Multi-concept Model Immunization through Differentiable Model Merging

Amber Yijia Zheng, Raymond A. Yeh

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

Abstract

Model immunization is an emerging direction that aims to mitigate the potential risk of misuse associated with open-sourced models and advancing adaptation methods. The idea is to make the released models' weights difficult to fine-tune on certain harmful applications, hence the name "immunized". Recent work on model immunization focuses on the *single-concept* setting. However, models need to be immunized against multiple concepts in real-world situations. To address this gap, we propose an immunization algorithm that, simultaneously, learns a single "difficult initialization" for adaptation methods over *a set of concepts*. We achieve this by incorporating a differentiable merging layer that combines a set of model weights adapted over multiple concepts. In our experiments, we demonstrate the effectiveness of multi-concept immunization by generalizing prior work's experiment setup of re-learning and personalization adaptation to multiple concepts.

Project page — https://www.amberyzheng.com/mima

1 Introduction

With the advancements in effective adaptation techniques, such as DreamBooth (Ruiz et al. 2023) or Textual Inversion (Gal et al. 2023), there are increasing risks of misuse for open-sourced text-to-image models. As the models are released, an ill-intended person can leverage adaptation methods to tune unsafe content into the models and perform malicious acts, *e.g.*, generating unsafe or sexual content (Harwell 2023).

To tackle these risks, Zheng and Yeh (2024) propose to "immunize" the open-sourced models before releasing them. The idea is to learn models that are resistant ("immuned") to adaptations on harmful concepts. Zheng and Yeh (2024) refer to their approach as *Immunizing text-to-image Models against Malicious Adaptation*, in short, IMMA. While IMMA shows a promising direction for mitigation, IMMA's method and experiments focus on the immunization of a *single concept.* However, in most practical settings, a model needs to be immunized against multiple harmful concepts.

To address this gap, we study Multi-concept Immunization against Malicious Adaptation (MIMA). We aim to make a *single model* resilient to adaptation on more than one concept. We propose a model immunization algorithm that meta-learns a "difficult initialization" for adaptation methods over *a set of concepts* formulated as a bi-level optimization with multiple lower-level tasks. We accomplish this by introducing a differentiable model merging layer that combines the individual lower-level task's weights from each target concept. The bi-level optimization is solved by backpropagating through this merging layer to immunize the model over a set of concepts. This approach is inspired by the success of model merging for multi-concept customization (Kumari et al. 2023b), we hypothesize that model merging would also benefit immunization as it captures the relationships among concepts.

Empirically, we experiment with several adaptation methods, including, Textual Inversion (Gal et al. 2023), DreamBooth (Ruiz et al. 2023), LoRA (Hu et al. 2022), and CustomDiffusion (Kumari et al. 2023b) over two applications: (a) restoring erased concepts such as artistic styles or object categories, and (b) learning personalized concepts. We found that MIMA successfully immunizes a model against *multiple* malicious concepts and outperforms IMMA-inspired baselines.

Our contributions are summarized as follows:

- We generalize the task of model immunization from a single concept to multiple concepts that more closely match the real-world scenario.
- We propose MIMA, a novel model immunization algorithm for multi-concept immunization. MIMA leverages a differentiable model merging layer that combines multiple adapted weights, enabling backpropagation to meta-learn an immunized model.
- We conduct experiments over two tasks and four adaptation methods demonstrating the efficacy of MIMA.

2 Related Work

Towards safer generative AI. Several directions have been proposed to make generative AI safer. One direction that has received attention is removing inappropriate content from pre-trained models (Schramowski et al. 2023; Gandikota et al. 2023, 2024; Zhang et al. 2024; Kumari et al. 2023a; Heng and Soh 2023). Another direction is to protect the data sources by using adversarial examples (Goodfellow, Shlens, and Szegedy 2015) to achieve data-poisoning (Biggio, Nelson, and Laskov 2011; Mei and Zhu 2015), such that when adapted on these protected images, the diffusion model fails (Shan et al. 2023; Liang et al. 2023; Liang and Wu 2023;

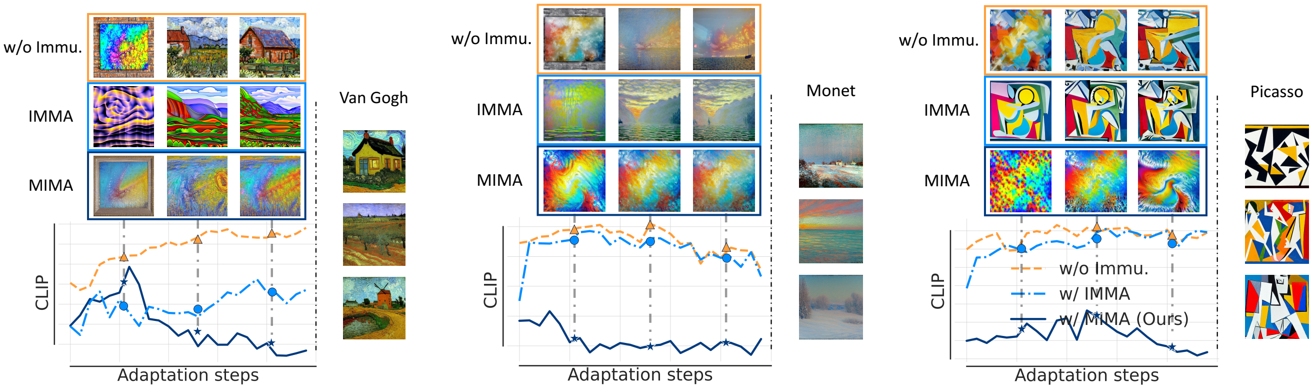

Figure 1: We propose MIMA an immunize algorithm that protects a model against the adaptation on harmful concepts. Here we show an experiment on immunization against the re-learning of multiple artistic styles and report the CLIP similarity between the generations and the target concept at different adaption steps. A lower CLIP similarity indicates a more effective immunization, as the images semantically differ more from the references. As can be seen, MIMA offers protection over all three concepts of Van Gogh, Monet, and Picasso. In comparison, IMMA (Zheng and Yeh 2024), designed to immunize over a single concept, only offers protection against the re-learning of Van Gogh.

Zhao et al. 2024). However, these approaches have limitations when dealing with open-sourced models. For content removal, adaptation methods can quickly relearn the removed content. For data poisoning techniques, it requires poisoning the content which may not be feasible depending on the content sources.

Most related to our work is the model immunization paradigm introduced by Zheng and Yeh (2024) which aims to learn poor initialization such that adaptation methods fail on a single concept. In this work, we generalize the study of model immunization to multi-concept and propose an algorithm MIMA that incorporates a differentiable model merging layer to enable multi-concept immunization. Also, model immunization has been considered in the few-shot classification setting (Zheng, Yang, and Yeh 2024).

