2P prompt is accompanied by metadata detailing the proportion of generated images classified as inappropriate, assessed by multiple classifiers such as Q16, NudeNet, and the

Stable Diffusion NSFW checker, along with an indication of whether at least half of the generated images are considered unsafe (“hard” prompts). The dataset also includes toxicity ratings, seeds, guidance scales, and links to the original prompts on Lexica. By providing standardized evaluation protocols and explicit estimates of inappropriate image generation, I2P facilitates rigorous experimentation on mitigating harmful outputs in diffusion-based T2I systems. In total, it has 4704 unique harmful prompts.

Collectively, these datasets create a robust testbed for investigating both implicitly adversarial and explicitly inappropriate content generation scenarios, thereby advancing the understanding of safety, fairness, and reliability in text-to-image models. To perform protective Lora training, LLAMA 3.2 was employed to generate text pairs that are non-harmful in response to harmful prompts.

3.2. Metrics

We have chosen as key metrics:

1. Attack Success Rate (ASR) - measures the proportion of adversarial prompts that successfully cause unintended, misleading, or harmful outputs. It is the primary metric for assessing the severity of the attack and the vulnerability of the model.
2. Minimum, maximum, avaerage NSWF score – to estimatate the range and distribution of inappropriate content generated by the model. This metric helps evaluate how consistently the model produces NSFW outputs in response to adversarial prompts, providing insights into the effectiveness of safety mechanisms and potential content moderation challenges.

4. Experiment

We employ Stable Diffusion 3.5 Medium, a prominent Multimodal Diffusion Transformer (MMDiT-X) developed by Stability AI, for high-quality text-to-image (T2I) generation tasks. This model features notable enhancements in image quality, typography, complex prompt handling, and resource efficiency through advancements such as dual attention blocks, QK-normalization, mixed-resolution training, and sophisticated pre-trained text encoders like OpenCLIP and T5-XXL.

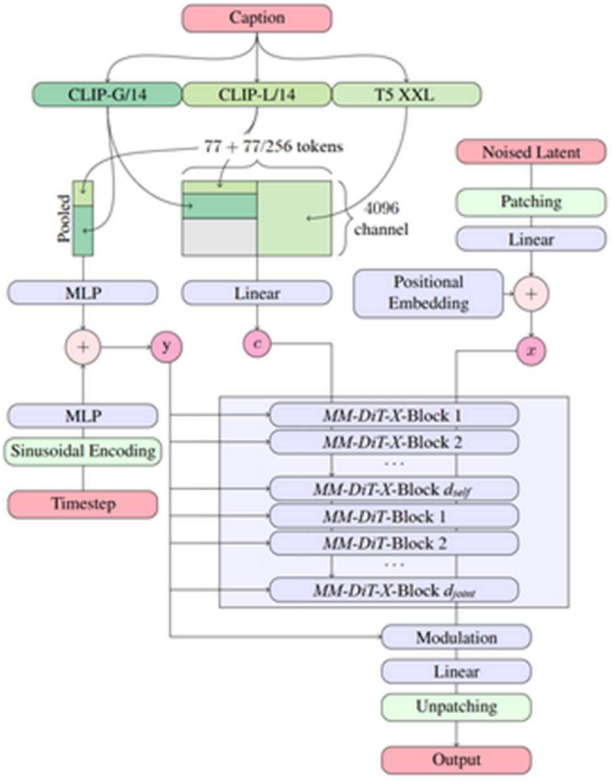

Figure 3: Stable Diffusion 3.5 Medium architecture [10]

Given its widespread use and adoption across various creative and research domains, it is essential to rigorously evaluate and ensure the safety of its generated outputs. Our assessments, utilizing datasets like Adversarial Nibbler and Inappropriate Image Prompts (I2P), provide a thorough analysis of potential risks and robustness, facilitating the responsible use and deployment of this powerful generative model.

