R.A.C.E. : Robust Adversarial Concept Erasure for Secure Text-to-Image Diffusion Model

Changhoon Kim[1(✉)], Kyle Min[2], and Yezhou Yang[1]

[1] Arizona State University, Tempe, USA
{kch,yz.yang}@asu.edu
[2] Intel Labs, Hillsboro, USA
kyle.min@intel.com

Abstract. In the evolving landscape of text-to-image (T2I) diffusion models, the remarkable capability to generate high-quality images from textual descriptions faces challenges with the potential misuse of reproducing sensitive content. To address this critical issue, we introduce **R**obust **A**dversarial **C**oncept **E**rase (RACE), a novel approach designed to mitigate these risks by enhancing the robustness of concept erasure method for T2I models. RACE utilizes a sophisticated adversarial training framework to identify and mitigate adversarial text embeddings, significantly reducing the Attack Success Rate (ASR). Impressively, RACE achieves a 30% reduction in ASR for the "nudity" concept against the leading white-box attack method. Our extensive evaluations demonstrate RACE's effectiveness in defending against both white-box and black-box attacks, marking a significant advancement in protecting T2I diffusion models from generating inappropriate or misleading imagery. This work underlines the essential need for proactive defense measures in adapting to the rapidly advancing field of adversarial challenges. Our code is publicly available: https://github.com/chkimmmmm/R.A.C.E.

Keywords: Concept Erasure · Responsible Image Generative Models · Secure T2I Diffusion Models

1 Introduction

The field of text-to-image (T2I) diffusion models has garnered significant attention for their ability to produce high-quality images that can be adaptively generated from textual descriptions [36,38]. This advancement is predicated on the training of T2I models with extensive datasets, often encompassing a range of

C. Kim and K. Min—These authors contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73010-8_27.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15141, pp. 461–478, 2025.
https://doi.org/10.1007/978-3-031-73010-8_27

Fig. 1. Comparative demonstration of concept erasure, red teaming, and robust erasure within T2I diffusion models. The ESD method [9] removes targeted concepts from the original SD outputs, yet these concepts can be reconstructed using UnlearnDiff [58]. Our proposed **R.A.C.E.** method showcases enhanced robustness against such red teaming reconstruction efforts.

content including copyrighted, explicit, and private materials [44–46]. Consequently, these models possess the capacity to inadvertently replicate protected images, potentially without user awareness [2,45,46]. The misuse of T2I models by malicious actors for misinformation or public opinion manipulation presents a significant concern [5,27].

In response to the challenges posed by the malicious exploitation of generative models, Stable Diffusion (SD) [38] has integrated a safety checker [37] and advocates for the utilization of a watermarking module [57]. Despite these initiatives, the reliance on post-hoc interventions presents limitations due to their potential for circumvention [7,18]. Consequently, the research community is pivoting towards formulating methodologies that embed safety protocols directly within the image generation pipeline. This approach ensures that content security is an intrinsic aspect of image creation [7,18,19,30,50,54,55]. Although these methods are valuable for identifying the source of content after an incident, their reactive nature highlights the necessity for proactive strategies. This includes pioneering techniques attempting early removal of sensitive content from T2I models [9,21], thereby preempting the production of inappropriate or harmful material.

To address the challenge of removing undesirable content from T2I models, even when users attempt to circumvent content restrictions, recent research has focused on the development of concept erasure techniques within T2I diffusion models [9,11,21,56]. These techniques primarily aim to eliminate specific concepts (e.g., "nudity") by altering the text embeddings associated with these concepts to neutral representations. Despite these efforts, there remains a vulnerability wherein erased concepts can be reconstructed. This is achieved by identifying text tokens that closely align with the visual embeddings of the targeted concepts, thus enabling the regeneration of prohibited content [3,47,49,53,58]. This issue is illustrated in Fig. 1, demonstrating that even with concept erasure, T2I models can be manipulated through prompt modification to regenerate the restricted content. This underscores the imperative for a more robust concept erasure methodology that can withstand such reconstruction attempts, ensuring the integrity of content generation within T2I models.

