IMMA: Immunizing text-to-image Models against Malicious Adaptation

Amber Yijia Zheng Raymond A. Yeh

Department of Computer Science, Purdue University
{zheng709,rayyeh}@purdue.edu

Abstract. Advancements in open-sourced text-to-image models and fine-tuning methods have led to the increasing risk of malicious adaptation, *i.e.*, fine-tuning to generate harmful/unauthorized content. Recent works, *e.g.*, Glaze or MIST, have developed data-poisoning techniques which protect the data against adaptation methods. In this work, we consider an alternative paradigm for protection. We propose to "immunize" the model by learning model parameters that are difficult for the adaptation methods when fine-tuning malicious content; in short IMMA. Specifically, IMMA should be applied before the release of the model weights to mitigate these risks. Empirical results show IMMA's effectiveness against malicious adaptations, including mimicking the artistic style and learning of inappropriate/unauthorized content, over three adaptation methods: LoRA, Textual-Inversion, and DreamBooth. The code is available at https://github.com/amberyzheng/IMMA.

1 Introduction

With the open-source of large-scale text-to-image models [8, 38] the entry barrier to generating images has been drastically lowered. Building on top of these models, methods such as Textual Inversion [12], DreamBooth [39], and LoRA [20] allow quick adaptation to generate personalized content. These newly introduced capabilities come with great responsibility and trust in individuals to do good instead of harm to society. Unfortunately, the capabilities of adapting text-to-image generative models have already had negative impacts, *e.g.*, the generation of sexual content [16], copying artists' work without consent [17, 32], duplicating celebrity images [30], *etc*. We broadly encapsulate all these harmful fine-tuning of models under the term *malicious adaptation*.

To countermeasure these adaptations, open-source models have users agree that they will not use the software for "the purpose of harming others" in their licenses [6] or implement safety checks that would censor inappropriate generated content [36]. Nonetheless, these approaches do not have real enforcing power. Users can trivially disregard the license and remove the safety filters [45].

To address these loopholes, recent data poisoning techniques have shown a promising path towards preventing malicious adaptations [27, 41, 44]. The main idea is to *protect images* by modifying them with imperceivable changes, *e.g.*, adversarial noise, such that adaptation methods confuse the style and content of

the images and fail to generalize. A key shortcoming of these methods is that the burden of enforcement is on *content creators.* The artists need to apply these techniques before releasing their artwork. Contrarily, we present an alternative paradigm that places this burden on the *releaser of open-source models.*

In this paper, we propose to "**I**mmunnize" **M**odels, *i.e.*, make them more resilient, against **M**alicious **A**daptation; in short **IMMA**. The goal of IMMA is to make adaptation more difficult for concepts that are deemed malicious while maintaining the models' adaptability for other concepts. At a high level, we propose an algorithm to learn model parameters that perform *poorly* when being adapted to malicious concepts, *e.g.*, mimicry artistic style. In Fig. 1, we illustrate the effect of IMMA when mimicking Picasso's style; the model with IMMA fails to mimic the style.

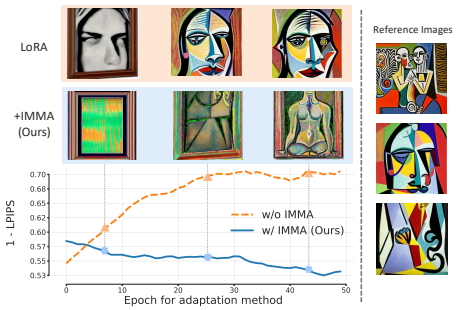

Fig. 1: IMMA on artistic style mimicry. Higher 1 − LPIPS indicates more similar to the reference images. IMMA successfully prevented the mimicking of the artistic style.

To demonstrate the effectiveness of IMMA, we consider three adaptation methods, namely, Textual Inversion [12], DreamBooth [39], and LORA [20]. We experimented with immunizing against several malicious settings, including, mimicking artistic styles, restoring erased concepts, and learning personalized concepts. Overall, IMMA successfully makes text-to-image models more resilient against adaptation to malicious concepts while maintaining the usability of the model. **Our contributions are as follows:**

- We propose a novel paradigm for preventing malicious adaptations. In contrast to the data poisoning for protection paradigm, we aim to protect the model instead of the data. See Fig. 2.
- We present an algorithm (IMMA) that learns difficult model initialization for the adaptation methods.
- We conduct extensive experiments evaluating IMMA against various malicious adaptations.

2 Related Work

Diffusion models. By learning to reverse a process of transforming data into noise, diffusion models achieve impressive generative capabilities [10, 19, 46]. With the aid of Internet-scale datasets [43], these models are capable of generating diverse images with high realism [7, 8, 31, 33, 35, 38, 40]. This progress led to new excitements in artificial intelligence generated content (AIGC) and interest in how to mitigate the associated risks, which we discuss next.

Preventing generative AI misuse. There are many potential risks associated with the advancement in generative capabilities [4, 34]. Recent works have started to address these risks. For example, Wang et al. [49] study how to detect

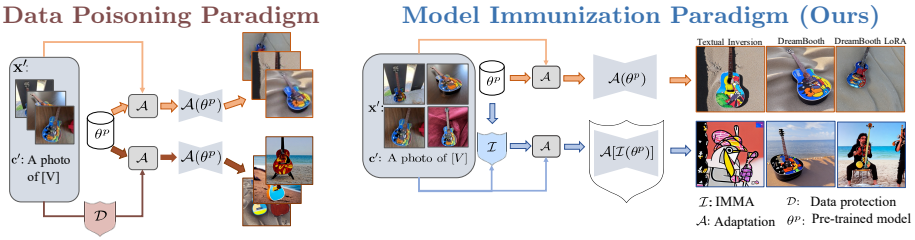

Fig. 2: Paradigms for preventing malicious adaptation. *Data poisoning:* modify training images $\mathbf{x}'$ with imperceivable changes, such that $\mathcal{A}$ fails to capture $\mathbf{c}'$ by training with modified images. *Model immunization (ours):* modify pre-trained model weights θ^p with immunization methods $\mathcal{I}$ before adaptation $\mathcal{A}$, such that $\mathcal{A}$ fails to capture $\mathbf{c}'$ by training on immunized model weights $\mathcal{I}(\theta^p)$.

unauthorized images that are used in the training set of these models. Schramowski et al. [42] study how to suppress the generation of inappropriate content, *e.g.*, nudity, self-harm, *etc.*, during diffusion's generation process. Other works [13, 14, 18, 23, 51], take a further step in trying to remove these inappropriate content from the diffusion models.

