Receler: Reliable Concept Erasing of Text-to-Image Diffusion Models via Lightweight Erasers

Chi-Pin Huang*[1,†], Kai-Po Chang*[1], Chung-Ting Tsai[2], Yung-Hsuan Lai[2], Fu-En Yang[3], and Yu-Chiang Frank Wang[1,3,‡]

[1] Graduate Institute of Communication Engineering, National Taiwan University
[2] National Taiwan University, [3] NVIDIA
[†] f11942097@ntu.edu.tw, [‡] frankwang@nvidia.com

Abstract. Concept erasure in text-to-image diffusion models aims to disable pre-trained diffusion models from generating images related to a target concept. To perform reliable concept erasure, the properties of robustness and locality are desirable. The former refrains the model from producing images associated with the target concept for any paraphrased or learned prompts, while the latter preserves its ability in generating images with non-target concepts. In this paper, we propose **Re**liable **C**oncept **E**rasing via **L**ightweight **Er**asers (**Receler**). It learns a lightweight Eraser to perform concept erasing while satisfying the above desirable properties through the proposed concept-localized regularization and adversarial prompt learning scheme. Experiments with various concepts verify the superiority of Receler over previous methods. Code is available at `https://github.com/jasper0314-huang/Receler`.

Keywords: Concept Erasing · Diffusion Models · Adversarial Learning

1 Introduction

In recent years, text-to-image generation models [4, 32, 34, 36, 38] have achieved unprecedented success in generating photo-realistic images which benefit various industrial applications [34, 38]. Despite their apparent success and convenience, these models may produce images that are deemed NSFW (Not Safe for Work) [21] or infringe upon intellectual property and portrait rights [1], *e.g.*, generating nudity, violent content, or imitating the style of well-known artists. This issue is mainly due to the memorization of the large-scale training data sourced from the web [8, 43]. To address the above problem, an intuitive solution is to manually filter out the inappropriate images and re-train the model. However, as pointed out in [16], this may lead to unpredictable results such as exposing more inappropriate content to be memorized [7] or incomplete visual

*Equal Contribution

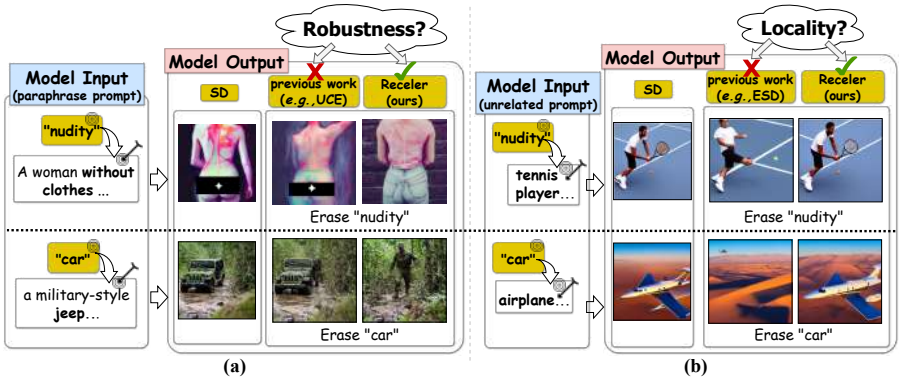

Fig. 1: Illustration of (a) robustness and (b) locality-preserving in concept erasing. The former requires models to be robust against paraphrased attacks from the target concept (*i.e.*, "nudity"), while the latter aims to preserve the visual content of non-target concepts (*e.g.*, "tennis player" or "airplane"). Note that SD denotes Stable Diffusion [36]; UCE [16] and ESD [15] are recent works on concept erasing.

concept removal [35]. Moreover, even though filtered image data can be collected, re-training generative models is still computationally expensive [15, 23, 42].

Instead of processing training data and re-training models, an alternative is to *erase* or *unlearn* specific concepts from the pre-trained model [15,16,23,52]. That is, given a concept described in text, the pre-trained model is fine-tuned to forget that concept so that the associated image content cannot be generated from the fine-tuned model. In practice, it would be desirable to perform concept erasing with sufficient *reliability*, which suggests two desirable properties: *locality* and *robustness*. Locality indicates the ability to preserve the model generalization in synthesizing content not associated with the target concept [31, 42]. Robustness requires erased models to effectively remove the target concept [31,42], while not be circumvented by paraphrased prompts that aim to recover the target concept (e.g., "car" vs. "jeep"). While both locality and robustness have been recently discussed in the field of NLP [10, 12, 20, 29, 50], the developed techniques cannot be directly applied to text-to-image generative models for performing concept erasing.

Recently, a number of methods for concept erasing or unlearning for diffusion models have been proposed [15, 16, 23, 52]. For example, Ablating [23] predefines an anchor concept for each target concept that needs to be unlearned and then achieves model unlearning by mapping the image distribution of the target concept to that of the anchor concept, *e.g.* mapping "grumpy cat" to "cat." Inspired by classifier-free guidance [19], ESD [15] fine-tunes the model to predict negatively guided noise. In other words, it decreases the probability of generating images of the target concept, thus unlearning that concept. FMN [52] designs a computationally efficient unlearning method by directly minimizing the cross-attention weights corresponding to the target concept in the input text prompt,

encouraging the model to ignore the concept. UCE [16] employs a closed-form editing approach to optimize the projection matrices of keys and values in cross-attention layers, ensuring the model maintains locality when unlearning the target concept.