Model adaptation and editing methods. With the open-source of high-quality text-to-image models, *e.g.*, Stalbe Diffusion (Rombach et al. 2022; DeepFloyd Lab at StabilityAI 2023), there is a surge of interest in how to adapt these pre-trained models for different applications, *e.g.*, adapting the model for a customized generation of personalized items (Ruiz et al. 2023; Gal et al. 2023; Kumari et al. 2023b), efficient fine-tuning of these models (Hu et al. 2022), or adding extra control to the generation (Zhang, Rao, and Agrawala 2023). Closely related is model editing which aims to directly modify model parameters to achieve new generative capabilities (Bau et al. 2020; Gal et al. 2022; Kumari et al. 2023b; Nitzan et al. 2023).

Optimization layers. We briefly discuss optimization layers as the differentiable model merging can be viewed as an optimization layer. The literature of optimization layers views an optimization problem as a differentiable function, *i.e.*, mapping its input to its exact solution (Domke 2012; Amos and Kolter 2017; Ren, Yeh, and Schwing 2020; Gould, Hartley, and Campbell 2021; Agrawal et al. 2019; Liu et al. 2023). Depending on the exact optimization program, the gradient of this mapping can either be computed analytically or via implicit differentiation. Optimization layers have found applications in AI (Amos and Kolter 2017; Tschiatschek, Sahin, and Krause 2018; Wang et al. 2019; Amos et al. 2018; Zheng et al. 2024) and computer vision (Yeh et al. 2022; Hu, Schwing, and Yeh 2023; Bai, Kolter, and Koltun 2019; Bai, Koltun, and Kolter 2020; Rolínek et al. 2020; Wang, Teng, and Wang 2023; Geng, Pokle, and Kolter 2023; Zheng, Yang, and Yeh 2024).

3 Background

Model immunization. Given pre-trained diffusion model weights θ^p, an adaptation method $\mathcal{A}$, and its corresponding loss function $L_\mathcal{A}$, IMMA (Zheng and Yeh 2024) aims to prevent $\mathcal{A}$ from fine-tuning θ^p, such that the fine-tuned model fails to generate images of a single target (harmful) concept $\mathbf{c}'$. IMMA is formulated as a bi-level optimization with the following objective:

$$\underbrace{\max_{\theta\in\mathcal{S}} L_\mathcal{A}(\mathbf{x}'_\mathcal{I},\mathbf{c}';\theta,\phi^\star)}_{\text{upper-level task}} \text{ s.t. } \phi^\star = \underbrace{\arg\min_{\phi} L_\mathcal{A}(\mathbf{x}'_\mathcal{A},\mathbf{c}';\theta,\phi)}_{\text{lower-level task}}. \quad (1)$$

Intuitively, the upper-level task aims to find the worst parameters θ for the adaptation algorithm $\mathcal{A}$ by taking into the $\mathcal{A}$'s update in the lower-level task. Note that, IMMA requires the following: (a) $\mathcal{A}$ is known at immunization time, and (b) there is a single-concept $\mathbf{c}'$ to be immunized. In this work, we propose an immunization algorithm that does not require $\mathcal{A}$ and immunizes a model over *a set* of concepts.

Model merging. To achieve a unified fine-tuned model capable of generating multiple target concepts, one approach is to fine-tune multiple models for each concept and combine them. Kumari et al. (2023b) propose to merge models by solving an optimization problem, where the keys and values of cross-attention weights (Dosovitskiy et al. 2021; Vaswani et al. 2017) are modified then merged. Recall, cross attention layers take in text embeddings $\mathbf{c} \in \mathbb{R}^{l\times c}$, where

l is the number of tokens with c denoting the embedding size, and project them into keys, values with the projection matrics $\mathbf{W}^k \in \mathbb{R}^{c\times d}$ and $\mathbf{W}^v \in \mathbb{R}^{c\times d'}$, where d and d' are the dimensions of the keys and values. We subsume these projection matrices into a compact notation $\mathbf{W}$.

Given N model weights $\{\mathbf{W}_{[n]}\}_{n=1}^{N}$, that has each been adapted to a concept embedding $\mathbf{c}_{[n]}$ from the target concept set $\mathcal{C}$, model merging aims to find a single optimal weight $\varphi^\star$ that mimics the mapping of all weights on their corresponding concepts while maintaining proximity to a set of N' regularization concepts $\mathcal{C}_{\text{reg}}$. This merging process is formulated as a constrained optimization program:

$$\varphi^\star = \arg\min\nolimits_\varphi ||\mathbf{C}_{\text{reg}}\varphi - \mathbf{C}_{\text{reg}}\mathbf{W}^p||_F^2 \;\text{ s.t. }\; \mathbf{C}\varphi = \mathbf{O}^*, \quad (2)$$

$$\text{where } \mathbf{C} \triangleq \texttt{Concat}(\mathcal{C}) \in \mathbb{R}^{(N\cdot l)\times c}\,,$$

$$\mathbf{C}_{\text{reg}} \triangleq \texttt{Concat}(\mathcal{C}_{\text{reg}}) \in \mathbb{R}^{(N'\cdot l)\times c}\,,$$

$$\text{and } \mathbf{O}^* \triangleq \texttt{Concat}\left(\{\mathbf{c}_{[n]}\mathbf{W}_{[n]}\}_{n=1}^{N}\right) \in \mathbb{R}^{(N\cdot l)\times d}.$$

Here $\texttt{Concat}(\cdot)$ concatenates a set $\mathcal{S}$ of embeddings into a stack of embeddings $\mathbf{S} = [\mathbf{s}_{[1]}^\top, \cdots, \mathbf{s}_{[N]}^\top]^\top$ and $\mathbf{W}^p$ corresponds to the pre-trained weight of a model before the adaptation.

The constraint matches the output of φ on target concepts fine-tuned for each of the $\boldsymbol{W}_{[n]}$. Next, the objective maintains the model's generation capability on the other concepts not in $\mathcal{C}$. It encourages the model φ to be similar to the corresponding pre-trained weight $\mathbf{W}^p$ on a set of regularization concepts $\mathcal{C}_{\text{reg}}$.

4 Approach

We introduce Multi-concept model Immunization against Malicious Adaptation (MIMA). As the name suggests, the goal is to protect a pre-trained model with weights θ^p from being fine-tuned by adaptation methods to generate images containing harmful concepts. Formulated as a bi-level optimization, MIMA meta-learns a difficult initialization for downstream adaptation methods on all the concepts within a target concept set. The key idea is to treat model merging as a differentiable optimization layer that allows for gradients to be backpropagated through model merging to provide updated directions for immunization. Please see overview in Fig. 2.