More closely related to this work, GLAZE [44], MIST [26, 27], and EUDP [53] studied the misuse of artistic mimicry, *i.e.*, preventing diffusion models from being used to copy artistic styles. Salman et al. [41] have also studied how to raise the cost of image manipulation using diffusion models. Notably, these works aim to protect the image, *i.e.*, a form of data-poisioning [3, 29] which modifies the data, *e.g.*, adding adversarial noise [15], such that the adaptation techniques fail. Different from these works, we *protect the model*, against misuse, *i.e.*, model immunization. We illustrate the difference between our proposed model immunization *vs*. data poisoning in Fig. 2.

Meta-learning. As our approach to model immunization is based on learning against an adaptation method, we briefly review meta-learning; also known as learning to learn. The area covers learning many aspects of a learning algorithm, *e.g.*, learning good initialization suitable to adapation [11], or other hyperparameters, *e.g.*, learning rate, or weight decay, via hypergradient [2, 25, 28, 37, 50, 54, 55]. Various hypergradients approximations have been proposed, *e.g.*, MAML [11] uses a single step gradient unrolling, with a summary provided by Lorraine et al. [28]. Different from these works, we aim to learn *poor* initializations for the adaptation methods to prevent the misuse of generative models.

3 Preliminaries

We briefly review the concepts necessary to understand our approach and introduce a common notation.

Text-to-image diffusion models. The goal of a text-to-image diffusion model [38] is to learn a conditional distribution of images $\mathbf{x}$ given concept embedding $\mathbf{c}$, *i.e.*, modeling $p(\mathbf{x}|\mathbf{c};\theta)$, where θ denotes the model parameters. The learning objective is formulated via variational lower bound [19, 21] or from a denoising

score-matching perspective [47, 48] which boils down to minimizing

$$L(\mathbf{x}, \mathbf{c}; \theta) = \mathbb{E}_{t, \epsilon \sim \mathcal{N}(0,I)} \left[w_t \| \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t) - \epsilon \|_2^2 \right], \tag{1}$$

where ϵ_θ denotes the denoising network, $\mathbf{x}_t$ and w_t denote the noised images and loss weights for a given time-step $t \sim \mathcal{U}\{0, T\}$ sampled from a discrete uniform distribution.

Concept erasing methods. To erase a target concept $\boldsymbol{c}'$ from a model, erasing algorithms [13, 23] fine-tunes a pre-trained model's parameters θ^p such that the model no longer generate images corresponding to that concept, *i.e.*,

$$p(\tilde{\boldsymbol{x}} | \mathbf{c}'; \theta^p_{-\mathbf{c}'}) \approx 0, \ \forall \tilde{\boldsymbol{x}} \sim p(\mathbf{x} | \mathbf{c}'; \theta^p), \tag{2}$$

where $\theta^p_{-\mathbf{c}'}$ denotes the parameters of the erased model after fine-tuning. The main idea is to train the target concept to generate images of some other concept. A common choice is an empty token, *i.e.*, the model performs unconditional generation when prompted with the target concept $\boldsymbol{c}'$.

Personalization of text-to-image diffusion models. In contrast to erasing a concept, personalization algorithms aim to *add* a novel concept to the pre-trained model. Given a set of images $\{\mathbf{x}'\}$ representative of the concept $\mathbf{c}'$. Personalization methods introduce a novel token $[V]$ in the word space and train the model to associate $[V]$ to the new concept. To learn this new concept, the model is trained using the data pair $(\mathbf{x}', \mathbf{c}')$, where $\mathbf{c}' \triangleq \Gamma([V])$ is extracted with the text encoder Γ using the novel word token $[V]$. We broadly use the notation $L_{\mathcal{A}}(\mathbf{x}', \mathbf{c}'; \theta, \phi)$ to denote the loss function of each adaptation method $\mathcal{A}$, where ϕ denotes the parameters that are being fine-tuned. For example, DreamBooth [24, 39] fine-tunes on a subset of parameters, *e.g.*, cross-attention layers, whereas Textual Inversion [12] only optimizes the new word token.

Another common approach to make the fine-tuning more data efficient is to use Low-Rank Adaptation (LoRA) [20]. Given a pre-trained weight $\theta^p \in \mathbb{R}^{n \times d}$, LoRA aims to learn an adaptor Δ, such that the final weights become $\hat{\theta} = \theta^p + \Delta$. LoRA specifically restricts the Δ to be low-rank, *i.e.*, $\Delta = \boldsymbol{A}\boldsymbol{B}$ where $\boldsymbol{A} \in \mathbb{R}^{n \times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r \times d}$ with $r \ll \min(n, d)$. In this work, we show that LoRA can easily learn back the concepts that were previously erased which highlights the need for our proposed research direction of model immunization.

4 Approach

Given a pre-trained text-to-image model with parameters θ^p, *e.g.*, StableDiffusion [38], existing adaptation methods $\mathcal{A}$ [12, 20, 24, 39] can fine-tune θ^p such that the model generates images of a concept $\mathbf{c}'$. Our goal is to *prevent* the adaptation methods from successfully doing so on harmful concepts, *e.g.*, unauthorized artistic style. To accomplish this, we present an algorithm IMMA $\mathcal{I}$ that takes pre-trained parameters θ^p as input, and outputs the immunized parameters θ^I. When applying adaptation $\mathcal{A}$ with θ^I, it should fail to learn the concept $\mathbf{c}'$, *i.e.*, the model is "immunized" against adaptation. At a high level, IMMA aims to learn a *poor model initialization* when being fine-tuned by the adaptation methods. We now describe the algorithmic details.

