Term Paper: IMMA - Immunizing Text-to-Image Models Against Malicious Adaptation

Anonymous submission

Abstract

The increasing access to large-scale text-to-image models accelerated personalization in content creation-what is commonly referred to as a double-edged sword. It opens avenues for misuse, ranging from unauthorized style mimicry to creating harmful content. While data-poisoning methods have tried to protect the images against such misuse, active protection places an unfair burden on content creators and often does not sufficiently work. Another alternative could come from the IMMA framework, or Immunizing Models against Malicious Adaptation, which in itself immunizes the model against harmful adaptation.

This paper is a detailed implementation of the IMMA with the aim to test its robustness and practicality. We perform an experimental analysis to gauge the effectiveness of IMMA in rendering models resilient against unauthorized adaptation. Our re-implementation validates several key findings of the original IMMA framework, which indeed proves to be remarkably resistant to malicious adaptations. However, we also point out some limitations regarding the design of IMMA, giving some possible future directions for enhancement in increasing the adaptiveness and applicability of IMMA. This work contributes to responsible open-sourced generative models by rigorously testing and validating model-level defenses.

Code — https://github.com/heesookiim/ECE570/upload/main

Introduction

Recent progress in generative AI has significantly increased the rate at which digital content can be created and customized. Large-scale text-to-image models, such as Stable Diffusion[4] and DALL-E, provide unprecedented capabilities to create high-quality imagery from textual prompts in a highly customized way. These open-source models reduce entry barriers into content creation, allowing users to fine-tune models for unique purposes including personalized artistic styles or particularized visual themes. While the number of applications for these models continues to increase, so does the number of ways in which the technology can be misused. The same adaptation techniques enabling personalized content, such as Textual Inversion[1], DreamBooth[5], and LoRA[2], can be utilized to produce unauthorized or harmful content, including protected artistic style copying and explicit/sensitive imagery without consent.

Some methods have attempted to secure the data itself as a way of mitigating these risks. The other methods, such as data poisoning, modify the images by making imperceptible changes in order to make the fine-tuning algorithms fail to replicate the content correctly. Examples of this include Glaze[6] and MIST[3], which try to safeguard original contents from unauthorized stylistic impersonation through disallowing models from learning from respective images.

These latter approaches put the onus of protection on content creators, having them modify their works preemptively to prevent misuse. In addition, the schemes with data poisoning are incomplete protections in many cases; they do not address vulnerabilities at the model level and can be easily evaded by determined attackers. In view of these limitations, Zheng and Yeh proposed the IMMA framework[8], a model-level defense that shifts the focus from protecting data to protecting the generative model itself. In this case, IMMA tries to immunize the model against harmful adapting by modifying its internal parameters in such a manner as would affect malicious fine-tuning.

IMMA offers a proactive solution independent of data modification and learns useful parameter settings that degrade the success of malicious adaptations while keeping the model useful for benign applications. Following this, IMMA leverages a bi-level optimization process that trains some of the model parameters to perform poorly on malicious tasks, such as unauthorized style replication or sensitive content generation, when adapted using methods like LoRA or DreamBooth. The framework of IMMA is comprehensively reimplemented in this paper to further validate its effectiveness, study its robustness under various adaptation scenarios, and gather any limitations that arise in practice. We evaluate the capability of IMMA to prevent unauthorized adaptations across a wide set of tasks and adaptation methods using experiments duplicating conditions from the original study. Our results agree with the findings in the original paper: IMMA strongly reduces the malicious adaptation fidelity. We point out where modifications are needed, for instance, improving scalability of IMMA to multiple target concepts, and study its effect on benign adaptations. In this work, we suggest that IMMA can support the development of safer open-source generative models facing ethical challenges due to the rapid advancement of AI-based content creation.

Related Work

Generative AI and Diffusion Models

Recent breakthroughs in generative AI, especially diffusion models, have achieved quality leaps in image synthesis. In a nutshell, diffusion models gradually develop noise into meaningful images with impressive visuals across diverse domains. By leveraging the power of internet-scale datasets, models such as Stable Diffusion or DALL-E have widely gained user appeal due to their functionality, which has been making the dream of complex visuals from text prompts a reality. The open-sourced nature of these models can also make them inherently vulnerable to abuse, especially via fine-tuning methods that allow adaptation for target-specific visual outputs.