To address safety concerns, the outputs of Stable Diffusion were automatically evaluated using a specialized NSFW detection model founded on a fine-tuned Vision Transformer (ViT) — specifically, the MMDiT-X variant. This classifier was pretrained on ImageNet-21k and fine-tuned on an annotated dataset containing approximately 80,000 images categorized as either normal or NSFW. With an impressive accuracy rate of 98%, it ensures effective automated identification of explicit or inappropriate content, thereby enhancing model safety throughout image generation workflows.

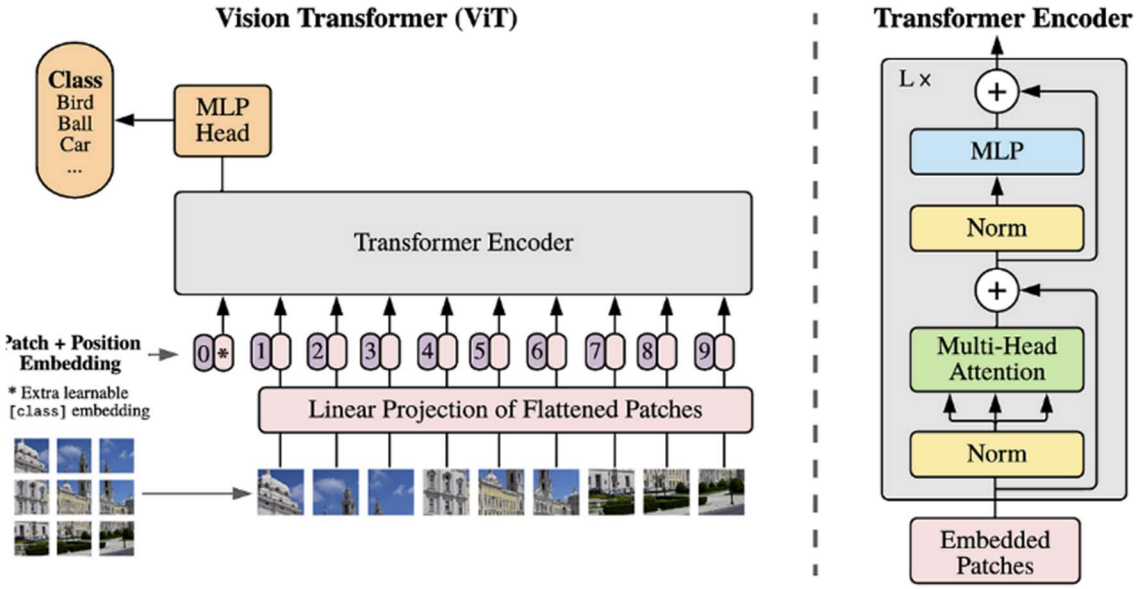

Figure 4: Vision Transformer architecture [25]

We conducted three distinct experiments utilizing two separate datasets, each divided into training, validation, and test subsets with proportions of 60%, 20%, and 20%, respectively. The primary goal of the first experiment was to establish baseline performance metrics for the Stable Diffusion model without incorporating any protective mechanisms against harmful content generation. For each input prompt, the model generated five unique images, each evaluated individually using a specialized NSFW (Not Safe For Work) classification model. Additionally some pictures were manually reviewed.

In the second experiment, we trained a Low-Rank Adaptation (LoRA) [26, 27, 28] adapter using the Subcenter ArcFace loss function. The main objective was to adjust the latent vector representations of harmful prompts to align as closely as possible with representations associated with non-harmful content. The training parameters included the AdamW optimizer, configured with a learning rate of 5e-4 specifically for the parameters of the loss function. Additionally, a OneCycleLR scheduler was implemented to dynamically adjust the learning rate throughout the training process, reaching a maximum of 5e-3 over a total of ten epochs.

In the third experiment, the LoRA adapter was trained to block one class of images at a time, specifically targeting “self-harm.” The adapter was trained with the same loss function, optimizers,

and hyperparameters, but only the prompts of the specified class were altered while others remained unchanged. Training was conducted in a Google Colab environment equipped with an NVIDIA A100 GPU. The training process exhibited a consistent reduction in the loss.