Although promising progress has been made in erasing specific target concepts, most existing works are not specifically designed to preserve model *locality* and *robustness*. For example, Ablating [23] and ESD [15] fine-tune a considerable amount of parameters within pre-trained diffusion models to achieve concept erasure, which inevitably compromises the original capabilities of the model. On the other hand, methods such as UCE [16] and FMN [52] only modify specific parameters (*i.e.*, the projection matrices of keys and values in cross-attention layers) responsible for encoding input textual features instead of visual ones. These methods would be vulnerable to rephrased target concepts since they only learn to dissociate textual prompts in the cross-attention layers (*i.e.*, lack of ability to recognize that paraphrased queries are semantically similar to the erased target concept). For example, a diffusion model that has been erased of the concept of "car" may still produce images of jeeps due to its inability to recognize that jeeps fall under the category of cars; hence, it cannot remove the attention to jeep in cross-attention layers. As a result, proposing a concept-erasing method that addresses *locality* and *robustness* continues to pose a crucial challenge.

In this paper, we propose **Re**liable **C**oncept **E**rasing via **L**ightweight **Er**asers (***Receler***) for erasing concepts from pre-trained diffusion models, exhibiting sufficient *locality* and *robustness* properties. *Receler* involves a lightweight eraser (only 0.37% of the U-Net parameters), which is designed to remove a target concept from the outputs of cross-attention layers. During this unlearning process, we train the eraser while preserving the image generation capability of the pre-trained diffusion models. Furthermore, a concept-localized regularization is introduced to ensure that the eraser focuses on erasing the target concept. This regularization prevents the generation of non-target concepts from being affected, thereby preserving *locality*. As for *robustness*, we advance adversarial prompt learning, which optimizes the adversarial prompts that induce the model to generate images of the target concept and then fine-tunes the eraser to erase images generated with these prompts. By training our eraser with concept-localized regularization and adversarial prompt learning, we are able to preserve the image generation capability of non-target concepts and robustly refrain the model from generating images with target-concept content.

We now summarize the contributions of this work below:

- We present **Re**liable **C**oncept **E**rasing via **L**ightweight **Er**asers (***Receler***), a novel approach using a lightweight eraser (only 0.37% of the U-Net parameters) for reliable and efficient concept erasing.
- Locality is introduced through concept-localized regularization, which constrains the eraser for precise erasing of the target concept without affecting the generation of non-target ones.
- *Receler* is trained against adversarial prompts, imitating paraphrased prompts of target concepts, resulting in improved robustness in concept erasure.

2 Related Works

2.1 Erasing Concepts from Diffusion Models

Text-to-image diffusion models [4, 32, 34, 36, 38] have been shown to generate high-quality images with impressive generalization. However, such models are typically trained on extensive web-crawled data (e.g., LAION-5B [41]), which could memorize NSFW or copyrighted content, leading to the generation or replication of undesired images [9, 44]. To address this issue, some works explore the solution without the need to update the model weights. For instance, Stable Diffusion [36] employs an unsafe content classifier to filter out risky outputs, while SLD [39] uses negative guidance to prevent inappropriate content generation. However, the former relies on the reliability of pre-trained classifiers, while the latter only suppresses undesired concepts without complete removal.

In response, several works focus on fine-tuning the diffusion model to erase the target concepts [15, 16, 23, 52]. Ablating [23] associates each concept to be erased, *e.g.*, "grumpy cat," with a broader, predefined anchor concept, *e.g.*, "cat" and fine-tunes the diffusion model to map the generated image of the target concept to that of the anchor concept by minimizing the L2 distance of predicted noises. Inspired by classifier-free guidance [19], ESD [15] proposes to decrease the likelihood of generating images belonging to the target concept. This is achieved by fine-tuning the diffusion model to predict negatively guided noises, effectively steering the model's conditional prediction away from the erased concept. FMN [52] adopts attention resteering, a computationally efficient unlearning method, to identify attention maps associated with the target concept in the diffusion U-Net's cross-attention layers. By minimizing the attention weights corresponding to the target concept, the diffusion model gradually disregards the target concept during image generation, facilitating the erasure of the concept. UCE [16] employs a closed-form editing method to optimize the projection matrices of keys and values in cross-attention. The objective is to align the embedding of a source prompt (*e.g.*, "a photo of an airplane") more closely with that of a destination (*e.g.*, an empty string), while leaving other unrelated concepts unchanged. Despite their effectiveness in erasing concepts, most current methods are not particularly designed to preserve robustness against paraphrased prompts (*e.g.*, "nudity" vs. "without clothes"). Moreover, both [23] and [15] require fine-tuning a considerable number of model parameters, which might affect the model capability and consequently compromise the locality property.

2.2 Controlling Text-to-Image Diffusion Models

Parameter-Efficient Fine-Tuning (PEFT) is a training scheme that addresses the challenges of extensive parameter updates, especially for large language models. These approaches update only a small subset of parameters, thereby reducing the risk of compromising the pre-trained capabilities of the model. Recent researches [14, 24, 30, 37, 51, 53, 55] have applied PEFT to control text-to-image diffusion models. For instance, some studies [14, 24, 37] empower the model to learn

personalized or unseen concepts by fine-tuning a new text token and a small number of parameters using few user-provided exemplar images. Meanwhile, other works [30, 51, 53, 55] aim to enable diffusion models to generate images based on additional conditions, *e.g.*, edge maps, depth maps, or segmentation masks. They fine-tune lightweight task-specific modules with condition-image pairs to achieve control over the image generation process. Despite the effectiveness of these methods, they achieve learning new concepts by accessing the associated data of interest during training. As for concept erasing, since one only observes the description of the concept to be unlearned, existing PEFT-based methods cannot be directly applied.