4.1 Model immunization

To achieve this goal of learning a poor initialization for adaptation, we propose the following bi-level program:

$$\overbrace{\max_{\theta \in \mathcal{S}} L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{I}}, \mathbf{c}'; \theta, \phi^\star)}^{\text{upper-level task}} \text{ s.t. } \phi^\star = \overbrace{\arg\min_{\phi} L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{A}}, \mathbf{c}'; \theta, \phi)}^{\text{lower-level task}}. \tag{3}$$

Here, the set $\mathcal{S}$ denotes a subset of θ that is trained by IMMA. This set is a hyperparameter that we choose empirically. Given a dataset of images $\mathcal{D} = \{\mathbf{x}'\}$ representative of the concept $\mathbf{c}'$, we sample the training data $\mathbf{x}'_{\mathcal{A}}$ and $\mathbf{x}'_{\mathcal{I}}$ independently. As reviewed in Sec. 3, ϕ denotes the parameters modified by the adaptation method $\mathcal{A}$. Note, ϕ may include newly introduced parameters by $\mathcal{A}$ or the parameters in the pre-trained model. We use the notation $\mathcal{U}$ to be the set of overlapping parameters between ϕ and θ.

Intuitively, the lower-level task simply performs the adaptation $\mathcal{A}$ given the model initialization θ by minimizing the loss function $L_{\mathcal{A}}$ with respect to (w.r.t.) ϕ. On the other hand, the upper-level task is *maximizing* the loss function of $\mathcal{A}$ w.r.t. θ. That is, the upper-level task aims for θ that results in *poor performance* when the adaptation method $\mathcal{A}$ is applied to the target concept $\mathbf{c}'$. The worse the performance is when adapted by $\mathcal{A}$, the more immunized a model is.

To solve the program in Eq. (3), we use gradient-based methods to solve the upper-level optimization. This leads to the following update steps:

$$\phi^\star \leftarrow \arg\min_{\phi} L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{A}}, \mathbf{c}'; \theta^{i-1}, \phi) \tag{4}$$

$$\theta^i \leftarrow \theta^{i-1} + \alpha \nabla_\theta L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{I}}, \mathbf{c}'; \theta^{i-1}, \phi^\star), \tag{5}$$

Algorithm 1 IMMA (Our method)

Input: pre-trained model θ^p, images $\mathcal{D} = \{\mathbf{x}'\}$ representative of the concept $\mathbf{c}'$, learning rate α, IMMA-modified parameters set $\mathcal{S}$, adaptation loss $L_{\mathcal{A}}$

Output: Immunized model θ^I

1: Initialize $\theta^0 = \theta^p$
2: Initialize ϕ^0 based on $\mathcal{A}$
3: **for** $i = 1$ to I **do**
4: Sample batch $\mathbf{x}'_{\mathcal{A}}$ and $\mathbf{x}'_{\mathcal{I}}$ from $\mathcal{D}$
5: $\phi \leftarrow \phi^{i-1}$ *# Initialize ϕ from the previous iteration*
6: $\phi^i \leftarrow \arg\min_\phi L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{A}}, \mathbf{c}'; \theta^{i-1}, \phi)$
7: $\theta^{i-1}_{\in \mathcal{U}} \leftarrow \phi^i$ *# Assign the overlaps between θ and ϕ*
8: $\theta^i_{\in \mathcal{S}} \leftarrow \theta^{i-1}_{\in \mathcal{S}} + \alpha \nabla_\theta L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{I}}, \mathbf{c}'; \theta^{i-1}, \phi^i)$
9: **end for**
10: **return** θ^I

where α denotes learning rate. We summarized the procedure in Alg. 1. Given the pre-trained model parameters θ^p, the algorithm returns an immunized parameter θ^I. We will next describe the subtle choices that were made regarding re-initialization and overlapping parameters during the optimization of θ and ϕ.

Details on updating ϕ and θ. In Alg. 1 line 5, in theory, we should reinitialize ϕ following the initialize scheme of $\mathcal{A}$, as the lower-level task is performing the adaptation. In practice, computing the lower-level task until convergence for each outer-loop iteration is prohibitively expensive. To reduce computation, we only solve the lower-level task with a fixed number of update steps. However, this leads to

lower-level tasks being not very well trained due to the small number of updates on ϕ. To address this, we initialize ϕ from ϕ^{i-1}, *i.e.*, the result from the previous outer-loop iteration. Empirically, this leads to faster convergence of the lower-level tasks; we suspect that this is because θ^i and θ^{i-1} remain quite similar after one outer-loop update.

The next subtle detail is when updating θ in Alg. 1 line 7. Recall, that depending on the adaptation method, the pre-trained parameters θ and adapted parameters ϕ may overlap. In such a scenario, there are two ways to compute the upper-level task's gradient: (a) we assume that θ^{i-1} and ϕ^i are separate parameters, *i.e.*, the result of ϕ^i will not change θ^{i-1}; (b) we update the overlapping parameters in $\mathcal{U}$ with the value from ϕ^i, leading to line 7 in Alg. 1. Empirically, we found (b) to perform better. We conduct an ablation study for Alg. 1 lines 5 and 7 in Sec. 5.3, where we found both to improve immunization quality.

Implementation details. To apply IMMA in practice, we approximate the solution of the lower-level task by taking a single gradient step for each of the adaptation methods. For the upper-level task, we use the Adam optimizer [22]. We choose $\mathcal{S}$ in Eq. (3) to contain only the cross-attention layer, *i.e.*, only these layers are being optimized. This choice follows the intuition that cross-attention layers are important as they mix the features of the target concept and image representation. Additional hyperparameters and experiment details are documented in the appendix.