Acknowledging the imperative for enhanced concept erasure methodologies within T2I models, we pose a critical question: *"Is it feasible to develop a concept erasure approach that is resilient against reconstruction efforts?"* In pursuit of this, we introduce **R.A.C.E.** (**R**obust **A**dversarial **C**oncept **E**rase), a novel strategy aimed at bolstering the resilience of concept erasure techniques against adversarial manipulations, as delineated in Algorithm 1. At the heart of RACE lies an adversarial training framework, leveraging insights from the effort of adversarial robustness [26]. Our method effectively identifies adversarial text embeddings capable of reconstructing erased concepts and then facilitates their integration into the T2I concept erasure workflow.

A pivotal aspect of RACE is its ability to efficiently uncover adversarial text embeddings within a single time step of the diffusion process, an approach elaborated in Sect. 3.2. This efficiency not only streamlines the process of identifying adversarial examples but also facilitates the integration of our adversarial attack mechanism into the concept erasure workflow. To demonstrate the robustness of RACE, we have carried out an extensive array of experiments. These experiments validate the effectiveness of RACE in countering diverse red teaming strategies, with detailed results presented in Sect. 4. Our empirical investigations underscore the capacity of RACE to significantly enhance the robustness of T2I

models against both white-box and black-box attacks across a broad spectrum of target concepts, including artistic, explicit, and object categories.

We summarize the three main contributions here:

- We are the first to present, to our knowledge, an adversarial training approach specifically designed to fortify concept erasure methods against prompt-based adversarial attacks without introducing additional modules.
- Our method, RACE, implements a computationally efficient adversarial attack method that can be plugged into the concept erasing method.
- We show RACE significantly improves T2I models' robustness against prompts based on white/black box attacks.

2 Related Works

For a comprehensive overview of additional related works, please refer to the supplementary material.

Text-to-Image Synthesis. The field of generative models has seen remarkable advancements, notably extending their capabilities beyond generating photorealistic images [6,17] to include Text-to-Image (T2I) synthesis [29,33,36,38,41]. This progress has led to the development of fine-tuning techniques that allow for the customization of T2I models to user-specific needs [8,22,32,40,48], thereby enabling the creation of highly realistic images that align closely with textual prompts. However, the potential for misuse by malicious entities, using these models for purposes such as spreading misinformation [5,27], raises significant concerns. This underscores the urgency of devising protective measures to mitigate the risk of such exploitations.

Advanced Techniques in Concept Erasure for T2I Diffusion Models. Within the realm of machine unlearning, concept erasure for T2I Diffusion models has recently emerged as a critical area of research, focusing on the removal of sensitive or copyrighted concepts from T2I models. Methods to achieve this include guiding the image generation process or adjusting the model's weights to exclude these elements [9,11,21,24,28,43,56]. Notably, techniques by Gandikota et al. [9] and Kumari et al. [21] involve mapping sensitive concepts to null entities or benign equivalents by fine-tuning the weights of Stable Diffusion (SD) [38] models, effectively preventing the generation of undesirable content. Despite these advancements, red teaming methods have exposed potential loopholes, indicating that erased concepts might be regenerated through meticulously designed text prompts. Addressing this issue, our work contributes an adversarial training strategy aimed at bolstering the resilience of Stable Diffusion models against such text-prompt-based attacks [3,49,58], thereby enhancing the security and integrity of the content generation.

Robustness Evaluation via Red Teaming in T2I Models. While various safety measures have been proposed to shield SD models from misuse, red teaming strategies reveal vulnerabilities that still allow for circumvention. Research

Algorithm 1 **R**obust **A**dversarial **C**oncept **E**rasure: RACE Algorithm

Input: Diffusion Model Φ_θ, frozen diffusion model $\Phi_{\theta*}$, scheduler $\mathcal{S}$, target concept c, training steps M, adversarial steps N, perturbation limit ϵ, attack step size α
for $i = 0, \ldots, M$ **do**
  Sample noise $n \sim \mathcal{N}(0, 1)$, timestep $t \sim \mathcal{U}(1, 1000)$
  Initialize $\delta \sim \mathcal{U}(-\epsilon, \epsilon)$
  Denoise $z_t = \mathcal{S}(n, t, c)$
  for $j = 0, \ldots, N$ **do** ▷ Perform targeted attack
    $\delta = \delta + \alpha \cdot sign(\nabla_\delta - L_{SD}(\Phi_\theta, z_t, t, c, \delta))$
    Clamp δ within $[-\epsilon, \epsilon]$
  end for
  $\theta = \theta - \nabla_\theta L_{RACE}(\Phi_\theta, \Phi_{\theta^*}, z_t, t, c, \delta)$
end for
return Φ_θ