4.1 Multi-concept Model Immunization

Problem formulation. As in IMMA (Zheng and Yeh 2024), the general immuimization process is formulated as a bi-level optimization problem. Given pre-trained model weights θ^p, a set of target concept embeddings $\mathcal{C} = \{\mathbf{c}_{[n]}\}_{n=1}^{N}$, and the image set $\mathcal{X} = \cup_n \mathcal{X}_{[n]}$ where $\mathcal{X}_{[n]} = \{\mathbf{x}_{[n]}\}$ is a set of images representative of the concept $\mathbf{c}_{[n]}$, we optimize:

$$\underbrace{\max_{\theta\in\mathcal{S}^u} \sum_{n=1}^{|\mathcal{C}|} L(\mathbf{x}^u_{[n]}, \mathbf{c}_{[n]}; \texttt{Merge}\left(\left\{\theta'_{[n]}\right\}\right)),}_{\text{upper-level task}}$$

$$\underbrace{\text{s.t. } \; \theta'_{[n]} \triangleq \arg\min_{\theta\in\mathcal{S}^l} L(\mathbf{x}^l_{[n]}, \mathbf{c}_{[n]}; \theta) \;\; \forall n,}_{\text{multiple lower-level tasks}} \quad (3)$$

where $\mathbf{x}^u_{[n]}$ and $\mathbf{x}^l_{[n]}$ are independently sampled from $\mathcal{X}_{[n]}$. The sets $\mathcal{S}^u$ and $\mathcal{S}^l$ denote the subset of model parameters that are being updated in the upper and lower tasks. Next, the $\texttt{Merge}$ function, defined formally in Eq. (5), combines a set of model weights into a single model, and L denotes the standard loss for training a diffusion model given by

$$L(\mathbf{x}, \mathbf{c}; \theta) = \mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,I)}\left[w_t\|\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t) - \epsilon\|_2^2\right]. \quad (4)$$

Here, ϵ_θ is the denoising network with weights θ conditioned on the timestep t sampled from a discrete uniform distribution, $\mathbf{x}_t$ is the noisy image, and w_t is a loss weight.

To achieve multi-concept immunization, the upper-level task aims to make the standard diffusion loss (Eq. (4)) high when being adapted to any of the target concepts. Hence, in the lower-level tasks, we perform a set of updates, one for each concept $\boldsymbol{c}_{[n]}$, leading to N different models $\theta'_{[n]}$.

However, model immunization requires a single model. It is unclear how to perform the maximization in the upper-level task given N separate models. Specifically, we need to merge these N models into one. The main challenges are: **(a)** How to merge these N models? The procedure proposed by Kumari et al. (2023b) only merges the projection matrices of keys and values, what about the other parameters? **(b)** How do we backpropagate through the model merging operation, such that gradient-based optimization can be performed? We now answer these two questions.

Differentiable model merging. To merge the weights that are fine-tuned on different concepts, we split the model parameters into two sets: key and value project matrices subsumed in $\mathcal{W}$ and the rest $\notin \mathcal{W}$. For parameters within $\mathcal{W}$, we combine the parameters following the optimization in Eq. (3), and for the other parameters, we perform a simple average. More formally, the $\texttt{Merge}$ operation is defined as:

$$\theta' \triangleq \texttt{Merge}\left(\{\theta'_{[n]}\}_{n=1}^{N}\right) \triangleq \begin{cases} \varphi^\star(\{\theta'_{[n],\in\mathcal{W}}\}) \\ \frac{1}{N}\sum_{n=1}^{N}\theta'_{[n],\notin\mathcal{W}} \end{cases}, \quad (5)$$

where we view the solution $\varphi^\star$ in Eq. (3) *as a function of the input* model weights to the optimization problem. The reason why we perform a simple average over the parameters $\notin \mathcal{W}$ is that they are shared across the different lower-level tasks. With a simple average, *i.e.*, the gradients will also be shared.

To backpropagate through Eq. (5), we need to compute the gradient through $\varphi^\star$. To obtain $\varphi^\star$ from Eq. (3), we can solve its Lagrange form of:

$$L(\varphi, \mathbf{M}) = \left\|\mathbf{C}_{\text{reg}}\varphi - \mathbf{C}_{\text{reg}}\mathbf{W}^p\right\|_F^2 - \text{tr}\big((\mathbf{C}\varphi - \mathbf{O}^*)\mathbf{M}^\top\big) \quad (6)$$

where $\mathbf{M} \in \mathbb{R}^{(N\times l)\times d}$ represents the matrix of Lagrange multipliers associated with the constraint. Taking a closer look at Eq. (7), we can express it as a linear system $\boldsymbol{Q}\varphi = \boldsymbol{t}$ with

$$L(\varphi, \mathbf{M}) = \left\|\mathbf{C}_{\text{reg}}\varphi - \mathbf{C}_{\text{reg}}\mathbf{W}^p\right\|_F^2 - \text{tr}\big((\mathbf{C}\varphi - \mathbf{O}^*)\mathbf{M}^\top\big) \quad (7)$$

In other words, the solution has the form

$$\varphi^\star = \boldsymbol{Q}^{-1}\boldsymbol{t} \;\text{ and }\; \frac{\partial\mathcal{L}}{\partial\varphi^\star} = \boldsymbol{Q}^\intercal\frac{\partial\mathcal{L}}{\partial\boldsymbol{t}} \quad (8)$$

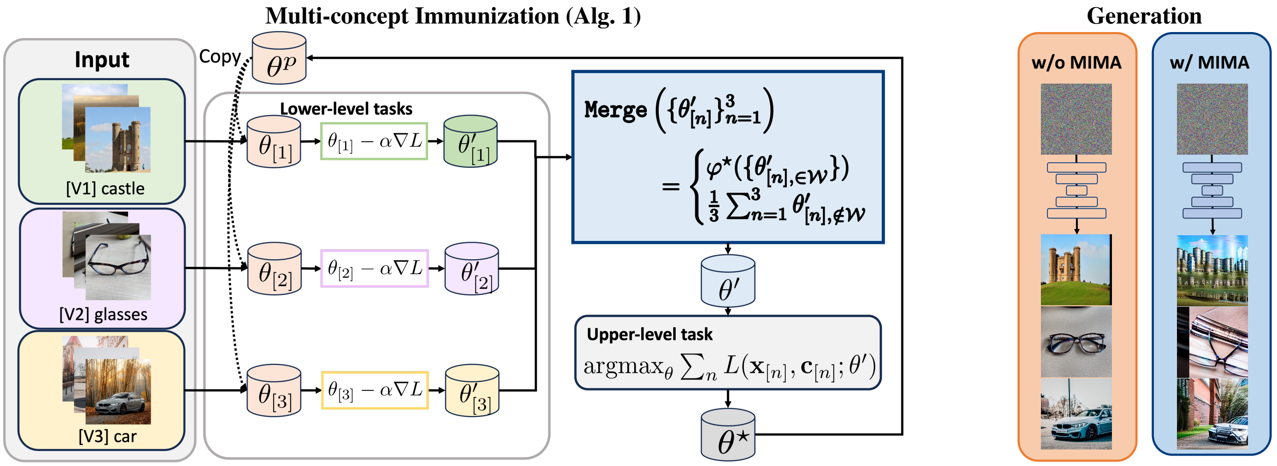
Multi-concept Immunization (Alg. 1)
Generation
Input
Copy
θ^p
Lower-level tasks
[V1] castle
[V2] glasses
[V3] car
$\theta_{[1]}$
$\theta_{[1]} - \alpha\nabla L$
$\theta'_{[1]}$
$\theta_{[2]}$
$\theta_{[2]} - \alpha\nabla L$
$\theta'_{[2]}$
$\theta_{[3]}$
$\theta_{[3]} - \alpha\nabla L$
$\theta'_{[3]}$
Merge $\left(\{\theta'_{[n]}\}_{n=1}^{3}\right) = \begin{cases}\varphi^\star(\{\theta'_{[n],\in\mathcal{W}}\}) \\ \frac{1}{3}\sum_{n=1}^{3}\theta'_{[n],\notin\mathcal{W}}\end{cases}$
θ'
Upper-level task
$\text{argmax}_\theta \sum_n L(\mathbf{x}_{[n]}, \mathbf{c}_{[n]}; \theta')$
$\theta^\star$
w/o MIMA
w/ MIMA