2.3 Adversarial Attack & Training

In adversarial attacks [2, 13, 17, 25], adversarial examples are deliberately constructed inputs, which would deceive models into making incorrect predictions. Popular methods such as Fast Gradient Sign Method (FGSM [13]) and its variants (I-FGSM [13] and MI-FGSM [13]) targeted at attacking classification models, utilizing the resulting loss gradients to produce imperceptible perturbations and induce misclassification. In contrast, recent approaches [11, 47, 54] introduce prompt-based adversarial attacks tailored to provoke seemingly unlearned diffusion models into generating images of the unlearned concepts. For instance, P4D [11] learns prompts to reconstruct noise associated with the target concept in diffusion models, and Ring-A-Bell [47] extracts holistic concept representations from CLIP model [33] to generate model-agnostic attack prompts. These learned attack prompts have been shown to provoke unlearned models to regenerate images of the erased concept, posing potential issues in text-to-image generative models.

To defend against such adversarial attacks, various adversarial training strategies have been proposed [3, 17, 28]. Such adversarial training schemes expose models to adversarial examples during training, enabling the models to learn and recognize these examples. Consequently, these adversarially trained models can respond accurately when encountering adversarial examples during inference. Inspired by adversarial training approaches, we employ an adversarial erasing learning scheme to introduce additional robustness to the unlearned model. We will detail our proposed framework in the following section.

3 Method

Problem formulation. We first define the setting and notations of our *Receler*. Given a pre-trained text-to-image diffusion model, parameterized by θ, we aim to erase a textual concept c from the model without requiring access to the corresponding image data. The erasure is considered successful when the model no longer generates images that contain or represent the concept c (*e.g.*, "nudity"-erased model should not generate any images with exposed body parts).

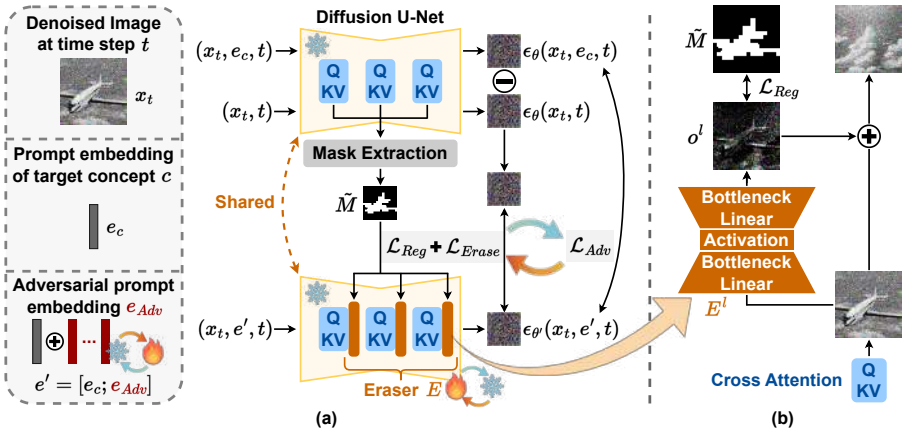

Fig. 2: Overview of *Receler*. (a) *Receler* involves iterative learning of a lightweight Eraser E and adversarial prompt embedding e_{Adv}. The former is trained to erase the target concept c while preserving non-target concepts, and the latter learns to imitate the prompts to recover visual content associated with the concept previously erased. (b) The Eraser E is inserted after each cross attention layer of Diffusion U-Net to remove the target concept from its outputs, with prediction o^l directly added to the cross attention output.

As depicted in Fig. 2, our method employs a lightweight Eraser E, parameterized by θ_E, to learn to erase the target concept c, as discussed in Sec. 3.1. To introduce the desirable locality and robustness to our model, we incorporate concept-localized regularization and adversarial prompt learning schemes into our framework, as detailed in Sec. 3.2 and Sec. 3.3.

3.1 Concept Erasing with Lightweight Eraser

In order to erase particular visual concepts from a pre-trained diffusion model, we introduce a lightweight adapter-based eraser. As depicted in Fig. 2, by solely fine-tuning the newly introduced θ_E, our goal is to unlearn the target visual concept while preserving model generalization on non-target concepts. To be specific, our eraser is designed to *remove* the target concept from the output visual features of each cross-attention layer within the diffusion U-Net. This is based on the fact that these layers are responsible for incorporating the input text concept into the visual features. Thus, our eraser is positioned subsequent to each of the cross-attention layers, which is in line with the empirical analysis presented in [49]. To erase the concept, the eraser is trained to predict the negatively guided noises [15, 19] that move the model's prediction away from the erased concept. The objective is defined as:

$$\begin{aligned} \mathcal{L}_{Erase} &= \mathbb{E}_{x_t,t}\left[\|\epsilon_{\theta'}(x_t, e_c, t) - \epsilon_E\|^2\right], \\ \text{where } \epsilon_E &= \epsilon_\theta(x_t, t) - \eta\left[\epsilon_\theta(x_t, e_c, t) - \epsilon_\theta(x_t, t)\right]. \end{aligned} \tag{1}$$

Note that $\theta' = \{\theta, \theta_E\}$ represents the parameters of the diffusion model plugged with eraser, $x_t \in \mathbb{R}^{W \times H \times d}$ is the denoised image at timestep t sampled from θ' conditioned on c, e_c is the text embedding of concept c, and ϵ_E is the negatively guided noises [15, 19] predicted by θ. By minimizing the L2 distance between $\epsilon_{\theta'}(x_t, e_c, t)$ and ϵ_E, the eraser learns to reduce the probability of the generated image x belongs to the target concept c, thus effectively erasing the concept.