has demonstrated that techniques like Textual Inversion [8] can be exploited to regenerate content previously erased from SD models [34], prompting the development of countermeasures aimed at safeguarding against such inversions [52,59]. In real-world applications, T2I services such as Midjourney predominantly rely on user-provided text prompts, making them susceptible to prompt-based red teaming attacks [3,47,49,53,58]. These methods employ sophisticated prompt optimization techniques to restore images containing erased content, with their efficacy contingent upon the level of model access-categorized into white-box approaches, which utilize SD's U-Net [39] for prompt optimization [3,58], and black-box strategies, where such access is restricted [47,49,53]. Both approaches establish formidable benchmarks in attack success rates, as detailed in Table 2. However, the landscape lacks robust defense mechanisms against prompt-based red teaming, primarily due to the prohibitive computational demands associated with identifying adversarial prompts-a challenge that renders traditional adversarial training approaches impractical. Addressing this gap, our work introduces a novel defense strategy tailored to counteract prompt-based red teaming attacks, marking a significant step forward in fortifying T2I diffusion models against adversarial threats.

3 Method

Our methodology aims to expunge target concepts from T2I diffusion models through an adversarial training framework [26]. Initially, in Sect. 3.1, we establish the foundation by formally introducing the notations and the rationale underpinning our approach. Following this, Sect. 3.2 details our proposed adversarial attack, specifically designed for robust concept erasure, and delineates its integration into the adversarial training regime.

3.1 Preliminaries

Stable Diffusion Models. Our method is built upon the Stable Diffusion Model (SD) [38], which operates as the foundational architecture for our concept

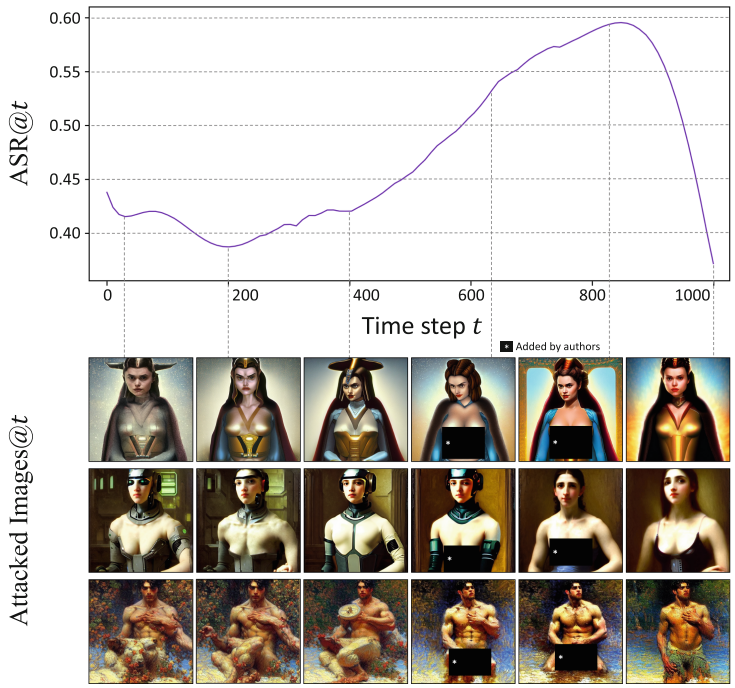

Fig. 2. Single-Timestep Adversarial Attack Efficacy. This figure illustrates the Attack Success Rate (ASR) across various timesteps, alongside representative images. Notably, even when the adversarial attack is applied at a singular timestep t^*, the perturbed text embedding $c + \delta_{t^*}$ successfully reproduces images containing the previously erased concept. For method details, see Sect. 3.2.

erasure technique. The SD model is composed of two primary elements: firstly, an image autoencoder that has been pre-trained on a diverse and extensive image dataset [6]. Within this autoencoder, an encoder function $\mathcal{E}(\cdot)$ transforms an input image x into a latent representation $z = \mathcal{E}(x)$. Conversely, a decoder function $\mathcal{D}(\cdot)$ aims to reconstruct the input image from its latent form, where $\mathcal{D}(z) = \hat{x} \approx x$.