Adaptation Methods in Generative Models

Adaptation techniques such as Textual Inversion, DreamBooth, and LoRA enable users to personalize generative models with new visual concepts. Textual Inversion gives models the ability to learn certain visual features from a few sample images and associate those with novel textual tokens, while DreamBooth uses fine-tuning to introduce unique visual styles or objects into generative models, effectively teaching the model a specific style or identity. LoRA improves model adaptation efficiency by constraining modifications to low-rank matrices, which makes adaptations faster and more effective. While these techniques are beneficial, they also create potential for misuse, such as unauthorized replication of protected artistic styles or the generation of explicit content.

Countermeasures: Data Poisoning and Content Protection

In response to these risks, data-centric approaches like data poisoning have been proposed. Data poisoning involves making subtle, often imperceptible, changes to images to confuse adaptation algorithms, making it harder for generative models to learn specific styles or content. Notable data poisoning methods include Glaze and MIST, designed to introduce noise or adversarial features into training data, thereby reducing the effectiveness of adaptation methods. Although useful, data poisoning requires content creators to proactively modify their work, placing a protective burden on them and often offering only partial solutions to model vulnerabilities. Additionally, these methods may be circumvented through preprocessing steps, limiting their real-world robustness.

Model-Level Protections: IMMA

IMMA, or Immunizing Models against Malicious Adaptation, offers a model-based alternative to data poisoning, targeting the model's adaptability to harmful content. IMMA avoids altering training data, instead adjusting the model's internal parameters to weaken its ability to adapt to malicious purposes, making it harder for the model to learn unauthorized styles or content. IMMA employs a bi-level optimization approach, tactically updating model parameters to reduce adaptation success rates while preserving flexibility for benign applications. This proactive model-level defense shifts the paradigm from protecting data alone to enhancing model robustness, potentially offering a more comprehensive solution against misuse.

Meta-Learning and Optimization in Model Adaptation

IMMA's optimization approach is inspired by meta-learning, where models are trained not only for specific tasks but to perform well across a range of adaptations. Meta-learning has been extensively studied for tasks requiring quick adaptation to new data, often using methods like Model-Agnostic Meta-Learning (MAML) and hyperparameter tuning for efficient generalization. IMMA takes a different approach by training model parameters to resist adaptation rather than enhance it, effectively flipping the typical meta-learning paradigm to create a "poor" initialization that hinders malicious fine-tuning. This represents a novel application of meta-learning focused on security and robustness in generative models.

In brief, IMMA's model-based approach capitalizes on advances in generative AI adaptation while addressing the limitations of data-poisoning techniques. Our reimplementation aims to validate and extend IMMA's role as a safeguard for open-source generative models in today's increasingly complex digital landscape.

Problem Definition

The goal of IMMA is to shift the focus from data-centric defenses to model-centric immunization. To do so, IMMA formulates model immunization as a bi-level optimization problem to weaken a model's adaptability to harmful tasks. Given a pre-trained model with parameters θ_p and a harmful target concept c' (e.g., unauthorized artistic style or explicit content), the goal of IMMA is to produce immunized parameters θ_I that limit the model's effectiveness in adapting to c'. This is achieved through the following optimization tasks: Lower-Level Optimization (Simulating Malicious Adaptation) and Upper-Level Optimization (Immunization).

First, an adaptation method A (e.g., Textual Inversion, DreamBooth, or LoRA) is applied to model parameters θ_I to learn the target concept c'. The adaptation minimizes the adaptation loss $L_A(x', c'; \theta_I, \varphi)$ over parameters φ, which may include modified tokens and weights specific to the adaptation task. This simulates an attacker's attempt to adapt the model for unauthorized purposes.

In response to the lower-level task, IMMA maximizes the adaptation loss with respect to θ, making it harder for the adaptation process to learn the target concept. Specifically, IMMA adjusts parameters θ to minimize the effectiveness of fine-tuning for c', ideally making the adaptation method unable to capture the harmful concept. This is formalized as:

$$\max_{\theta \in S} L_A(x'_I, c'; \theta, \varphi^*) \tag{1}$$

where φ^* is the solution to the lower-level problem, and S is the subset of model parameters chosen for immunization. The objective is to learn a model initialization θ that, when

adapted by malicious methods, performs poorly at generating harmful outputs.