5. Results

Figure 5 depicts the test loss curves for the LoRA adapters utilized in Experiments 2 and 3, illustrating a consistent decline in loss over time. This trend suggests that the adapters effectively learn to reduce harmful content generation.

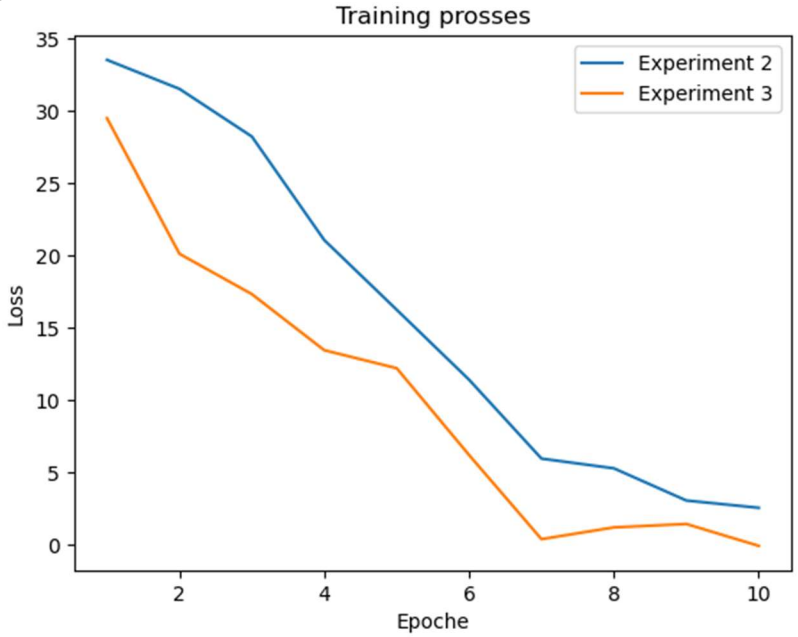

Figure 5: Training processes of both LoRA adapters

Figure 6 provides a comparative assessment of the baseline model and our LoRA-based approach, revealing a significant reduction in unsafe outputs while maintaining overall image quality. Figure 7 – shows metrics comparison against one category “self-harm”