4.2 Applications of model immunization

Immunizing concept erased models. Recent works [13, 23, 51], reviewed in Sec. 3, have shown that they can erase concepts from diffusion models without retraining the model from scratch. After a target concept $\mathbf{c}'$ is erased, the erased model $\theta^p_{-\mathbf{c}'}$ can no longer generate that object or style given the concept's text prompt. However, in our experiments, we show that the model can easily *re-learn* the target concept again in just a few training epochs by using LoRA [20]. In Fig. 3, we illustrate that the erased stable diffusion (ESD [13]) successfully removed a target concept of artists' style.

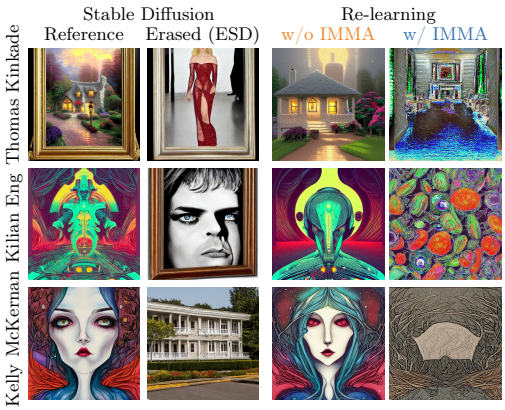

Fig. 3: IMMA's result against re-learning.

This motivated us to immunize the concept erased model $\theta^p_{-\mathbf{c}'}$ to make the adaptation of re-learning $\mathbf{c}'$ more difficult. Ideally, the immunized model is no longer able to generate images of the target concept, *i.e.*,

$$p(\tilde{\mathbf{x}}|\mathbf{c}'; \mathcal{A}[\mathcal{I}(\theta^p_{-\mathbf{c}'})]) \approx 0, \ \forall \tilde{\mathbf{x}} \sim p(\mathbf{x}|\mathbf{c}'; \theta^p). \tag{6}$$

Immunizing against personalization adaptation. Another potential for misuse is with personalization adaptation. Methods such as DreamBooth [39] or Textual Inversion [12] allow for a stable diffusion model to quickly learn to generate a personalized/unique concept given a few images of the unique concept. This motivated us to study immunization against these personalization adaptations. In practice, IMMA will be applied before a pre-trained model's release such that it fails to generate the unique concept even after adaptation.

5 Experiments

To evaluate IMMA, we conduct comprehensive experiments over multiple applications and adaptation methods.

5.1 Immunizing erased models against re-learning

Experiment setup. We employ LoRA as the adaptation method, and the pre-trained erased models are from ESD [13] using their publicly released code base. Following ESD, we consider experiments on eight artistic styles including both well-recognized and modern artists, and ten object classes from a subset of ImageNet [9]. For immunization, we use 20 images generated by Stable Diffusion (SD V1-4), prior to erasing, with the prompt of the target artistic style or object class. For adaptation, we generate *different* 20 images for each dataset and use LoRA to fine-tune the erased model for 20 epochs. For style, the prompt used in the evaluation is *"An artwork by {artist name}"*. For object, the prompt used in evaluation is *"a {concept}"*, for example, *"a parachute"*. We also evaluate on re-learning Not-Safe-For-Work (NSFW) content where we followed I2P prompts proposed by Schramowski et al. [42].

Evaluation metrics. To measure the effect of immunization against re-learning, we aim for a metric that measures the *performance gap* with and without IMMA, where the larger value indicates a stronger effect of IMMA, and vice versa. For this, we propose *Similarity Gap Ratio* (`SGR`). Let $\mathbf{x}_\mathcal{I}$ and $\mathbf{x}_\mathcal{A}$ denote the generated images with and without IMMA, and $\mathbf{x}_r$ to be the reference images of the target concept then `SGR` is defined as follows:

$$\texttt{SGR}(\mathbf{x}_\mathcal{I}, \mathbf{x}_\mathcal{A}, \mathbf{x}_r) = \frac{\mathcal{M}(\mathbf{x}_r, \mathbf{x}_\mathcal{A}) - \mathcal{M}(\mathbf{x}_r, \mathbf{x}_\mathcal{I})}{\mathcal{M}(\mathbf{x}_r, \mathbf{x}_\mathcal{A})}, \tag{7}$$

where $\mathcal{M}$ is an image similarity metric. Common choices, following prior works [12, 13, 39], include Learned Perceptual Image Patch Similarity [52] (LPIPS), and similarity measured in the feature space of CLIP [12] or DINO [5] each denoted as `SGR (L)`, `SGR (C)`, and `SGR (D)`. For consistency, we report *one minus* LPIPS, such that larger values for all three metrics mean higher image similarity.

User study. To check the quantitative metrics against human perception, we prepare four reference images and one pair of generated images (w/ and w/o IMMA) for the participants to select the generated image that is more similar to

the reference images in terms of content and quality.

Results on style. In Tab. 1, we report SGR for re-learning artist styles for erased models. Over eight artists, we consistently observe a positive gap among all three SGR based on LPIPS, CLIP, and DINO with averages of 21.84%, 5.5%, and 23.35%.

In Fig. 4, we directly show the LPIPS, CLIP, and DINO metrics at each epoch during the LoRA fine-tuning to visualize the gap. A lower value represents that the generated images are *less similar* to the reference images. Across all of the plots, we observe models without IMMA ↑ (orange) are more similar to the reference images than models with IMMA ↓ (blue). For artistic styles, the gap remains steady throughout the epochs. Overall, models with IMMA struggle to generate images containing the target concepts.

In Fig. 3, we provide qualitative comparisons. We observe that LoRA successfully re-learned the target concept (third column). On the other hand, the model with IMMA (last column) either generates an image with lower quality or an unrelated image. Our user study also validates this observation. All of the 30 respondents selected generations without IMMA as the one with high similarity and quality across all compared samples. This shows that models with IMMA generate worse images of the target styles.

Results on objects. As in artistic styles, we report SGR of re-learning target objects for the erased models in Tab. 2. The average SGR (L), SGR (C), and SGR (D) across ten classes are 21.84%, 11.41%, and 41.62%. We also

Table 1: SGR ↑(%) on artistic styles for ESD with LoRA adaptation.