Algorithm: IMMA Training Process

The iterative training process for IMMA is summarized in **Algorithm 1**.

Algorithm 1: IMMA: Immunizing Model against Malicious Adaptation

Require: Pre-trained model weights θ^p, dataset $D = \{x'\}$ representing the target concept c', learning rate α, set of parameters S to be adjusted by IMMA, adaptation loss function L_A

Ensure: Immunized model weights θ^I

1: Initialize model parameters $\theta^0 = \theta^p$
2: Initialize adaptation parameters ϕ^0 according to the adaptation method A
3: **for** each iteration $i = 1$ to I **do**
4: Sample a batch of data x'_A for adaptation and x'_I for immunization from D
5: Set $\phi \leftarrow \phi^{i-1}$ (carry over adaptation parameters from the previous iteration)
6: Update $\phi^i \leftarrow \arg\min_\phi L_A(x'_A, c'; \theta^{i-1}, \phi)$ (minimize adaptation loss to simulate malicious fine-tuning)
7: Adjust immunized parameters $\theta_S^i \leftarrow \theta_S^{i-1} + \alpha \nabla_\theta L_A(x'_I, c'; \theta^{i-1}, \phi^i)$ (maximize adaptation loss with respect to model parameters)
8: **end for**
9: **return** Final immunized parameters θ^I

In this algorithm, we alternate between lower-level adaptation and upper-level immunization updates. At each iteration, the lower-level task minimizes L_A with respect to ϕ to simulate adaptation, while the upper-level task maximizes L_A with respect to θ, aiming to create a poor initialization for malicious adaptations.

To summarize, the success of IMMA's bi-level optimization approach allows it to directly counteract fine-tuning efforts by attackers, preventing malicious adaptation. This model-centric defense shifts focus from data protection to robust model initialization, offering a scalable solution for managing the risks associated with open-source generative models.

Methodology

IMMA Implementation and Codebase

We used the official IMMA repository[7] for this reimplementation. The repository contains essential resources and code for setting up and running the IMMA framework, including scripts for model training, adaptation, and evaluation needed to reproduce the experiments in the original paper. We employed the bi-level optimization framework provided by the repository to accurately simulate adaptation tasks and apply IMMA's immunization technique.

Reference Image Selection

To achieve comparable results to those in the original IMMA paper, we selected similar prompts for model evaluation. Specifically, for experiments on the relearning of personalized content, we used the prompt "a purse on a beach." This prompt was chosen for its resemblance to examples in the original IMMA evaluation, allowing a more direct comparison between the immunized and non-immunized model results. This consistent prompt ensured we could accurately measure IMMA's effectiveness in preventing unauthorized concept adaptation.

Computational Resources

Running IMMA's bi-level optimization and experiments involving high-resolution text-to-image generation required significant computational resources. We conducted all experiments on Purdue University's Gilbreth A10 GPUs, which provided the necessary computation power to run the IMMA framework and adaptation methods efficiently. These GPUs were instrumental in managing the high computation needs for both adaptation simulations and immunization training cycles.

Generative Model: Stable Diffusion

For all experiments, we used the Stable Diffusion model as the base generative model, accessed via the Hugging Face Model Hub. Stable Diffusion's architecture closely resembles the setup in the original IMMA paper and is open-source. Initial images were generated using Stable Diffusion as baselines, which were then adapted to test IMMA's resistance to unauthorized content generation.

Experimental Setup

We considered two main scenarios as the following: Relearning of artistic styles and DreamBooth adaptation for personalized content.

Training Images and Model Preparation To measure the efficiency of IMMA in preventing the relearning of some artistic styles, we created 100 images representatives of Van Gogh's style using the prompt **"An artwork by Van Gogh"**. Such images were utilized as training data for the relearning experiments. Moreover, we have considered pre-trained weights of the erased model for the Van Gogh style downloaded from https://erasing.baulab.info. These deleted weights therefore became a baseline on which to measure how IMMA altered the model's ability to relearn a previously deleted style.