Figure 2: Method overview. *Left:* MIMA is formulated as a bi-level optimization program. For the lower-level, we unroll loss L for the copied weights of each concept. Next, we combine the individual weights $\theta'_{[n]}$ via our proposed Merge layer defined in Eq. (5). For the upper-level, we maximize the diffusion loss L with respect to the parameters θ by backpropagating through θ'. *Right:* During generation, a model $\theta^\star$ immunized with MIMA fails to be adapted by $\mathcal{A}$ on all of the target concepts, *i.e.*, the generations do not contain good quality images of castles, glasses, or cars.

```
Algorithm 1: MIMA (Our method)
Input: pre-trained model $\theta^p$, images $\mathcal{X} = \cup_n \mathcal{X}_{[n]}$ concepts
    $\mathcal{C} = \{\mathbf{c}_{[n]}\}_{n=1}^{N}$, learning rates $\alpha$ and $\beta$, modified parameters set $\mathcal{S}^l$ and $\mathcal{S}^u$ in lower and upper tasks, loss function
    $L$, Merge layer, training epochs $K$
Output: Immunized model $\theta^\star$
Initialize $\theta^0 = \theta^p$
for $k = 1$ to $K$ do
    Sample batches of each concept $\{(\mathbf{x}^u_{[n]}, \mathbf{c}_{[n]})\}_{n=1}^{N}$
    from $\mathcal{X}$ and $\mathcal{C}$
    # Solve the lower-level tasks for one step.
    for $n = 1$ to $N$ do
        Sample batch $\mathbf{x}^l_{[n]}$ from $\mathcal{X}_{[n]}$
        $\theta'_{[n],\in\mathcal{S}^l} \leftarrow \theta^{k-1}_{\in\mathcal{S}^l} - \alpha\nabla_\theta L(\mathbf{x}^l_{[n]}, \mathbf{c}_{[n]}; \theta^{k-1})$
    end for
    # Each $\theta'_{[n]}$ is a function of $\theta^{k-1}$
    $\theta' \leftarrow$ Merge$(\{\theta'_{[1]}, \dots, \theta'_{[N]}\})$
    $\theta^k_{\in\mathcal{S}^u} \leftarrow \theta^{k-1}_{\in\mathcal{S}^u} + \beta\nabla_\theta L(\mathbf{x}^u_{[n]}, \mathbf{c}_{[n]}; \theta')$
end for
$\theta^\star \leftarrow \theta^K$
return $\theta^\star$
```

from chain-rule. We note that this gradient has been previously studied in the optimization layer literature (Amos and Kolter 2017; Barratt and Boyd 2021) in more generic forms.

Putting everything together, from chain rule, the gradient of L w.r.t. to θ is:

$$\frac{\partial L(\theta')}{\partial\theta} = \frac{\partial L}{\partial\theta'} \cdot \sum_{n=1}^{N} \frac{\partial\theta'}{\partial\theta'_{[n]}} \cdot \frac{\partial\theta'_{[n]}}{\partial\theta}, \tag{9}$$

where $\frac{\partial\theta'}{\partial\theta'_{[n],\in\mathcal{W}}}$ is computed through $\varphi^\star$ and $\frac{\partial\theta'}{\partial\theta'_{[n],\notin\mathcal{W}}}$ is a scaled identity matrix.

Solving bi-level optimization. We solve the bi-level optimization program in Eq. (3) using gradient-based methods (Maclaurin, Duvenaud, and Adams 2015; Shaban et al. 2019) commonly used in meta-learning (Finn, Abbeel, and Levine 2017). We provide a summary in Alg. 1. The lower-level tasks in Eq. (3) are solved approximately per $\theta'_{[n]}$ with a single step of gradient update. After collecting all $\{\theta'_{[n]} \forall n\}$, we aggregate these weights with the proposed Merge layer, which leads to an aggregated parameter θ' as a function of the original θ. Next, we iteratively solve the upper-level task using gradient descent by backpropagating through θ' to update θ, *i.e.*, unrolled gradient.

5 Experiments

As in IMMA (Zheng and Yeh 2024), we consider two categories of malicious adaptation: ❶ immunization for protecting against re-learning concepts from an erased model and ❷ immunization against personalized content. Different from IMMA, we generalize immunization experiment settings to multiple concepts.

Baselines. In our experiments, we compare MIMA against two baselines extended from IMMA:

- Joint (JT) performs multi-concept immunization by joint training all concepts by combining the training datasets into one. For the $\mathcal{A}$, we choose DreamBooth as the inner loop adaptation algorithm, *i.e.*, modifying the whole U-Net, which gives the best immunization performance among the different adaptation methods used in IMMA.
- Compose (CP) only aggregates the cross-attention key and value weights via Eq. (3) and freezes all the other weights of θ during immunization training. This is equivalent to IMMA choosing Custom Diffusion (Kumari et al. 2023b) as the adaptation algorithm $\mathcal{A}$ during immunization. As