SGR	Van Gogh	Pablo Picasso	Tyler Edlin	Kelly Mckernan	Kilian Eng	Claude Monet	Thomas Kinkade	Kirbi Fagan
(L)	17.61	28.08	28.67	23.34	26.78	31.14	18.47	16.18
(C)	4.77	4.56	5.87	9.34	4.59	7.27	9.44	1.18
(D)	18.4	21.81	26.05	14.99	25.73	31.48	39.59	15.07

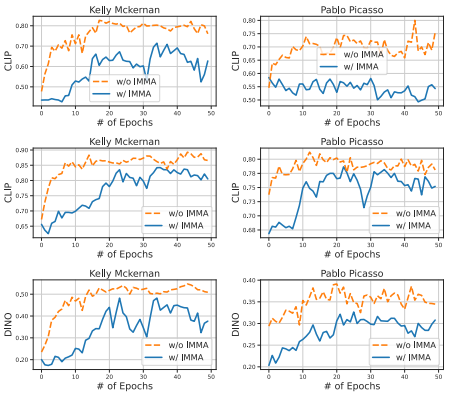

Fig. 4: Similarity *vs.* epochs for LoRA on styles. Models with IMMA achieve lower similarity throughout LoRA's epochs.

Table 2: SGR ↑(%) on objects for ESD with LoRA adaptation.

SGR	Cass. player	Garbage truck	Gas pump	Chain saw	En. springer	Golf ball	church	French horn	Parachute	Tench
(L)	11.66	7.48	16.64	27.33	22.42	41.96	10.09	6.88	50.97	39.14
(C)	3.66	5.18	19.72	8.72	15.08	10.67	12.51	8.11	19.34	11.07
(D)	12.16	20.0	31.31	50.86	68.85	58.09	14.33	30.25	68.56	61.84

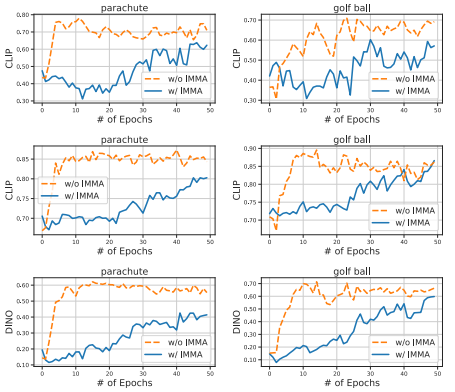

Fig. 5: Similarity *vs.* epochs for LoRA on objects. Models with IMMA achieve lower similarity throughout LoRA's epochs.

Table 3: Acc. (%) of object erased models on 500 images. *Col. 1:* Original ESD model with the target concept erased. *Col. 2:* ESD without IMMA after LoRA. We observe that the target concept is sucessfully relearned. *Col. 3:* ESD with IMMA after LoRA. *Col. 4 & 5:* Acc. of other objects of ESD before and after IMMA.

Class name /	Acc. of LoRA's target (↓)			Acc. of others (↑)	
Methods	ESD	w/o IMMA	w/ IMMA	ESD	w/ IMMA
Cassette player	0.2	2.0	0.2	72.0	46.4
Garbage truck	3.0	40.8	7.6	63.8	25.5
Gas pump	0.0	63.4	0.0	62.7	39.8
Chain saw	0.0	15.2	0.8	79.3	52.5
EN springer	0.4	15.6	0.8	67.2	49.6
Golf ball	0.4	22.4	0.0	56.4	36.7
Church	5.6	73.4	11.8	82.3	70.0
French horn	0.2	80.2	0.4	64.9	57.6
Parachute	2.0	91.0	0.0	78.8	58.9
Tench	1.0	50.8	0.4	78.0	55.5
Average	1.3	45.5	2.2	70.6	49.3

visualize LPIPS, CLIP, and DINO in Fig. 5. Overall, we observe the same trend as in the result for artistic styles. Similarity metrics across all the plots drop for models with IMMA. That is, with the same number of fine-tuning epochs, generations from models with IMMA exhibit lower quality or are less related to the object.

As the target concept contains objects, we consider classification accuracy (ResNet50 pre-trained on ImageNet) for evaluation reported in Tab. 3. First, without IMMA, ESD can relearn to generate the object. With a mere three epochs of LoRA, the average accuracy of the target concept increased from 1.3% to 45.5%. On the other hand, ESD with IMMA, the average accuracy remains low at 2.2%, demonstrating the effectiveness of IMMA at preventing relearning.

Thus far, the evaluation has focused on the *target concept* for IMMA models. We are also interested in how well IMMA preserves the *other concepts*. We define "other concepts" to be the remaining nine object categories beside the target object that is being adapted. In Tab. 3 (rightmost two columns), we observe that the original ESD has an average of 70.6% and after IMMA the accuracy dropped to 49.3%. We fully acknowledge that this is *a limitation* of IMMA. Prevention against certain target concepts may degrade other concepts. IMMA roughly trades off 43% in the target concept with 20% in other concepts. Finally, qualitative comparisons are shown in Fig. 6, where we observe the same conclusion that IMMA is effective against re-learning.

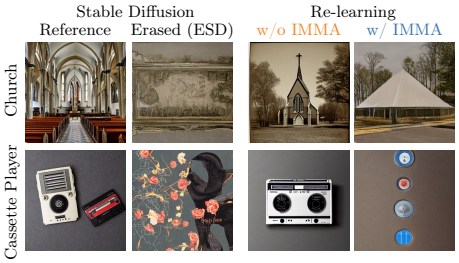

Fig. 6: Qualitative result of IMMA against re-learning.

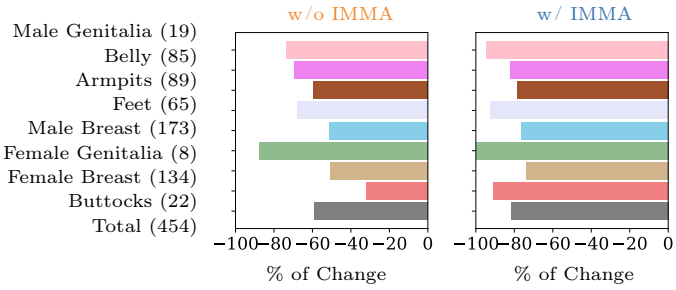

Fig. 7: IMMA on NSFW content. We report the % of change, relative to SD V1-4, in the number of detected nudity content after LoRA on the nudity-erased model.