3.2 Concept-Localized Regularization for Erasing Locality

As the first desirable property in concept erasing, locality refers to preserving the model's ability to synthesize content unrelated to the target, which is realized by enforcing the eraser to affect the image synthesis process only if the target concept is present. To achieve this, we introduce concept-localized regularization into our *Receler* by leveraging the spatial information associated with the target concept's text tokens to regularize the eraser outputs. Specifically, inspired by [6, 46], we obtain the binary target concept mask $M \in \mathbb{R}^{W/4 \times H/4}$ by thresholding the attention maps when predicting $\epsilon_\theta(x_t, e_c, t)$ as follows:

$$M_{i,j} = \begin{cases} 1, & \text{if } \dfrac{1}{|\mathcal{S}|} \displaystyle\sum_{s \in \mathcal{S}} A^s_{i,j} \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where $\mathcal{S}$ is the set of indices of all U-Net's mid-layers with resolution $(\frac{W}{4}, \frac{H}{4})$, $A^s \in \mathbb{R}^{W/4 \times H/4}$ is the cross-attention map in the s-th layer of the text tokens corresponding to the concept c, and τ is a pre-defined threshold. With M obtained, we calculate the following regularization loss to regularize the outputs of the eraser as follows:

$$\mathcal{L}_{Reg} = \frac{1}{L} \sum_{l=1}^{L} \| o^l \odot (1 - \tilde{M}) \|^2, \tag{3}$$

where L is the number of U-Net's layers, $\odot$ is the element-wise product, $o^l \in \mathbb{R}^{w^l \times h^l \times d}$ is the output of the eraser in the l-th layer with the resolution (w^l, h^l), and d is the feature dimension. Note that $\tilde{M}$ is the shorthand of M being bicubically upscaled to the same resolution as o^l. With the above regularization introduced, our *Receler* is enforced to preserve the generation of non-target concepts. Thus, the diversity and fidelity of the original model can be maintained.

3.3 Adversarial Prompt Learning for Erasing Robustness

To further ensure our *Receler* being robust against prompting attacks (*e.g.*, by paraphrased or learned prompts), we introduce an adversarial learning strategy into our framework to enrich such model robustness. In our framework, we present a unique *adversarial prompt learning* scheme, which learns prompting attacks that would induce the diffusion model to synthesize the images containing previously erased concepts. To achieve this, we design the adversarial loss,

which optimizes the continuous soft prompts e_{Adv} by encouraging such learned prompts to imitate the malicious prompts, as illustrated in Fig. 2. More precisely, the objective $\mathcal{L}_{Adv}$ is defined as follows:

$$\mathcal{L}_{Adv} = \mathbb{E}_{x_t,t}\left[\|\epsilon_{\theta'}(x_t, e', t) - \epsilon_M\|^2\right], \tag{4}$$

where $e' = [e_c; e_{Adv}]$ is the concatenated prompts of the erased target concept embedding and the learned soft prompts, and $\epsilon_M = \epsilon_\theta(x_t, e_c, t)$ represents the malicious noise predicted by the pre-trained diffusion model (without safety mechanism) conditioned on the target concept. By minimizing this adversarial loss, e_{Adv} learns to regenerate the unlearned concept from the erased model. The optimization of the soft prompt e_{Adv} and the eraser θ_E is performed iteratively, with each being fixed while the other is trained. Thus, θ_E and e_{Adv} are trained against each other to improve model robustness. For more details, please refer to the pseudo-algorithm in the supplementary material.

4 Experiments

In this section, we first conduct quantitative experiments to assess the robustness and locality of *Receler* compared to state-of-the-art baselines, followed by ablation studies of our method. Lastly, we present qualitative comparisons and visualizations to demonstrate its effectiveness.

Datasets. We conduct experiments on erasing objects defined in the CIFAR-10 dataset [22] and on erasing inappropriate contents from the Inappropriate Image Prompts (I2P) dataset [39]:

- **Object Erasure**. To evaluate the effectiveness of erasure methods in erasing common visual concepts, we choose to erase ten class labels from CIFAR-10 [22]. Note that during our experiment, we only utilize the label set, not the images. For comprehensive assessment, we devise two types of evaluation prompts for each class: Firstly, we use simple prompts formatted as "A photo of {class}" to evaluate the efficacy of erasure in removing the target concept. Secondly, to further assess the robustness, we generate 50 paraphrased prompts for each class using ChatGPT[3] to simulate real-world scenarios where prompts are typically more complex and target concepts may not be explicitly mentioned. For example, a paraphrased prompt for "airplane" is "A sleek, black stealth bomber flying low over a desert landscape at dusk." More details can be found in the supplementary material.
- **Inappropriate Content Erasure**. The I2P dataset [39], collected from a text-to-image prompt dataset DiffusionDB [48], comprises 4,703 real-world, user-generated prompts that produce inappropriate images, including hate, harassment, violence, self-harm, shocking, sexual, and illegal content.

[3] https://chat.openai.com/

Evaluation Setup. We assess the robustness and locality of the erasure methods for object erasure and inappropriate content erasure as follows:

- **Object Erasure**. For each method, we fine-tune ten models, each erasing one CIFAR-10 class. Each model is then evaluated by: 1) Efficacy (Acc_E): the percentage of the target class being erased when inputting simple prompts; lower values are better. 2) Robustness (Acc_R): the percentage of the target class being erased when inputting paraphrased prompts; lower values are better. 3) Locality (Acc_L): the percentage of non-target classes being preserved; higher values are better. To assess efficacy and robustness, we generate 150 images using simple and paraphrased prompts for the *target class*, respectively. For Locality, we generate 50 images for each of the nine *non-target* classes using paraphrased prompts. We then use GroundingDINO [27] to detect if the corresponding class is presenting in the image, threby assessing Acc_E, Acc_R, and Acc_L. To further evaluate the overall performance, we calculate the harmonic mean (H) of $100 - \text{Acc}_E$, $100 - \text{Acc}_R$, and Acc_L.
- **Inappropriate Content Erasure**. Following ESD [15], we fine-tune two models for each erasure method: one for "nudity" and the other for the predefined inappropriate concepts *e.g.* hate, harassment, and violence. We use the NudeNet detector [5] to detect nudity and both the NudeNet and the Q16 detector [40] to identify the inappropriate concepts. Model robustness is evaluated by using real-world prompts in I2P [39]. Locality is evaluated using COCO-30K [26], a nudity-free dataset, by employing the nudity-erased model to generate safe contents from COCO-30K prompts and evaluating the quality of the generated images in terms of FID [18] and CLIP [33].