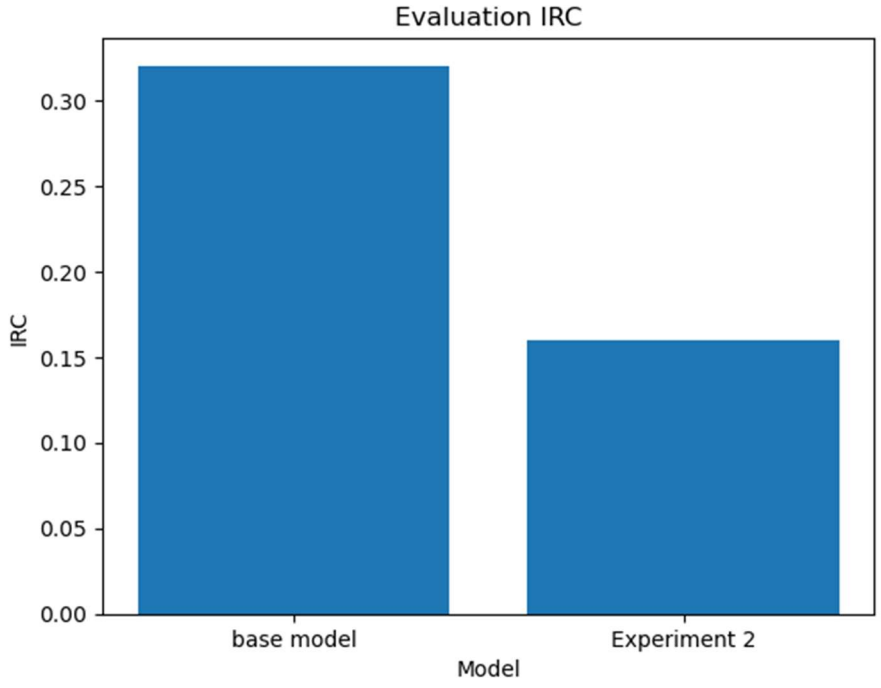

Figure 6: Evaluation result on all categories

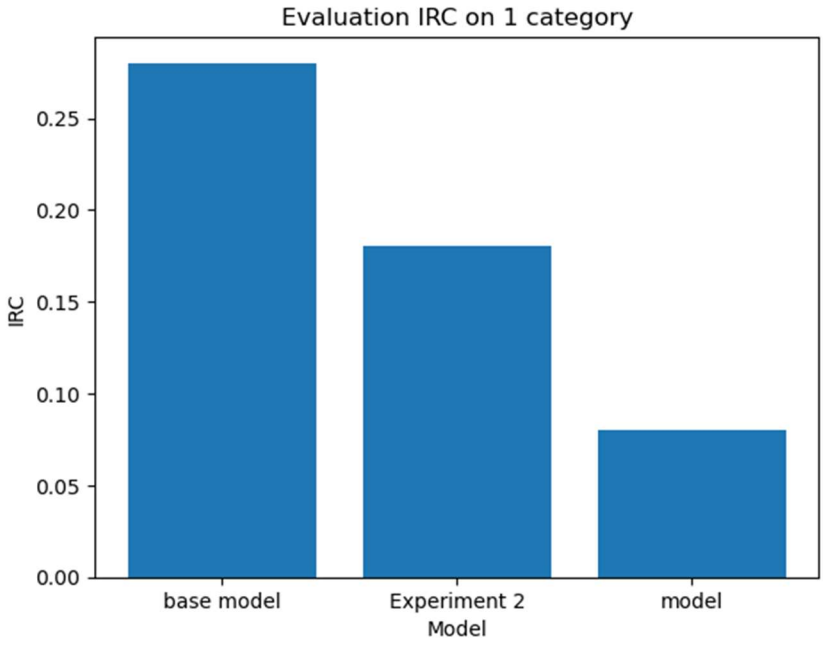

Figure 7: Evaluation result on one category

All metrics aimed to compare both approaches and the base model are presented in Table 1.

Table 1
Frequency of Special Characters

Approach	ASR (%)	NSFW Score (Avg)
SD 3.5	31	0.45
Experiment 2 - SD 3.5 + General LoRA	16	0.21
Experiment 3 - SD 3.5 + Specific LoRA	8	0.12

In Figure 8, we present representative samples generated by the baseline model, highlighting its vulnerability to producing inappropriate content when confronted with adversarial prompts.

a) b) c)

Figure 8: Generation examples by each version of the model a) Original model b) experiment 2 c) experiment 3

Across all figures, the results indicate that the integration of LoRA fine-tuning with safety-oriented loss functions leads to considerable advancements in minimizing harmful or NSFW content while preserving the model's generative capabilities.

6. Discussions

Upon completing the training process, we assessed the model's effectiveness using a test subset and compared its performance against a baseline that did not incorporate our newly introduced safety mechanisms. This outcome highlights the potential of our method to deliver robust safety features that rival existing industry standards. Both approaches showed significant ASR reduction – 48% for general LoRA and 74% for specialized LoRA on one category.

A significant finding emerged from our investigations into class-specific LoRA training, such as for self-harm imagery. Models equipped with these specialized adapters exhibited superior visual fidelity compared to those utilizing a broad, generalized NSFW filter. This suggests that a modular, domain-targeted approach can more effectively address nuanced safety requirements. In practice, multiple specialized LoRA adapters could be deployed simultaneously to block various categories of harm—including self-harm and explicit content—while minimizing negative impacts on benign image generation.

However, in each experiment, after seven epochs of training with the adapters, the learning process slowed down. This may indicate that the model is unable to extract meaningful insights from the data beyond that point.

Looking ahead, further validation of this modular approach will require extensive experimentation with various model architectures, training datasets, and cultural contexts. Specifically, rigorous testing against white-box adversarial prompts is crucial to evaluate how effectively these adapters can withstand carefully engineered attacks designed to circumvent standard safety mechanisms. By systematically expanding our defenses and refining their implementation, we aim to advance the development of safe, scalable image generation models that can be responsibly utilized across diverse real-world applications.