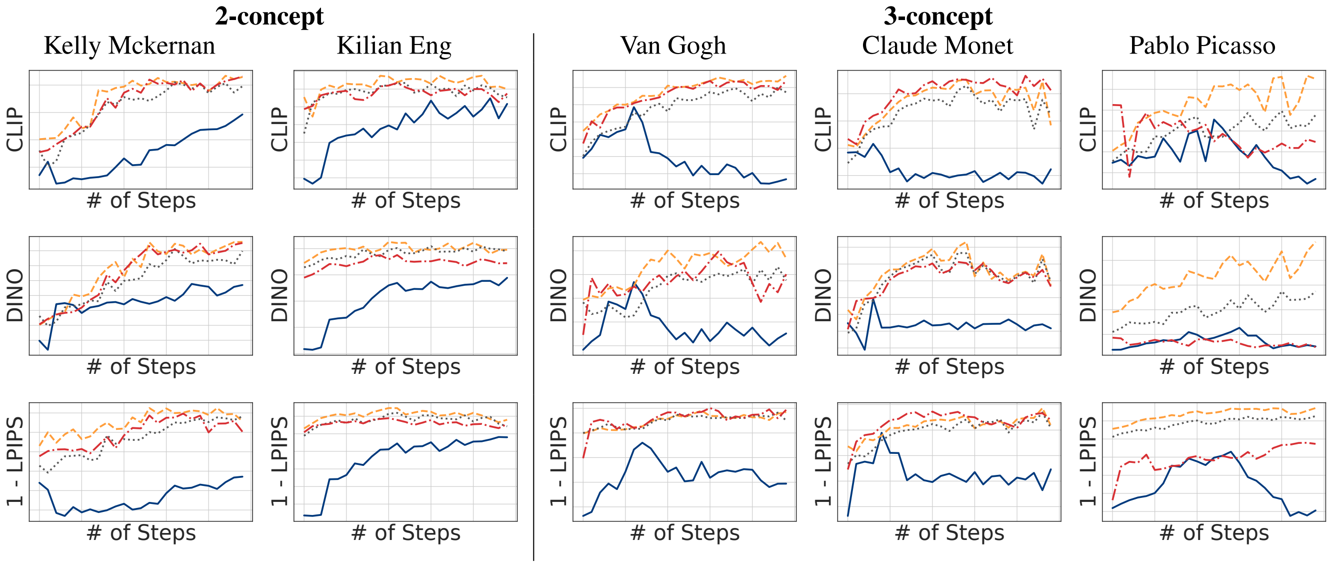

Figure 3: Similarity *vs*. epochs for LoRA on styles. Each row shows one metric. Models with MIMA achieve lower similarity throughout LoRA's steps. This means that on the target concepts, MIMA generates images less similar to the references.

Custom Diffusion supports customization on multiple concepts, `CP` is a natural generalization of IMMA to multi-concept immunization.

5.1 Multi-concept Re-learning Immunization

Following IMMA (Zheng and Yeh 2024), we perform experiments on eight artistic styles and ten classes spanning various categories from a subset of ImageNet (Deng et al. 2009).

Experiment details. We choose the concept sets by randomly sampling two or three concepts from the eight styles or the ten classes. The pre-trained weights are from UCE (Gandikota et al. 2024), an algorithm that erases multiple concepts from a pre-trained Diffusion model. For immunization, we generate 20 images for each target concept from Stable Diffusion V1-4 (SD V1-4) with the prompts of the target artistic styles and objects. Specpficially, the prompts are *"an artwork of* {*artist name*}*"* and *"a photo of* {*object name*}*"* rescptively.

As in IMMA, we consider the risk of re-learning the concept using the efficient adaptation method of LoRA (Hu et al. 2022). We generate another 20 images to be used as the training images for LoRA. To maintain the model capability of being finetuned to learn other concepts, we generate 200 regularization images for `Merge` using either the prompt *"artwork"* or *"object"* for each of the corresponding settings. The results for re-learning style are presented in this section, and the results for objects can be found in the appendix.

Evaluation metrics. IMMA explained that the effectiveness of an immunization can be evaluated by quantifying the performance *gap* with and without the immunization. Following this intuition, we propose *Mean Similarity Gap Ratio* (`MSGR`) between the generation with and without MIMA for *all target concepts* from $\mathcal{C}$ as an evaluation metric.

Group #		2-concept					3-concept				
		1	2	3	4	5	1	2	3	4	5
	`(C)`	1.26	1.80	0.81	3.13	2.16	0.55	3.45	-1.32	3.11	5.38
`JT`	`(D)`	9.87	10.6	6.00	21.0	1.88	6.34	29.2	0.61	15.8	12.3
	`(L)`	2.82	0.31	4.23	4.65	2.80	1.51	5.78	-1.17	2.18	5.90
	`(C)`	5.31	1.15	0.82	**7.38**	**7.45**	6.13	7.44	6.43	3.28	7.10
`CP`	`(D)`	24.8	-5.30	7.40	31.1	19.5	19.1	26.1	31.3	29.5	**26.3**
	`(L)`	24.9	3.02	6.96	8.32	8.37	13.3	17.4	32.3	9.73	16.0
	`(C)`	**6.66**	**12.2**	**6.84**	3.89	7.26	**6.92**	**13.2**	**11.5**	**14.8**	7.67
Ours	`(D)`	**25.3**	**22.0**	**27.6**	**42.4**	**20.0**	**50.4**	**46.9**	**39.4**	**46.6**	**19.5**
	`(L)`	**41.4**	**12.3**	**22.2**	**11.7**	**28.7**	**38.9**	**44.8**	**38.1**	**42.6**	**19.6**

Table 1: `MSGR` ↑(%) on artistic styles for UCE with LoRA. MIMA shows an average `MSGR` improvement of 18.95% over `JT` and 10.94% over `CP` across all three similarity metrics.

Given a metric $\mathcal{M}$ that captures image similarity, $\texttt{MSGR}\left(\{\mathbf{x}^{\mathcal{I}}_{[n]}\}, \{\mathbf{x}^{\mathcal{A}}_{[n]}\}, \{\mathbf{x}^{r}_{[n]}\}\right)$ is defined as

$$\frac{1}{|\mathcal{C}|}\sum_{n=1}^{|\mathcal{C}|}\frac{\overbrace{\mathcal{M}(\mathbf{x}^{r}_{[n]},\mathbf{x}^{\mathcal{A}}_{[n]})}^{\text{w/o immunization}(\uparrow)} - \overbrace{\mathcal{M}(\mathbf{x}^{r}_{[n]},\mathbf{x}^{\mathcal{I}}_{[n]})}^{\text{w/ immunization}(\downarrow)}}{\mathcal{M}(\mathbf{x}^{r}_{[n]},\mathbf{x}^{\mathcal{A}}_{[n]})}. \tag{10}$$

Here, $\mathbf{x}^{\mathcal{I}}_{[n]}$ and $\mathbf{x}^{\mathcal{A}}_{[n]}$ denote the generated images with and without immunization of the n-th target concept, and $\mathbf{x}^{r}_{[n]}$ denotes the corresponding reference images of the target concept. A larger `MSGR` indicates a stronger effect of MIMA as the performance gap is larger.