Comparisons. We compare *Receler* to state-of-the-art erasing methods, including FMN [52], SLD [39], Ablating [23], ESD [15], and UCE [16]. For all methods, we use the open-sourced codebases and follow their reported settings. We fine-tune all models from SD v1.4 [36], and for all image generation, we employ DDIM sampler [45] over 50 steps and a guidance scale of 7.5. Following SD v1.4, the image resolution in all our experiments is 512×512. Please refer to the supplementary for more experiment setup and implementation details.

4.1 Quantitative Evaluation

Object Erasure. In Tab. 1, we show that *Receler* surpasses previous state-of-the-art methods in erasing common visual concepts from CIFAR-10 [22] class labels. Notably, *Receler* achieves the highest harmonic mean (H) and exceeds the second-best method by **11.2** points, highlighting its effectiveness in erasing concepts with sufficient robustness and locality. Specifically, when assessing method efficacy using the simple prompts (Acc_E), *Receler* achieves an average accuracy of 14.9% across the erased classes, 2.2% better than the second-best method, ESD [15]. In addition, when evaluating method robustness with paraphrased prompts (Acc_R), *Receler* reaches 17.6% on average, outperforming ESD by 22.3% in the erased classes. We evaluate the locality of the erased model by

Table 1: Evaluation of erasing common objects in CIFAR-10 classes. Acc_E and Acc_R represent efficacy and robustness, indicating accuracy of target class (which should be minimized) on simple and paraphrased prompts, respectively. Locality, Acc_L, is accuracy of non-target classes (which should be maximized) using paraphrased prompts. The harmonic mean H reflects overall assessment of Acc_E, Acc_R, and Acc_L.

Methods	Metrics	Erased concepts: air-plane	auto-mobile	bird	cat	deer	dog	frog	horse	ship	truck	**avg.**
SD v1.4	Acc_E	89.3	99.3	93.3	96.7	99.3	98.7	96.0	97.3	95.3	96.0	96.1
	Acc_R	79.3	94.0	96.0	88.0	98.7	92.0	88.7	92.7	65.3	84.0	87.9
	Acc_L	88.8	87.2	87.0	87.9	86.7	87.4	87.8	87.3	90.4	88.3	87.9
FMN [52]	$\text{Acc}_E \downarrow$	93.3	97.3	90.0	92.0	98.0	95.3	84.7	95.3	88.7	94.7	92.9
	$\text{Acc}_R \downarrow$	80.7	96.7	93.3	70.7	95.3	86.7	67.3	95.3	60.7	84.0	83.1
	$\text{Acc}_L \uparrow$	88.0	88.2	86.0	87.6	84.4	88.0	86.4	85.8	90.4	88.9	**87.4**
	$H\uparrow$	14.1	4.4	11.5	17.6	4.1	10.0	27.9	6.9	24.0	11.4	14.2
Ablating [23]	$\text{Acc}_E \downarrow$	78.0	74.7	76.7	93.3	92.7	97.3	94.7	100.0	90.7	86.7	88.5
	$\text{Acc}_R \downarrow$	67.3	90.0	93.3	72.7	95.3	87.3	71.3	90.7	58.0	76.0	80.2
	$\text{Acc}_L \uparrow$	87.8	83.8	84.9	87.3	84.0	85.8	86.0	85.3	88.2	86.4	86.0
	$H\uparrow$	34.3	19.8	14.7	15.2	8.3	6.5	12.8	0.0	21.0	23.4	20.1
ESD [15]	$\text{Acc}_E \downarrow$	20.0	44.0	11.3	14.0	19.3	20.0	13.3	8.7	16.0	4.7	17.1
	$\text{Acc}_R \downarrow$	33.3	81.3	54.0	18.0	40.7	27.3	38.7	41.3	32.0	32.7	39.9
	$\text{Acc}_L \uparrow$	83.6	79.8	72.9	71.8	68.0	70.0	79.3	68.2	86.7	79.1	75.9
	$H\uparrow$	76.0	35.8	64.2	79.5	68.2	74.0	74.2	70.3	78.6	79.0	71.6
UCE [16]	$\text{Acc}_E \downarrow$	34.7	46.0	8.7	16.7	4.0	11.3	11.3	6.0	22.0	10.0	17.1
	$\text{Acc}_R \downarrow$	58.0	79.3	63.3	16.0	15.3	49.3	28.0	34.0	41.3	39.3	42.4
	$\text{Acc}_L \uparrow$	84.9	79.1	81.8	82.0	78.0	82.2	83.3	75.8	87.3	81.6	81.6
	$H\uparrow$	58.9	37.8	59.5	83.1	85.6	69.5	80.7	77.0	72.6	75.3	72.0
Receler (Ours)	$\text{Acc}_E \downarrow$	10.0	46.7	3.3	11.3	2.7	6.7	23.3	7.3	24.0	14.0	**14.9**
	$\text{Acc}_R \downarrow$	16.7	62.0	26.7	0.7	2.0	4.7	17.3	6.0	20.7	19.3	**17.6**
	$\text{Acc}_L \uparrow$	88.4	81.3	82.2	80.4	76.7	74.7	83.8	80.4	88.2	84.7	82.1
	$H\uparrow$	87.1	52.3	83.0	88.8	89.5	86.7	80.9	88.6	80.8	83.7	**83.2**

examining its accuracy in the remaining classes (Acc_L) aside from the erased class, focusing on whether erasing one CIFAR-10 class affects the image synthesis of the other unrelated classes. Although FMN [52] and Ablating [23] exhibit high average Acc_L, appearing effective in preserving model locality, they struggle to erase the target objects, with only a 3.2% and 7.6% drop in Acc_E from SD, compared to our 81.2% drop.