7. Conclusions

This work presents a robust and modular framework aimed at enhancing safety within text-to-image (T2I) generative models, with a particular emphasis on Low-Rank Adaptation (LoRA) fine-tuning utilizing Subcenter ArcFace loss. Through comprehensive experiments conducted on the Adversarial Nibbler and Inappropriate Image Prompts (I2P) datasets, our proposed methodology demonstrates significant effectiveness in reducing harmful outputs—evidenced by decreased Attack Success Rate (ASR) and minimized NSFW content—while preserving high image fidelity.

Importantly, the class-specific LoRA adapters trained to target individual harm categories, such as self-harm, show superior visual quality compared to broader safety filters, underscoring the advantages of targeted interventions. These findings indicate that future safety solutions could benefit from adopting a multi-adapter strategy that seamlessly incorporates domain-specific defenses without compromising overall creative flexibility. Moreover, our results reveal competitive alignment with established safety mechanisms within the industry, such as Safe Stable Diffusion, positioning LoRA-based methods as a resource-efficient and adaptable alternative.

Despite these encouraging outcomes, certain challenges persist. First, real-world implementations must take into account the evolving and culturally nuanced definitions of harm, necessitating continuous updates to safety annotations and a more comprehensive consideration of cultural contexts. Second, future white-box adversarial evaluations are vital to verify the robustness of LoRA adapters against intentionally designed bypasses. Lastly, although our approach maintains generative quality, ensuring long-term resilience against new and increasingly sophisticated attacks requires ongoing research.

In conclusion, this study establishes a strong foundation for modular, targeted, and parameter-efficient safety interventions in advanced T2I models. By integrating LoRA fine-tuning with purpose-driven loss functions, practitioners can effectively address a variety of harm categories, uphold generative excellence, and promote responsible AI deployment. We encourage the community to build upon this work by expanding our framework to include broader model architectures and real-

world applications, ultimately paving the way for safer, more inclusive, and high-fidelity generative AI systems.

Declaration on Generative AI

During the preparation of this work, the authors used Stable Diffusion 3.5 medium for Figure 8 in order to generate all 3 images in order to showcase how the approach proposed in this paper affects image generation. After using these tools, the author reviewed the content as needed and took full responsibility for the publication's content.

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete Problems in AI Safety. arXiv preprint arXiv:1606.06565.

[2] Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2017). AI Safety Gridworlds. arXiv preprint arXiv:1711.09883.

[3] Saichyshyna, N., Maksymenko, D., Turuta, O., Yerokhin, A., Babii, A., & Turuta, O. (2023). Extension Multi30K: Multimodal Dataset for Integrated Vision and Language Research in Ukrainian. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP) (pp. 54–61). Dubrovnik, Croatia: Association for Computational Linguistics.

[4] Maksymenko, D., & Turuta, O. (2024). Interpretable Conversation Routing via the Latent Embeddings Approach. Computation, 12(12), 237. https://doi.org/10.3390/computation12120237

[5] Erdem, E., Kuyu, M., Yagcioglu, S., Frank, A., Parcalabescu, L., Plank, B., Babii, A., Turuta, O., Erdem, A., Calixto, I., Lloret, E., Apostol, E.-S., Truică, C.-O., Šandrih, B., Martinčić-Ipšić, S., Berend, G., Gatt, A., & Korvel, G. (2022). Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability, and Learning. Journal of Artificial Intelligence Research, 73. https://doi.org/10.1613/jair.1.12918

[6] Panchenko, D., Maksymenko, D., Turuta, O., Luzan, M., & Tytarenko, S. (2022). Ukrainian News Corpus as Text Classification Benchmark. In Ignatenko, O., et al. (Eds.), ICTERI 2021 Workshops. ICTERI 2021 (Vol. 1635). Springer, Cham. https://doi.org/10.1007/978-3-031-14841-5_37

[7] Improving Speaker Verification Model for Low-Resources Languages. CEUR Workshop Proceedings, 3403, 99–113. 7th International Conference on Computational Linguistics and Intelligent Systems (CoLInS 2023), Kharkiv, 20–21 April 2023.