Following IMMA, we choose $\mathcal{M}$ to be one minus the Learned Perceptual Image Patch Similarity (Zhang et al. 2018) (LPIPS), cosine similarity measured in the feature space of CLIP (Radford et al. 2021) or DINO (Caron et al. 2021) each denoted as `MSGR(L)`, `MSGR(C)` and `MSGR(D)`.

Style results. In Tab. 1, we report the `MSGR` of re-learning sets of concepts after they were erased by UCE (Gandikota et al. 2024). We provide results on five groups of concepts for

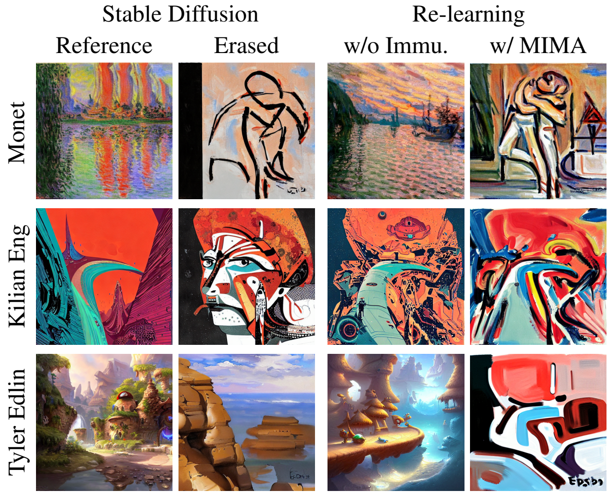

Figure 4: Qualitative result of MIMA against re-learning artistic styles. Both Erased and MIMA are adapted to all three concepts on a single model.

Group #		2-concept					3-concept				
		1	2	3	4	5	1	2	3	4	5
JT	(C)	3.34	3.77	3.30	2.08	5.38	2.77	3.72	2.23	6.50	3.97
	(D)	11.3	16.5	14.3	7.94	17.1	14.1	10.3	14.1	20.2	13.6
CP	(C)	1.26	0.21	2.34	1.88	5.04	2.21	1.25	1.58	2.07	0.77
	(D)	3.08	5.11	9.28	10.7	22.7	12.0	14.6	**21.6**	5.45	1.22
Ours	(C)	**6.57**	**5.49**	**6.23**	**5.57**	**7.48**	**4.97**	**3.75**	**3.35**	**6.64**	**6.43**
	(D)	**14.7**	**18.1**	**17.5**	**28.5**	**24.8**	**17.1**	**19.4**	13.6	**22.7**	**18.7**

Table 2: MSGR ↑(%) on personalized adaptation. MIMA shows an average MSGR improvement of 3.75% over JT and 6.36% over CP across all groups.

each of the two or three concept combinations, *e.g.*, group 1 of two concepts corresponds to the artistic style of Monet, and Fagan. All the numbers are reported at the 400^{th} step of LoRA with a batch size of four.

We observe that the MSGR of MIMA is generally greater than zero. A positive gap between the similarity without and with MIMA indicates the effectiveness of immunization. Overall, we observe that MIMA outperforms JT by 18.95% and CP by 10.94% averaging across all groups and metrics.

To further study the effectiveness of MIMA, we visualize the CLIP, DINO, and LPIPS metrics at each training step for LoRA in Fig. 3. The gaps between the lines and the dashed orange line illustrate the MSGR. A larger gap means that the immunization method performs better. We observe that MIMA outperforms to two compared baselines.

In Fig. 4, we provide qualitative results and observe the following: (a) LoRA can train back the erased concepts on a model without immunization; (b) With the immunization of MIMA, a shows a degree of resistance to LoRA, *i.e.*, the model fails to generate artwork in the style of the multiple protected artists. These observations are consistent with our quantitative findings.

Group #		2-concept					3-concept				
		1	2	3	4	5	1	2	3	4	5
JT	(C)	1.82	4.26	1.28	0.08	4.96	2.52	3.14	2.23	3.84	1.73
	(D)	14.2	16.6	7.02	0.01	8.43	7.33	-10.9	8.73	4.81	6.48
CP	(C)	1.10	3.52	-0.03	-0.67	4.05	0.49	-0.58	1.58	1.83	-0.55
	(D)	7.46	9.83	6.88	-7.95	2.18	-5.93	-5.76	-0.30	2.36	0.77
Ours	(C)	**5.08**	**5.60**	**3.16**	**4.38**	**5.11**	**5.79**	**7.10**	**4.36**	**4.03**	**1.92**
	(D)	**22.4**	**22.9**	**15.0**	**15.7**	**9.36**	**20.7**	**14.8**	**15.7**	**16.2**	**7.20**

Table 3: MRSGR ↑(%) on personalized adaptation. MIMA shows an average MRSGR improvement of 5.89% over JT and 9.31% over CP across all groups.

5.2 Multi-concept Personalization Immunization

Following IMMA, we evaluate MIMA against learning unique/personalized concepts under four adaptation methods: Textual Inversion (TI) (Gal et al. 2023), DreamBooth (Ruiz et al. 2023) (DB), DreamBooth LoRA, and Custom Diffusion (CD) (Kumari et al. 2023b).

Experiment details. We conduct the experiments on thirteen unique concepts from Kumari et al. (2023b), including pets, furniture, scenes, decor items, *etc*. Each of them contains four to six real-world images of a personalized/unique concept. To form the concept sets, we randomly select two to three concepts among them. For MIMA training, we pair each unique concept with a unique token in the prompt. We train MIMA with personalized images and the prompt containing the unique token. For adaptation, we consider the four aforementioned adaptation methods on top of the same immunized weights to study the effect. The evaluation prompt for all concepts is "A photo of $[V^*]$", with *different* special tokens during MIMA training phrases. The regularization concept is set to each target concept's category name, *e.g.*, *"cat"* or *"plant"*. As in the re-learning task, we generated 200 images for the regularization of MIMA.

Evaluation metrics. Beyond MSGR, we also want to show that the model maintains its capacity of being fine-tuned to generate *other* concepts. Hence, we introduce the Mean Relative Similarity Gap Ratio (MRSGR). This metric measures the performance gap between the target and other concepts for models with and without immunization, where the performance is measured as the average similarity between generations from models with and without MIMA.