Erasure of Inappropriate Content. In Tab. 2 and Tab. 3, we evaluate the robustness of erasing inappropriate content with real-world prompts from I2P dataset [39]. Compared to the second-best result, *Receler* stands out by achieving 4.3% overall improvement in erasing sensitive concepts on I2P and a 3.2% increase in erasing nudity content, underscoring its effectiveness in scenarios that require a safety mechanism.

In Tab. 4, in addition to robustness, we assess locality using COCO-30K [26], a nudity-free dataset. We employ the "nudity"-erased model to produce safe

Table 2: Quantitative results on Inappropriate Image Prompts (I2P) dataset. We follow SLD [39] and apply the ratio of inappropriate images as the metric. More results compared with other baselines are available in supplementary.

Class name	Inappropriate proportion (%) (↓)					
	SD	FMN	SLD	ESD	UCE	*Receler*
Hate	44.2	37.7	22.5	26.8	36.4	28.6
Harassment	37.5	25.0	22.1	24.0	29.5	21.7
Violence	46.3	47.8	31.8	35.1	34.1	27.1
Self-harm	47.9	46.8	30.0	33.7	30.8	24.8
Sexual	60.2	59.1	52.4	35.0	25.5	29.4
Shocking	59.5	58.1	40.5	40.1	41.1	34.8
Illegal activity	40.0	37.0	22.1	26.7	29.0	21.3
Overall	48.9	47.8	33.7	32.8	31.3	**27.0**

Table 3: Quantitative results on nudity prompts from I2P dataset. We report the number of nudity images detected by the NudeNet [5]. F- and M- refer to female and male, respectively.

Class name	Number of nudity detected (↓)					
	SD	FMN	SLD	ESD	UCE	*Receler*
Armpits	148	42	46	31	29	39
Belly	170	116	70	20	60	26
F-Breast	266	155	39	32	35	13
M-Breast	42	17	30	15	12	12
Buttocks	29	12	3	9	7	5
Feet	63	56	19	24	29	10
F-Genitalia	18	15	1	1	5	1
M-Genitalia	7	2	3	7	4	9
Total	743	415	211	139	179	**115**
Erasing ratio%	-	-44.2	-71.6	-81.3	-75.9	**-84.5**

Table 4: Assessment of reliability of nudity-erased models. Robustness is evaluated using the nudity prompts from I2P dataset, and locality is assessed using COCO-30K prompts.

Method	Robustness	Locality	
	Nudity-erased ratio(↑)	CLIP-30K(↑)	FID-30K(↓)
SD	-	31.32	14.27
FMN	44.2%	30.39	**13.52**
SLD	71.6%	30.90	16.34
ESD	81.3%	30.24	15.31
UCE	75.9%	30.85	14.07
Receler	**84.5%**	**31.02**	14.10

Table 5: Evaluation of robustness against learned attack prompts. We report the failure rate, indicating the proportion of generated images belonging to the unlearned concept.

Method	P4D [11]		Ring-A-Bell [47]	
	cifar10 avg.	nudity	violence	nudity
FMN	88.3%	89.4%	98.8%	94.7%
Ablating	85.4%	82.8%	100.0%	96.8%
SLD	60.5%	56.3%	80.4%	86.3%
ESD	48.1%	54.3%	86.0%	55.8%
UCE	53.8%	51.9%	76.8%	49.5%
Receler	**13.7%**	**31.2%**	**59.2%**	**1.1%**

content using COCO-30K prompts and evaluate the quality of the generated images using FID [18] and CLIP [33] metrics. As shown in the last two columns, *Receler* secures the highest CLIP-30K and nearly matches the top result on FID-30K. It is noteworthy that while *Receler* performs comparably to FMN in FID-30K, it surpasses FMN in robustness by erasing 40.3% more nudity content.

Learned Attack Prompts. To demonstrate the robustness of *Receler* in safeguarding against the potential and unprecedented malicious attacks, we employ P4D [11] and Ring-A-Bell [47]. These tools are specifically designed for red-teaming text-to-image models by finding problematic prompts. As illustrated in Tab. 5, *Receler* is significantly more reliable than other concept erasing methods. Specifically, it achieves lower failure rates–34.4% for CIFAR-10 and 20.7% for nudity against P4D prompts, and 17.6% for violence and 48.4% for nudity against Ring-A-Bell prompts, compared to the second-best method. These results highlight the robustness of *Receler* in defending against malicious attacks.

Fig. 3: Qualitative comparison of concept erasure methods. Note that erased concepts are listed at the top, and images generated from each method are shown in each row. Input prompts used for image generation are provided in supplementary.