Training IMMA Weights To vaccinate the model against that, we first trained IMMA weights on the training images of Van Gogh. For all experiments below, we utilized the following key hyperparameters:

- Resolution: 512
- Batch Size: 1
- IMMA Learning Rate: 1×10^{-5}
- Epochs: 50
- LR Scheduler: Constant with no warmup

Adapting DreamBooth on IMMA-Immunized Model In addition to relearning, we employed Lora adaptation to simulate the model's ability to generate personalized content. Using the prompt "a purse on a beach," we tested both immunized and non-immunized versions of the model to assess whether IMMA could prevent the model from adapting to new, highly specific content. DreamBooth, which fine-tunes multiple model parameters to capture fine details, provided a challenging test for IMMA's effectiveness in limiting unauthorized adaptations while preserving the model's versatility for benign uses.

- Learning Rate: 1×10^{-4}
- Epoch: 50
- Mixed Precision: FP16

Computational Environment All experiments were performed with the help of a A10 GPU cluster at Purdue Gilbreth server, which provided all the computational resources required to train and evaluate the IMMA framework. The stable diffusion model from Hugging Face was used, and all experiments were visualized and tracked with sinteractive and sbatch function from Gilbreth in looking at training and validation performance.

By following these procedures, we were able to reimplement and rigorously evaluate the IMMA framework within a controlled environment. This setup ensured a faithful replication of the conditions in the original study, allowing us to compare results accurately and derive insights into IMMA's strengths and limitations as a safeguard against harmful adaptations.

Experimental Results

Experiment Overview

Accordingly, we designed two major experiments: which estimates the resistance of IMMA against content-specific adaptation when using LoRA for relearning an erased artistic style. The second experiment evaluates the effectiveness of IMMA against DreamBooth adaptation with the goal of generating personalized content.

Both experiments quantified the ability of IMMA to degrade the model's performance on unauthorized content generation while preserving the adaptability for benign uses. Conclusions validating the consistency and robustness that IMMA provides in terms of immunization are drawn by comparing our results with results from the original study of IMMA.

Experiment 1: Relearning of Artistic Styles

Setup In this experiment, we tested how well IMMA could prevent the relearning of Van Gogh's style. We took a Stable Diffusion model and used an erased model checkpoint-a model in which Van Gogh's style had been selectively removed-to serve as our baseline. Using this erased model, we applied LoRA to fine-tune the model on a dataset of Van Gogh style images under two conditions: with IMMA and without IMMA.

In all the experiments, the prompt was "An artwork by Van Gogh"; hence, we could see how closely the model could mimic the style of Van Gogh both with and without IMMA.

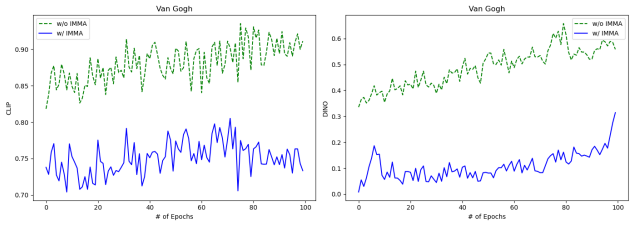

Figure 1: Similarity vs. epochs for LoRA on Van Gogh styles

Metrics The main metrics used to measure the relearning process include CLIP similarity and DINO similarity. These metrics estimate semantic similarities between generated images and reference Van Gogh images while quantifying how well the model can actually mimic the artistic style. Higher scores mean greater similarity, showing that the model has relearned the erased style.

Results According to Figure 1, the model without IMMA obtained a monotonically increasing CLIP and DINO similarity score over 100 training epochs, which would imply successful adaptation back to the Van Gogh style. Then the model with IMMA had consistently much lower similarity scores, suggesting that it strongly resisted learning the style again. Finally,

- **CLIP Similarity**: The model without IMMA reached a CLIP score of almost 0.90 by the 100th epoch, whereas the model with IMMA remained at around 0.75, hardly adapting to the style of Van Gogh.
- **DINO Similarity:** Analogously, DINO scores were higher for the model with no IMMA, reaching about 0.6 toward the end of training, and much lower for the model with IMMA, struggled to adapt at around 0.1–0.3.