Results on NSFW. For experiments on NSFW concepts, we use SD V1-4 to generate 4,703 unsafe images, among which 351 images contain 454 nudity counts based on NudeNet [1] with a threshold of 0.05. We randomly select two sets of 50 images containing nudity and their corresponding prompts for re-learning adaptation and the other for IMMA. The evaluation is conducted on the remaining 251 unsafe prompts.

In Fig. 7, we compare the percentage of change before and after IMMA adaptation with respect to the base SD V1-4 model. We observe that IMMA successfully generated less nudity content after LoRA, *i.e.*, w/ IMMA achieved a negative 80% of change in the detected nudity content compared to the negative 60% change for the model without IMMA. This shows that IMMA successfully immunized the model making it more difficult to re-learn nudity from unsafe images/prompts.

5.2 Immunizing against personalized content

Experiment setup. We consider three adaptation methods for learning new unique/personalized concepts: Textual Inversion (TI) [12], DreamBooth (DB) [39], and DreamBooth LoRA. For DreamBooth LoRA (DB LoRA), instead of modifying all of the parameters during fine-tuning, LoRA is applied on top of the cross-attention layers. We follow the exact adaptation procedures following prior works [12, 39], *i.e.*, adding a special token for the new unique concept. Note, we use *different* novel tokens during adaptation and IMMA. This is because, for a realistic evaluation, we would not know the novel token that would be used during the adaptation.

We perform experiments on ten different datasets released by Kumari et al. [24] which include a variety of unique objects. Each of them contains four to six images taken in the real world. The evaluation prompt for all concepts in this section is *"A [V] on the beach"* following DreamBooth.

Evaluation metrics. In this task, we report `SGR` in Eq. (7) with CLIP and DINO following DreamBooth [39]. Next, to show that the model maintains its capability to be personalized for *other concepts*, we propose *Relative Similarity Gap Ratio* (`RSGR`), which is given by

$$\texttt{RSGR}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{A}}, \mathbf{x}^o_{\mathcal{I}}, \mathbf{x}^o_{\mathcal{A}}) = \frac{\mathcal{M}(\mathbf{x}^o_{\mathcal{A}}, \mathbf{x}^o_{\mathcal{I}}) - \mathcal{M}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{I}})}{\mathcal{M}(\mathbf{x}^o_{\mathcal{A}}, \mathbf{x}^o_{\mathcal{I}})}, \tag{8}$$

Table 4: SGR (%) on personalized content adaptation. Higher positive values indicate better immunization quality.

		castle	car	chair	glasses	instrument	woodenpot	lighthouse	motorbike	houseplant	purse	**Average**
TI	SGR(C)	9.35	8.32	8.55	13.78	8.61	0.75	7.82	12.16	15.35	5.80	9.05
	SGR(D)	53.49	37.08	40.19	28.50	48.85	7.04	36.33	48.36	41.23	24.43	36.55
DB	SGR(C)	2.03	14.44	-8.60	0.91	2.18	5.80	-1.19	2.37	1.21	3.59	2.27
	SGR(D)	12.56	27.01	-14.67	13.37	24.64	29.31	-7.41	0.75	9.21	24.97	11.97
DB	SGR(C)	8.60	17.49	0.25	11.73	13.66	2.16	7.67	0.98	0.05	3.58	6.62
Lora	SGR(D)	43.21	34.37	8.94	37.63	40.38	32.70	48.36	6.40	24.88	0.52	27.74

Table 5: RSGR (%) on personalized content adaptation. Higher positive values indicate better performance at maintaining other concepts.

		castle	car	chair	glasses	instrument	woodenpot	lighthouse	motorbike	houseplant	purse	**Average**
TI	RSGR (C)	15.24	20.19	8.49	9.78	19.14	8.89	20.57	18.59	16.32	0.99	13.82
	RSGR (D)	70.69	63.57	22.5	40.1	57.66	34.15	51.85	67.18	55.22	8.88	47.18
DB	RSGR (C)	19.66	13.63	6.62	4.97	4.44	4.61	16.53	6.08	5.12	10.36	9.20
	RSGR (D)	36.99	35.98	17.72	32.04	24.59	23.28	35.06	2.60	25.05	50.1	28.34
DB	RSGR (C)	17.55	21.00	4.49	3.89	12.24	8.42	24.28	8.67	11.77	4.83	11.71
LoRA	RSGR (D)	49.66	41.57	18.9	25.41	29.92	35.92	53.12	28.63	41.02	24.74	34.89

where $\mathbf{x}^o_{\mathcal{A}}$ and $\mathbf{x}^o_{\mathcal{I}}$ are generated images without and with IMMA on *other unique concepts*.

The term $\mathcal{M}(\mathbf{x}^o_{\mathcal{A}}, \mathbf{x}^o_{\mathcal{I}})$ captures image similarity for other concepts with and without IMMA. Ideally, this term should be high as IMMA should not affect other concepts. The term $\mathcal{M}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{I}})$ captures the image similarity for target concepts with and without IMMA. In this case, the similarity should be low. RSGR reports the difference between these two terms as a ratio. Intuitively, larger RSGR indicates IMMA is better at preserving the other concepts while removing the target concept.

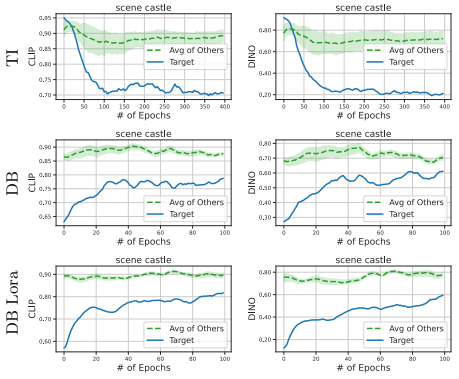

Fig. 8: CLIP and DINO similarity on other concept *vs*. target concept. The gap between the two lines shows RSGR.