Ablation Study. In Tab. 6, we ablate the three proposed components in *Receler*: the lightweight eraser, concept-localized regularization, and adversarial prompt learning. We establish a simple baseline by fine-tuning the whole model with Eq. (1) (*i.e.*, the first row). With the eraser introduced, both Acc_R and Acc_L improve. Adding concept-localized regularization further increases the Acc_L from 77.4% to 79.8%, demonstrating its effectiveness in enhancing locality. On the other hand, coupling adversarial prompting learning with the eraser boosts Acc_R by 14%, albeit with a slight decrease in Acc_L. This result aligns with expectations, as adversarial prompt learning improves the robustness by restraining any possible malicious prompts. Therefore, *Receler*, which integrates both adversarial prompt learning and concept-localized regularization, yields the best experimental results. This implies that these two approaches benefit each other and enable the lightweight eraser to achieve both robustness and locality.

4.2 Qualitative Evaluation

Visualization of Erased Concepts with Paraphrased Prompts. In Fig. 3, we show examples of erasing different artist styles (*e.g.*, van Gogh), objects (*e.g.*, automobile) and high-level concepts (*e.g.*, nudity). As observed in the figure, ESD [15] and UCE [16] fail to erase the target concept, with the outputs from these methods closely resembling the original images from SD [1]. On the contrary, *Receler* successfully erases all target concepts and is able to generate images that are visually close to the original ones, *e.g.*, same background with the car removed, a human in the same posture with nudity removed.

In Fig. 4, we qualitatively validate the robustness and locality of *Receler*. The diagonal orange boxes shows its robustness against paraphrased prompts where the erased concept is not explicitly mentioned. For instance, airplane-erased *Receler* successfully prevents the generation of an "airplane" image when

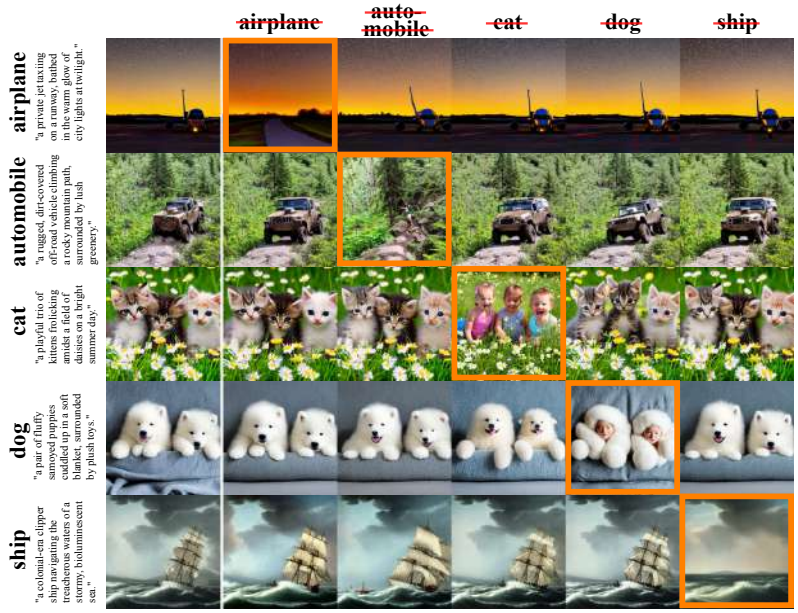

Fig. 4: Visualization of robustness and locality from *Receler* on CIFAR-10. The red strikethrough at the top indicates the erased concepts. On the left, the input paraphrased prompts are provided. Images enclosed within the diagonal orange borders shows robustness while others shows the locality.

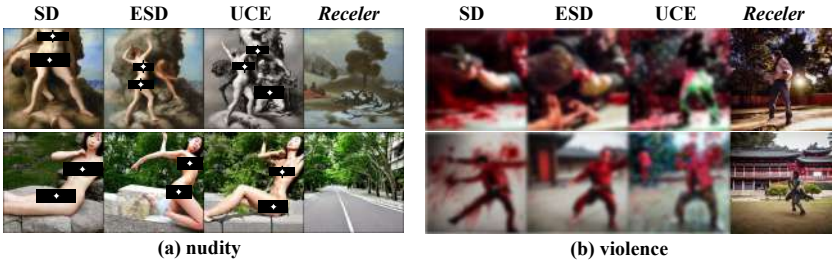

Fig. 5: Visualization of erasure methods against learned attack prompts. We use Ring-A-Bell [47] to generate adversarial prompts for nudity and violence concepts.

the prompt is paraphrased to "jet". Concerning locality, as shown in the non-diagonal boxes, *Receler* generates images that not only faithfully adhere to the descriptions but also closely resemble the images from the original diffusion model. Notably, it can be seen that *Receler* is able to replace the unlearned objects with reasonable alternatives rather than simply removing them, *e.g.*, two white puppies are substituted with two people wearing white furry clothes and gloves (see the fourth row in Fig. 4).

Table 6: Ablation study on CIFAR-10. We ablate the components of *Receler* and report the robustness and locality metrics. The first row refers to fine-tuning all parameters with only $\mathcal{L}_{Erase}$ in Eq. (1).

Components			Metrics	
Eraser	$\mathcal{L}_{\text{Reg}}$	$\mathcal{L}_{\text{Adv}}$	$\text{Acc}_R(\downarrow)$	$\text{Acc}_L(\uparrow)$
✗	✗	✗	39.4	75.2
✓	✗	✗	34.9	77.4
✓	✓	✗	26.3	79.8
✓	✗	✓	20.9	76.2
✓	✓	✓	**17.6**	**82.1**

Fig. 6: Examples of erasing multiple concepts. Instead of training an eraser for multiple concepts from scratch, we combine existing erasers for multi-concept erasure.

Visualization of Robustness against Learned Attack Prompts. In addition to the quantitative results shown in Tab. 5, we further qualitatively demonstrate the robustness of *Receler* against learned attack prompts in Fig. 5. The attack prompts are learned to induce the concepts of "nudity" and "violence" from models that should have erased these concepts using Ring-A-Bell [47]. For both nudity and violence attack prompts, *Receler* successfully prevents the generation of erased concepts, whereas other methods like ESD and UCE fail and generate images with the supposedly forbidden concepts (*e.g.*, nudity or blood).