These results are in line with the findings from the original IMMA paper that IMMA is effective at preventing unauthorized style adaptation, even for a very characteristic style like that of Van Gogh.

Qualitative Analysis Qualitative results are shown in Figure 2, which compares the images generated by the model with and without IMMA against the reference Van Gogh images. Employing a model without IMMA was able to successfully relearn and generate Van Gogh's style images similar to the original "Starry Night" aesthetic. As opposed, the IMMA model outputs reveal an art far from similar to the authentic nature of Van Gogh's style, hence the employment of IMMA is effective in hindering unauthorized adaptation.

Experiment 2: DreamBooth LoRA for Product Content Personalisation

Setup The second experiment demonstrates the capability of IMMA to curb the unauthorized generation of personalized content using DreamBooth adaptation. The personalized content used in this experiment includes a concept of

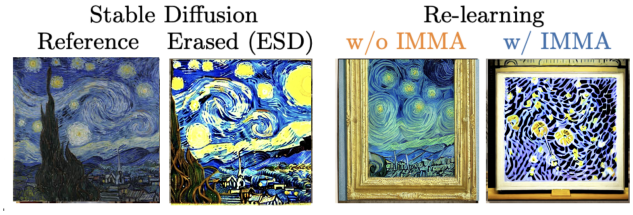

Figure 2: Results against relearning using IMMA

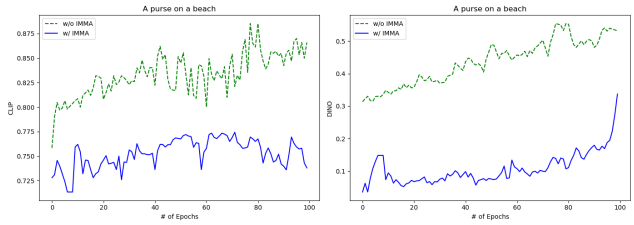

Figure 3: Similarity vs. epochs for LoRA on a purse

luggage purse, where training of the model is done to generate variations of a purse on the beach. It was trained with and without IMMA protection on the prompt "A purse on a beach" to produce validation images.

Metrics As with the first experiment, we will calculate CLIP similarity and DINO similarity to measure adaptation success. These metrics quantify semantic alignment between generated images and reference images of a purse on a beach, measuring how closely the model adapts to the personalized content.

Results Indeed, Figure 3 shows the results for a clear difference in effectiveness of adaptation:

- **CLIP Similarity:** The model indeed had a very high score for CLIP similarity at about 0.875 without IMMA, which proves successful adaptation to the personalized content. On the other side, the model with IMMA keeps the much lower CLIP similarity score at approximately 0.75, which indicates high resistance to adaptation.
- **DINO Similarity:** In the case of the DINO scores, the trend was the same. The model without IMMA had a linear growth in the similarity, while the IMMA-protected model showed constrained growth in the similarity itself, with the score remaining very low even after longer training. These results further establish IMMA as a very effective preventive measure against unauthorized personalization, in which the model cannot generate tailored outputs without drastically degrading general usability.

Conclusion

In this paper, we propose an in-depth reimplementation and investigation into the recently proposed IMMA framework for immunizing text-to-image models against unauthorized adaptations. The rapidly increasing accessibility to open-source generative models, along with powerful adaptation techniques, is an ethical and security concern since the models can be fine-tuned and used to generate harmful or unauthorized content. IMMA neutralizes such risks by immunizing the model itself, without recourse to data-based defenses; it thus adopts a proactive approach that constrains the model's vulnerability to malicious applications while maintaining its usability for benign ones.

We verified the efficacy of IMMA in two adaptation scenarios: relearning an erased artistic style, Van Gogh, and personalization unauthorized by the use of DreamBooth on a particular product concept-a purse on a beach. The results also indicate that the models immunized using IMMA consistently resist malicious adaptation attempts, which is manifested in lower similarity scores for CLIP and DINO compared to non-immunized models. Sketched below is a relearning experiment in the Van Gogh style. Results for the model protected by IMMA indicate that it had very little success in adapting and at the same time keeping low similarity, thus its generated images were highly divergent from the target style.