[8] O. Zolotukhin, V. Filatov, A. Yerokhin, O. Lanovyy, M. Kudryavtseva, V. Semenets, An approach to the selection of behavior patterns autonomous intelligent mobile systems, in: Proc. IEEE Int. Conf. Problems Infocommun. Sci. Technol. (PIC S&T), 2021, pp. 349–352. doi:10.1109/PICST54195.2021.9772110.

[9] O. Zolotukhin, V. Filatov, A. Yerokhin, M. Kudryavtseva, The methods for the prediction of climate control indicators in the Internet of Things systems, CEUR Workshop Proc., 2021. doi:10.5281/zenodo.14526027.

[10] Schramowski, P., Numao, M., Deiseroth, B., Adilova, L., Bitterwolf, J., Kutyniok, G., & Kersting, K. (2023). Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In Proceedings of CVPR 2023. arXiv preprint arXiv:2211.05105.

[11] Perez, F., & Ribeiro, I. (2022). Ignore Previous Prompt: Attack Techniques for Language Models. In Proceedings of the ML Safety Workshop, NeurIPS 2022. arXiv preprint arXiv:2211.09527.

[12] Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., et al. (2024). PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. arXiv preprint arXiv:2306.04528.

[13] Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., et al. (2024). Prompt Injection Attack Against LLM-integrated Applications. arXiv preprint arXiv:2306.05499.

[14] Kumar, A., Agarwal, C., Srinivas, S., Li, A. J., Feizi, S., & Lakkaraju, H. (2024). Certifying LLM Safety against Adversarial Prompting. arXiv preprint arXiv:2309.02705.
[15] Liu, Y., Jia, Y., Geng, R., Jia, J., & Gong, N. Z. (2024). Formalizing and Benchmarking Prompt Injection Attacks and Defenses. Proceedings of the 33rd USENIX Security Symposium, Philadelphia, PA, 2024.
[16] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.
[17] Wang, Y., Liu, X., Li, Y., Chen, M., & Xiao, C. (2024). AdaShield: Safeguarding Multimodal Large Language Models from Structure-based Attack via Adaptive Shield Prompting. arXiv preprint arXiv:2403.09513.
[18] Xiong, C., Qi, X., Chen, P. Y., & Ho, T. Y. (2024). Defensive Prompt Patch: A Robust and Interpretable Defense of LLMs against Jailbreak Attacks. arXiv preprint arXiv:2405.20099.
[19] Howe, N., McKenzie, I., Hollinsworth, O., Zajac, M., Tseng, T., Tucker, A., Bacon, P.L., & Gleave, A. (2024). Effects of Scale on Language Model Robustness. arXiv preprint arXiv:2407.18213.
[20] Akhtar, N., Mian, A., Kardan, N., & Shah, M. (2021). Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. IEEE Access 9: 155161-155189. DOI: 10.1109/ACCESS.2021.3127960
[21] Akhtar, N., & Mian, A. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access 6: 14410-14430. DOI: 10.1109/ACCESS.2018.2807385
[22] Stable Diffusion 3.5 Medium: https://huggingface.co/stabilityai/stable-diffusion-3.5-medium
[23] Adversarial Nibbler Dataset: https://github.com/google-research-datasets/adversarial-nibbler
[24] Inappropriate Image Prompts (I2P): https://huggingface.co/datasets/AIML-TUDA/i2p
[25] Falconsai. (n.d.). Fine-Tuned Vision Transformer (ViT) for NSFW Image Classification. Hugging Face Model Hub. Retrieved March 13, 2025, from https://huggingface.co/Falconsai/nsfw_image_detection
[26] IBM LoRA Guide: https://www.ibm.com/think/topics/lora
[27] Nexla Enterprise AI LoRA Guide: https://nexla.com/enterprise-ai/low-rank-adaptation-of-large-language-models/
[28] SafetyDPO Website: https://safetydpo.github.io