Formally, we denote $(\mathbf{x}^{\mathcal{I}}_{[n]}, \mathbf{x}^{\mathcal{A}}_{[n]})$ as the generated images, after adaptation, with and without MIMA for n-th target concept set $\mathcal{C}$. $(\mathbf{x}^{\mathcal{I}}_{o,[n']}, \mathbf{x}^{\mathcal{A}}_{o,[n']})$ are generated images with and without MIMA on n'-th *other unique concept* in the set of other concepts $\mathcal{C}_o$. We define $\texttt{MRSGR}(\{(\mathbf{x}^{\mathcal{I}}_{[n]}, \mathbf{x}^{\mathcal{A}}_{[n]})\}, \{(\mathbf{x}^{\mathcal{I}}_{o,[n']}, \mathbf{x}^{\mathcal{A}}_{o,[n']})\})$ as

$$\frac{\overbrace{\bar{\mathcal{M}}(\{(\mathbf{x}^{\mathcal{I}}_{o,[n']}, \mathbf{x}^{\mathcal{A}}_{o,[n']})\})}^{\text{Other concept}(\uparrow)} - \overbrace{\bar{\mathcal{M}}(\{(\mathbf{x}^{\mathcal{I}}_{[n]}, \mathbf{x}^{\mathcal{A}}_{[n]})\})}^{\text{Target concepts}(\downarrow)}}{\bar{\mathcal{M}}(\{(\mathbf{x}^{\mathcal{I}}_{o,[n']}, \mathbf{x}^{\mathcal{A}}_{o,[n']})\})}, \qquad (11)$$

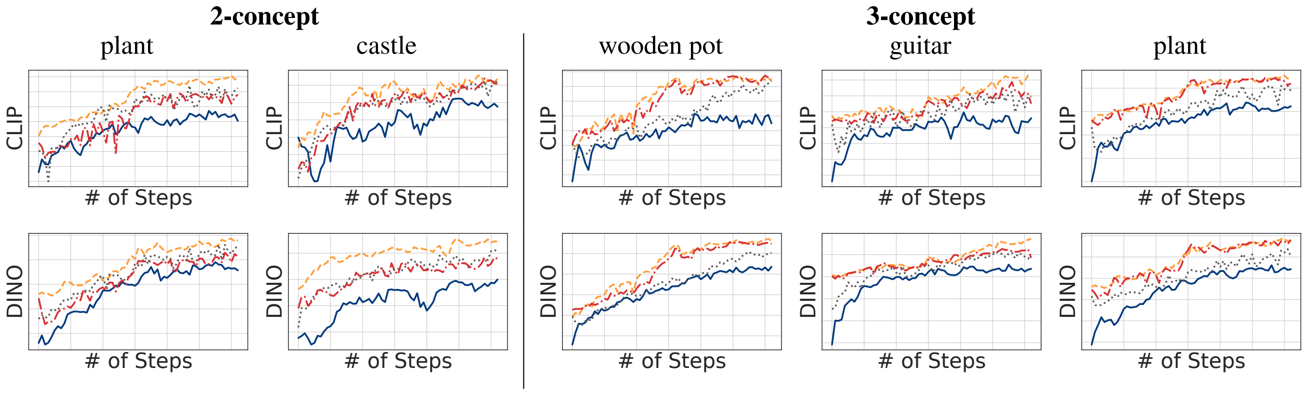

Figure 5: CLIP and DINO similarity on personalization concepts. The *gaps* between the dashed line and solid lines show `MSGR` ↑(%) of different methods. That is, a larger gap indicates stronger immunization.

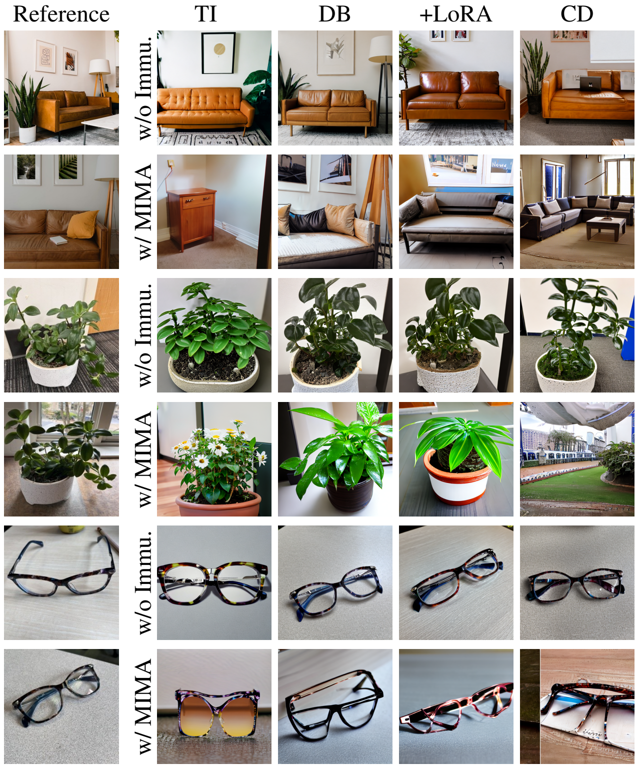

Figure 6: Qualitative results with and without MIMA against three concepts across four personalization methods.

with average similarity $\bar{\mathcal{M}}$ over the image pairs defined as:

$$\bar{\mathcal{M}}(\{(\mathbf{x}^{\mathcal{I}}_{[n]}, \mathbf{x}^{\mathcal{A}}_{[n]})\}) = \frac{1}{|\mathcal{C}|}\sum_{n=1}^{|\mathcal{C}|}\mathcal{M}(\mathbf{x}^{\mathcal{I}}_{[n]}, \mathbf{x}^{\mathcal{A}}_{[n]}) \text{ and } \tag{12}$$

$$\bar{\mathcal{M}}(\{(\mathbf{x}^{\mathcal{I}}_{o,[n']}, \mathbf{x}^{\mathcal{A}}_{o,[n']})\}) = \frac{1}{|\mathcal{C}_o|}\sum_{n'=1}^{|\mathcal{C}_o|}\mathcal{M}(\mathbf{x}^{\mathcal{I}}_{o,[n']}, \mathbf{x}^{\mathcal{A}}_{o,[n']}). \tag{13}$$

A larger `MRSGR` indicates a better effect at preserving the other concepts when immunizing the model against the target concepts. To show one single immunized model is effective against multiple adaptation methods, we report the averaged `MSGR` and `MRSGR` over the four adaptation methods.

Personalization results. In Tab. 2, we report `MSGR` of immunization against personalization adaptation on five 2-concept and 3-concept sets. We observe positive ratios across all sets and all evaluation metrics. Overall, MIMA has the largest ratios across different sets and evaluation metrics, which indicates that MIMA most effectively protects the pre-trained model. All the results in the tables are reported at the $40^{\texttt{th}}$ step for all adaptations.

To show that MIMA maintains the capacity to learn other personalized concepts with the adaptation methods, we report `MRSGR` in Tab. 3. We denote concepts other than the target concepts as "other concepts" within the thirteen personalization concepts. As we can see, `MRSGR` of MIMA is the largest across all concept sets and metrics, which means MIMA is better at preserving the model's ability to personalize other concepts. Additionally, we provide the `MSGR` metric against adaptation training steps in Fig. 5. We can observe a solid gap between with and without MIMA. The gap is larger than that of `JT` and `CP`, which shows that MIMA is more effective than the baselines at immunizing the model.