Compositional Concept Erasure. In Fig. 6, we showcase examples of *Receler* in performing compositional concept erasure. This is achieved by combining outputs from separately trained erasers, each targeting a specific unlearned concept, during inference. By averaging these outputs, compositional concept erasure is accomplished without necessitating retraining. This approach notably offers the flexibility to selectively determine which concepts are to be erased, as required.

5 Conclusion

In this paper, we proposed **Re**liable **C**oncept **E**rasing via **L**ightweight **Er**asers (***Receler***) to erase target concepts entirely from the pre-trained diffusion model against prompting attacks (*i.e.*, robustness), while preserving its image generation ability of other concepts (*i.e.*, locality). In *Receler*, we employ concept-localized regularization to enforce the eraser to only affect the visual features related to the target concept. To enhance model robustness to paraphrased or attack prompts, we present an adversarial prompt learning scheme to induce the model to produce images of previously erased concepts, followed by optimizing the model against such image generation. We conducted extensive quantitative and qualitative evaluations on *Receler*, validating its superior robustness and locality-preserving ability over previous concept-erasing methods.

Acknowledgements. This research was supported in part by the National Science and Technology Council via grant NSTC 112-2634-F-002-007, NSTC 113-2640-E-002-003 and the Featured Area Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education 113L900902. We also thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

1. Sarah andersen. et al v. stability ai ltd. et al. case no.3:2023cv00201. us district court for the northern district of california. (2023)
2. Akhtar, N., Mian, A., Kardan, N., Shah, M.: Advances in adversarial attacks and defenses in computer vision: A survey. IEEE Access **9**, 155161–155196 (2021)
3. Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q.: Recent advances in adversarial training for adversarial robustness. arXiv preprint arXiv:2102.01356 (2021)
4. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
5. Bedapudi Praneeth, b.k., lireza Ayinmehr: Nudenet: Neural nets for nudity classification, detection and selective censoring (2019)
6. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
7. Carlini, N., Jagielski, M., Zhang, C., Papernot, N., Terzis, A., Tramer, F.: The privacy onion effect: Memorization is relative. In: NeurIPS (2022)
8. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 5253–5270 (2023)
9. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 5253–5270 (2023)
10. Chen, J., Yang, D.: Unlearn what you want to forget: Efficient unlearning for LLMs. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (Dec 2023)
11. Chin, Z.Y., Jiang, C.M., Huang, C.C., Chen, P.Y., Chiu, W.C.: Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. arXiv preprint arXiv:2309.06135 (2023)
12. De Cao, N., Aziz, W., Titov, I.: Editing factual knowledge in language models. arXiv preprint arXiv:2104.08164 (2021)
13. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: CVPR (2018)
14. 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. arXiv preprint arXiv:2208.01618 (2022)
15. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. arXiv preprint arXiv:2303.07345 (2023)
16. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. arXiv preprint arXiv:2308.14761 (2023)

17. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
20. Huang, Z., Shen, Y., Zhang, X., Zhou, J., Rong, W., Xiong, Z.: Transformer-patcher: One mistake worth one neuron. arXiv preprint arXiv:2301.09785 (2023)
21. Hunter, T.: Ai porn is easy to make now. for women, that's a nightmare. The Washington Post pp. NA–NA (2023)
22. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto (2009)
23. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: ICCV (2023)
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023)
25. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial intelligence safety and security, pp. 99–112. Chapman and Hall/CRC (2018)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
27. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
28. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
29. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. In: NeurIPS (2022)
30. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
31. Ni, M., Wu, C., Wang, X., Yin, S., Wang, L., Liu, Z., Duan, N.: Ores: Open-vocabulary responsible visual synthesis. arXiv preprint arXiv:2308.13785 (2023)
32. Nichol, A., 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. arXiv preprint arXiv:2112.10741 (2021)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
35. Rombach, R.: Stable diffusion 2.0 release (Nov 2022)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)

37. 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: CVPR (2023)
38. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
39. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: CVPR (2023)
40. Schramowski, P., Tauchmann, C., Kersting, K.: Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 1350–1361 (2022)
41. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
42. Sinitsin, A., Plokhotnyuk, V., Pyrkin, D., Popov, S., Babenko, A.: Editable neural networks. arXiv preprint arXiv:2004.00345 (2020)
43. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: CVPR (2023)
44. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: CVPR (2023)
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
46. Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., Ture, F.: What the daam: Interpreting stable diffusion using cross attention. arXiv preprint arXiv:2210.04885 (2022)
47. Tsai, Y.L., Hsu, C.Y., Xie, C., Lin, C.H., Chen, J.Y., Li, B., Chen, P.Y., Yu, C.M., Huang, C.Y.: Ring-a-bell! how reliable are concept removal methods for diffusion models? arXiv preprint arXiv:2310.10012 (2023)
48. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. arXiv preprint arXiv:2210.14896 (2022)
49. Xiang, C., Bao, F., Li, C., Su, H., Zhu, J.: A closer look at parameter-efficient tuning in diffusion models. arXiv preprint arXiv:2303.18181 (2023)
50. Yao, Y., Wang, P., Tian, B., Cheng, S., Li, Z., Deng, S., Chen, H., Zhang, N.: Editing large language models: Problems, methods, and opportunities. arXiv preprint arXiv:2305.13172 (2023)
51. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
52. 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:2303.17591 (2023)
53. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
54. Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., Liu, S.: To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. arXiv preprint arXiv:2310.11868 (2023)

55. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. arXiv preprint arXiv:2305.16322 (2023)