We also conducted a user study for personalized adaptation using the setting as in Sec. 5.1.

Results on personalization. In Tab. 4 we observe a positive ratio among most of the SGR metrics, except for "furniture chair" and "lighthouse" with Dreambooth highlighted in red. We show the generations with negative SGR in the appendix. Overall, IMMA effectively prevents the pre-trained model from learning personalization concepts across the three adaptation methods.

Next, Tab. 5 reports RSGR to evaluate how well IMMA preserves the ability to personalize other concepts. As we can see, the RSGR values are consistently positive across all the datasets. This indicates IMMA immunizes against the target concept without hurting the adaptability for personalizing for other concepts. We directly visualize this relative gap in Fig. 8. As shown, the lines of the nine concepts

Fig. 9: Personalization adaptation w/ and w/o IMMA.

($\mathcal{M}(\mathbf{x}^o_{\mathcal{A}}, \mathbf{x}^o_{\mathcal{I}}) \uparrow$ in green) and the adaptation of the target concept ($\mathcal{M}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{I}}) \downarrow$ in blue) can be easily distinguished; consistent with results in Tab. 5.

Lastly, we show the generated images in Fig. 9. Comparing the generated images with and without IMMA, we observe models with IMMA are either unable to learn the target concept or generate unrealistic images. We also conducted a user study to validate this observation. All of the 30 participants selected generation without IMMA to be more similar to reference images, *i.e.*, models after IMMA fail to generate the target concepts.

5.3 Additional discussion

Comparison with data poisoning. We compare IMMA with MIST [27], one of the data poisoning (DP) methods in the personalized content setup. In Fig. 10 (top-row) we show both MIST and IMMA successfully prevent the model from learning the target concept after Textual Inversion.

This observation is also supported by a user study. On seven out of ten evaluated datasets, the majority of the 30 users found the generation without MIST to be of higher quality and more similar to the reference images.

Adaption on images with JPEG compression. As reported in MIST [27], one way to weaken the effect of poisoned data is by compressing the images with JPEG after adaptation. On the other hand, IMMA can defend against JPEG compression by including JPEG images in the training data. In Fig. 10 (bottom row), we observe that after compression, MIST fails to prevent generating the target concept, while IMMA remains robust against JPEG.

Fig. 10: MIST vs. IMMA. Generation with Textual Inversion on training images w/ and w/o JPEG compression.

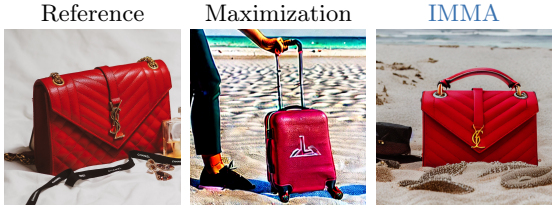

Fig. 11: Ablation on direct maximization. DreamBooth adaptation on *"luggage purse"* after immunization on *"woodenpot"*.

Ablation studies. We conduct ablations using the DreamBooth personalization setup in Sec. 5.2. First, we experiment with a direct maximization baseline, *i.e.*, only the upper-level task in Eq. (3). Results are shown in Fig. 11. We observe that direct maximization ruins the immunized model, *i.e.*, low image quality when being adapted for another concept.

Next, we ablate line 5 and line 7 in Alg. 1 and report CLIP similarity. Shown in Fig. 12 (left), without line 5, the adaptation can learn the target concept, *i.e.*, high CLIP similarity. Next, we ablate whether to update the overlapping parameters in ϕ by removing line 7. The result is shown in Fig. 12 (right). We observe that without line 7, the adaptation successfully learns the target concept. These results show the necessity of the proposed steps in Alg. 1.

Crossed adaptation with IMMA. Thus far, we report results for IMMA by immunizing the model against *the same* adaptation method $\mathcal{A}$. We now investigate whether IMMA remains effective under a *different* adaptation method during IMMA and adaptation. In other words, we consider IMMA with *crossed adaptation methods*, where we immunize the pre-trained model using $\mathcal{A}_1$ and perform malicious adaptation on $\mathcal{A}_2$. Fig. 13 shows the qualitative results across DB and TI. We observe that the model with IMMA against DreamBooth is also effective when being adapted with Textual Inversion, and vice versa.

Limitations. Our proposed IMMA and experiments focus on the immunization of a single concept. In this work, we methodically study a variety of adaptation

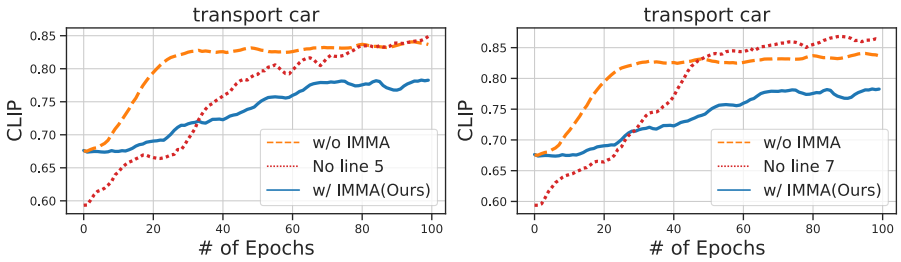

Fig. 12: Ablation on line 5 and line 7 of Alg. 1.

Fig. 13: Results on crossed adaptation immunization. *First row:* DB after IMMA with either DB or TI. *Second row:* TI after IMMA the model with either DB or TI.

settings and leave the support for multiple concepts for future work. We believe that results in a single concept demonstrate a convincing step towards model immunization. If the reader is interested, we present preliminary results in the Appendix showing that IMMA can generalize to multiple concepts, albeit, not as comprehensive as the single concept study.