Finally, we show the generated images with and without MIMA's adaptation in Fig. 6. Compared with the reference images in the first column, models with MIMA do not generate the exact personal item or generate an unrelated image. In other words, MIMA protects the model from being adapted to personal concepts.

6 Conclusion

In this work, we aim to mitigate the risk associated with the open-sourcing of text-to-image models by studying the mitigation method based on the model immunization paradigm of IMMA (Zheng and Yeh 2024). We generalized the setting by considering multi-concept immunization. We then propose MIMA, a multi-concept immunization algorithm that makes a pre-trained model difficult to fine-tune on *multiple* harmful concepts. MIMA leverages a differentiable merge layer that combines model weights to achieve multi-concept immunization.

Acknowledgements

This project is supported in part by an NSF Award #2420724 and the Ross-Lynn Research Scholar Grant.

References

Agrawal, A.; Amos, B.; Barratt, S.; Boyd, S.; Diamond, S.; and Kolter, J. Z. 2019. Differentiable Convex Optimization Layers. In *Proc. NeurIPS*.

Amos, B.; Jimenez, I.; Sacks, J.; Boots, B.; and Kolter, J. Z. 2018. Differentiable MPC for End-to-end Planning and Control. In *Proc. NeurIPS*.

Amos, B.; and Kolter, J. Z. 2017. OptNet: Differentiable optimization as a layer in neural networks. In *Proc. ICML*.

Bai, S.; Kolter, J. Z.; and Koltun, V. 2019. Deep Equilibrium Models. In *Proc. NeurIPS*.

Bai, S.; Koltun, V.; and Kolter, J. Z. 2020. Multiscale Deep Equilibrium Models. In *Proc. NeurIPS*.

Barratt, S. T.; and Boyd, S. P. 2021. Least squares auto-tuning. *Engineering Optimization*.

Bau, D.; Liu, S.; Wang, T.; Zhu, J.-Y.; and Torralba, A. 2020. Rewriting a deep generative model. In *Proc. ECCV*.

Biggio, B.; Nelson, B.; and Laskov, P. 2011. Support vector machines under adversarial label noise. In *Proc. ACML*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proc. CVPR*.

DeepFloyd Lab at StabilityAI. 2023. DeepFloyd IF. https://github.com/deep-floyd/IF.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*.

Domke, J. 2012. Generic methods for optimization-based modeling. In *Proc. AISTATS*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. ICML*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *Proc. ICLR*.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. StyleGan-NADA: Clip-guided domain adaptation of image generators. *ACM TOG*.

Gandikota, R.; Materzyńska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing Concepts from Diffusion Models. In *Proc. ICCV*.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proc. WACV*.

Geng, Z.; Pokle, A.; and Kolter, J. Z. 2023. One-Step Diffusion Distillation via Deep Equilibrium Models. In *Proc. NeurIPS*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *Proc. ICLR*.

Gould, S.; Hartley, R.; and Campbell, D. J. 2021. Deep declarative networks. *IEEE TPAMI*.

Harwell, D. 2023. AI-generated child sex images spawn new nightmare for the web. *The Washington Post*.

Heng, A.; and Soh, H. 2023. Selective Amnesia: A Continual Learning Approach to Forgetting in Deep Generative Models. In *Proc. NeurIPS*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. ICLR*.

Hu, Y.-T.; Schwing, A.; and Yeh, R. A. 2023. Surface Snapping Optimization Layer for Single Image Object Shape Reconstruction. In *Proc. ICML*.

Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023a. Ablating Concepts in Text-to-Image Diffusion Models. In *Proc. ICCV*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023b. Multi-Concept Customization of Text-to-Image Diffusion. In *Proc. CVPR*.

Liang, C.; and Wu, X. 2023. Mist: Towards Improved Adversarial Examples for Diffusion Models. arXiv:2305.12683.

Liang, C.; Wu, X.; Hua, Y.; Zhang, J.; Xue, Y.; Song, T.; Xue, Z.; Ma, R.; and Guan, H. 2023. Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples. In *Proc. ICML*.

Liu, Z.; Liu, L.; Wang, X.; and Zhao, P. 2023. Differentiable Frank-Wolfe Optimization Layer. *arXiv preprint arXiv:2308.10806*.

Maclaurin, D.; Duvenaud, D.; and Adams, R. 2015. Gradient-based hyperparameter optimization through reversible learning. In *Proc. ICML*.

Mei, S.; and Zhu, X. 2015. Using machine teaching to identify optimal training-set attacks on machine learners. In *Proc. AAAI*.

Nitzan, Y.; Gharbi, M.; Zhang, R.; Park, T.; Zhu, J.-Y.; Cohen-Or, D.; and Shechtman, E. 2023. Domain expansion of image generators. In *Proc. CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*.

Ren, Z.; Yeh, R. A.; and Schwing, A. 2020. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. In *Proc. NeurIPS*.

Rolínek, M.; Swoboda, P.; Zietlow, D.; Paulus, A.; Musil, V.; and Martius, G. 2020. Deep graph matching via blackbox differentiation of combinatorial solvers. In *Proc. ECCV*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*.

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 *Proc. CVPR*.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proc. CVPR*.

Shaban, A.; Cheng, C.-A.; Hatch, N.; and Boots, B. 2019. Truncated back-propagation for bilevel optimization. In *Proc. AISTATS*.

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 *USENIX Security Symposium*.

Tschiatschek, S.; Sahin, A.; and Krause, A. 2018. Differentiable Submodular Maximization. In *IJCAI*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. NeurIPS*.

Wang, P.-W.; Donti, P.; Wilder, B.; and Kolter, J. Z. 2019. SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *Proc. ICML*.

Wang, S.; Teng, Y.; and Wang, L. 2023. Deep equilibrium object detection. In *Proc. CVPR*.

Yeh, R. A.; Hu, Y.-T.; Ren, Z.; and Schwing, A. G. 2022. Total Variation Optimization Layers for Computer Vision. In *Proc. CVPR*.

Zhang, E.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2024. Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. In *Proc. CVPR Workshop*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proc. ICCV*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*.

Zhao, Z.; Duan, J.; Hu, X.; Xu, K.; Wang, C.; Zhang, R.; Du, Z.; Guo, Q.; and Chen, Y. 2024. Unlearnable Examples for Diffusion Models: Protect Data from Unauthorized Exploitation. In *Proc. ICLR Workshop*.

Zheng, A. Y.; He, T.; Qiu, Y.; Wang, M.; and Wipf, D. 2024. Graph Machine Learning through the Lens of Bilevel Optimization. In *Proc. AISTATS*, volume 238.

Zheng, A. Y.; Yang, C.-A.; and Yeh, R. A. 2024. Learning to obstruct few-shot image classification over restricted classes. In *Proc. ECCV*.

Zheng, A. Y.; and Yeh, R. A. 2024. Imma: Immunizing text-to-image models against malicious adaptation. In *Proc. ECCV*.