Similarly, in DreamBooth personalized content adaptation, the model protected by IMMA demonstrated high resistance, constraining adaptation even after an extended number of training epochs.

Future Directions

While our reimplementation and analysis of IMMA demonstrate its effectiveness in preventing unauthorized adaptations, there remain several areas for improvement and further exploration to enhance its robustness and versatility. Below, we outline potential directions for future research and development.

1. Extending IMMA to Handle Multiple Target Concepts In realistic settings, the models could require protection against multiple forms of misuse. Possibly the future work may be focused on methodology of expanding the framework of IMMA for multiple target concepts handled in one stroke, and further tuning of its parameters towards resisting a broader diversity of unauthorized adaptations with no performance compromise for benign tasks. This could be with regard to developing evermore complex optimization strategies toward achieving immunity across styles, objects, and various categories of content.

2. Improving Scalability and Computational Efficiency IMMA makes use of the bi-level optimization approach, which, although effective, is computationally expensive and requires considerable time for training, especially in high-resolution text-to-image models. The ability to make IMMA more scalable and computationally efficient could be explored further. For instance, through the application of meta-learning techniques, which include few-shot learning, and faster gradient-based optimization, it is possible to reduce the training time substantially without compromising the robustness of the framework. Besides, the exploration of improved performance in adaptation algorithms would ensue and make it more viable to apply IMMA on an extensive basis to larger models and datasets.

3. Enhancing Adaptability for Benign Applications One of the key research challenges in the current design of IMMA is how to balance immunity against harmful adaptations with flexibility for benign applications. In providing good performance degradation in malicious tasks, ideally, IMMA should have minimal impact on performance with respect to safe and intended uses. Development of selective applications of immunization, such as adaptive immunization, which only detects and responds to particular forms of misuse for which no legitimate adaptations exist, could be one future direction. Fine-grained control of adaptation parameters could further be explored to improve the model's constructive and ethical usability.

4. Integration with Other Defense Mechanisms IMMA is a model-centered defense strategy against malicious adaptation and thus hugely complementary to data-centered strategies such as poisoning of the data. Future work may explore hybrid mechanisms that combine IMMA with possible data-based strategies to create a potential layered security approach. For example, data poisoning together with model immunization may provide an extra layer of security, making it even harder for an attacker to fine-tune the models for unauthorized use.

5. Exploring IMMA's Application to Other Generative Models Analyzing IMMA in Other Generative Models While this work focused on adapting vulnerabilities regarding text-to-image models, the vulnerabilities concerning adaptation are common to other generative models, including text-to-text and even text-to-video models, and maybe music generation models. The principles of immunization in IMMA can be potentially extended to other domains, which makes it a versatile model-level security tool that can be applied to a wide range of generative AI applications. Future work will consider how to adapt the framework of IMMA to apply to other modalities, potentially opening directions toward cross-domain model immunization. 6. Robustness Against Advanced Adaptation Techniques The efficiency of IMMA against sophisticated and aggressive adaptation techniques is still an open problem. Future work may move along the lines of evaluation and resilience enhancement of IMMA against emerging techniques that are designed to be against model immunization. Testing could be done by advanced adaptation methods with either adversarial optimization or meta-adaptation strategies. In other words, prospects for further scaling up, making it adaptive, and integrating with other methods are ample; this can all come together to further fine-tune and extend the capabilities of IMMA toward what someday perhaps might come to finally make a safer, more responsible generative model possible.

Further investigation into combining these methods can lead to the development of a broader framework of protection for generative AI models.

References

[1] Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

[2] Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[3] Liang, C.; and Wu, X. 2023. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*.

[4] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

[5] 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.

[6] Shan, S.; Cryan, J.; Wenger, E.; Zheng, H.; Hanocka, R.; and Zhao, B. Y. 2023. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2187–2204.

[7] Zheng, A. Y.; and Yeh, R. A. 2024. IMMA: Immunizing Models against Malicious Adaptation. https://github.com/amberyzheng/IMMA.

[8] Zheng, A. Y.; and Yeh, R. A. 2025. Imma: Immunizing text-to-image models against malicious adaptation. In *European Conference on Computer Vision*, 458–475. Springer.