6 Conclusion

We propose to Immunize Models against Malicious Adaptation (IMMA). Unlike the data-poisoning paradigm, which protects images, our method focuses on protecting pre-trained models from being used by adaptation methods. We formulate "Immunization" as a bi-level optimization program to learn a poor model initialization that would make adaptation more difficult. To validate the efficacy of IMMA, we conduct extensive experiments on relearning concepts for erased models and immunizing against the adaptation of personalized content. We believe that model immunization is a promising paradigm for combatting the risk of malicious adaptation, and that IMMA is an encouraging first step. We are hopeful that the advancement of IMMA will result in safer open-source text-to-image models, benefiting both the research community and society.

Acknowledgements

This project is supported in part by an NSF Award #2420724. We thank Renan A. Rojas-Gomez for proofreading and helpful discussions.

References

1. Bedapudi, P.: NudeNet: Neural nets for nudity classification, detection and selective censoring. https://github.com/platelminto/NudeNetClassifier (2019)
2. Bengio, Y.: Gradient-based optimization of hyperparameters. Neural Computation (2000)
3. Biggio, B., Nelson, B., Laskov, P.: Support vector machines under adversarial label noise. In: Proc. ACML (2011)
4. Bird, C., Ungless, E., Kasirzadeh, A.: Typology of risks of generative text-to-image models. In: Proc. AIES (2023)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proc. CVPR (2021)
6. CreativeML Open RAIL-M: Stable Diffusion LICENSE File (2023), URL https://github.com/CompVis/stable-diffusion/blob/main/LICENSE
7. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., Yu, M., Kadian, A., Radenovic, F., Mahajan, D., Li, K., Zhao, Y., Petrovic, V., Singh, M.K., Motwani, S., Wen, Y., Song, Y., Sumbaly, R., Ramanathan, V., He, Z., Vajda, P., Parikh, D.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
8. DeepFloyd Lab at StabilityAI: DeepFloyd IF. https://github.com/deep-floyd/IF (2023)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc. CVPR (2009)
10. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Proc. NeurIPS (2021)
11. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proc. ICML (2017)
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: Proc. ICLR (2023)
13. Gandikota, R., Materzyńska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proc. ICCV (2023)
14. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proc. WACV (2024)
15. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proc. ICLR (2015)

16. Harwell, D.: AI-generated child sex images spawn new nightmare for the web. The Washington Post (2023), URL https://www.washingtonpost.com/technology/2023/06/19/artificial-intelligence-child-sex-abuse-images/
17. Heikkilä, M.: This artist is dominating ai-generated art. and he's not happy about it. MIT Technology Review. Retrieved March **16**, 2023 (2022)
18. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. arXiv preprint arXiv:2305.10120 (2023)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proc. NeurIPS (2020)
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proc. ICLR (2022)
21. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. In: Proc. NeurIPS (2021)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. ICLR (2015)
23. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proc. ICCV (2023)
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proc. CVPR (2023)
25. Larsen, J., Hansen, L.K., Svarer, C., Ohlsson, M.: Design and regularization of neural networks: the optimal use of a validation set. In: IEEE Signal Processing Society Workshop (1996)
26. Liang, C., Wu, X.: Mist: Towards improved adversarial examples for diffusion models (2023)
27. Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., Guan, H.: Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In: Proc. ICML (2023)
28. Lorraine, J., Vicol, P., Duvenaud, D.: Optimizing millions of hyperparameters by implicit differentiation. In: Proc. AISTATS (2020)
29. Mei, S., Zhu, X.: Using machine teaching to identify optimal training-set attacks on machine learners. In: Proc. AAAI (2015)
30. Moore, S.: Can the law prevent AI from duplicating actors? It's complicated. Forbes (Jul 2023), URL https://www.forbes.com/sites/schuylermoore/2023/07/13/protecting-celebrities-including-all-actors-from-ai-with-the-right-of-publicity/?sh=5c56ba4159ec
31. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: Proc. ICML (2022)
32. Noveck, J., O'Brien, M.: Visual artists sue ai companies in sf federal court for repurposing their work. Associated Press (2023)
33. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

34. Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., Zhang, Y.: Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. arXiv preprint arXiv:2305.13873 (2023)
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)
36. Rando, J., Paleka, D., Lindner, D., Heim, L., Tramèr, F.: Red-teaming the stable diffusion safety filter. In: Proc. NeurIPS ML Safety Workshop (2022)
37. Ren, Z., Yeh, R., Schwing, A.: Not all unlabeled data are equal: Learning to weight data in semi-supervised learning (2020)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. CVPR (2022)
39. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proc. CVPR (2023)
40. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Proc. NeurIPS (2022)
41. Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., Madry, A.: Raising the cost of malicious AI-powered image editing. In: Proc. ICML (2023)
42. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proc. CVPR (2023)
43. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models. In: Proc. NeurIPS (2022)
44. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by text-to-image models. In: USENIX Security Symposium (2023)
45. SmithMano: Tutorial: How to remove the safety filter in 5 seconds. Reddit (2022)
46. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proc. ICML (2015)
47. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Proc. NeurIPS (2019)
48. Vincent, P.: A connection between score matching and denoising autoencoders. Neural computation (2011)
49. Wang, Z., Chen, C., Liu, Y., Lyu, L., Metaxas, D., Ma, S.: How to detect unauthorized data usages in text-to-image diffusion models. arXiv preprint arXiv:2307.03108 (2023)

50. Yeh, R.A., Hu, Y.T., Hasegawa-Johnson, M., Schwing, A.: Equivariance discovery by learned parameter-sharing. In: Proc. AISTATS (2022)
51. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2211.08332 (2023)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. CVPR (2018)
53. Zhao, Z., Duan, J., Hu, X., Xu, K., Wang, C., Zhang, R., Du, Z., Guo, Q., Chen, Y.: Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. arXiv preprint arXiv:2306.01902 (2023)
54. Zheng, A.Y., He, T., Qiu, Y., Wang, M., Wipf, D.: Graph machine learning through the lens of bilevel optimization. In: Proc. AISTATS (2024)
55. Zheng, A.Y., Yang, C.A., Yeh, R.A.: Learning to obstruct few-shot image classification over restricted classes. In: Proc. ECCV (2024)