The second element is a U-Net [39]-based diffusion model trained to craft latent representations within the acquired latent space. This model facilitates the conditioning on either class labels or text embeddings derived from training data. Let us denote by $c = \mathcal{E}_{txt}(y)$ the textual embedding encoded from a conditioning text prompt y, where $\mathcal{E}_{txt}$ symbolizes the text encoder, such as CLIP [35]. Under these constructs, the SD training objective is encapsulated by the loss function:

$$L_{SD} := \mathbb{E}_{n\sim\mathcal{N}(0,1),z,c,t}\left[||n - \Phi_\theta(z_t, t, c)||_2^2\right], \tag{1}$$

where t indexes the time step, n represents a noise sample drawn from a standard Gaussian distribution, z_t is the perturbed version of z up to time step t, and Φ_θ is the denoising network based on a U-Net architecture. During inference,

a random noise sample is procured from a Gaussian distribution and denoised using Φ_θ following a scheduler $\mathcal{S}$, operating over a sequence of predetermined time steps T. The resulting denoised latent, z_0, is then decoded to produce the final image, $\hat{x} = \mathcal{D}(z_0)$.

Erase and Reconstruction of Target Concept. The objective of concept erasure is to remove specific concepts, such as "nudity", from the latent space of a pre-trained diffusion model. The Erased Stable Diffusion (ESD) approach [9] introduces a method for excising these concepts from the latent representations within the Stable Diffusion framework. The erasure loss is formalized as follows:

$$L_{erase} := ||\Phi_\theta(z_t, t, c) - (\Phi_{\theta^*}(z_t, t) - \eta(\Phi_{\theta^*}(z_t, t, c) - \Phi_{\theta^*}(z_t, t)))||_2^2, \quad (2)$$

where Φ_{θ^*} represents the frozen denoising U-Net, and η denotes the guidance scale associated with classifier-free guidance [15]. ESD fine-tunes L_{erase} with respect to θ, guiding Φ_θ to produce outputs where the target concept is effectively nullified. Crucially, this process does not necessitate additional datasets; it operates successfully with only a concise textual description.

Conversely, red teaming efforts aim to counteract concept erasure by crafting adversarial text prompts $\tilde{y}$ capable of resurrecting the erased concept in the generated image. White-box methods [3,58] engage in this adversarial prompt optimization by leveraging the gradients of Φ_θ. Meanwhile, black-box approaches [47,49,53] aim to achieve comparable outcomes without reliance on the gradients of Φ_θ. Both approaches pose significant computational demands, which presents challenges for their integration into an adversarial training framework.

3.2 Adversarial Training on Concept Erased Diffusion Models

Motivation. The primary aim of T2I diffusion models is to generate high-quality images conditioned on specific prompts. Intriguingly, the SD model's loss function, as depicted in Eq.(1), can also facilitate image classification tasks [23]. This classification capability is derived by applying Bayes' Theorem to the model's predictions $p_\theta(x|c_i)$ and the prior distribution $p(c)$ across a set of conditions c_i, where each $c_i = \mathcal{E}_{txt}(y_i)$ represents a textual embedding of the prompt y_i:

$$p_\theta(c_i|x) = \frac{p(c_i)p_\theta(x|c_i)}{\sum_j p(c_j)p_\theta(x|c_j)}. \quad (3)$$

Notably, the prior terms $p(c)$ can be disregarded when they are uniformly distributed over the prompts c_i (i.e., $p(c_i) = \frac{1}{N}$). In the context of diffusion models, directly computing $p_\theta(x|c_i)$ is computationally challenging, leading to the reliance on the computation of $\log p_\theta(x|c_i)$ and the utilization of the Evidence Lower Bound (ELBO) for optimization purposes. Leveraging approximations introduced in [14], we can approximate the posterior distribution over prompts c_i as follows:

$$p_\theta(c_i|x) = \frac{exp\{-\mathbb{E}_{z,n,t}\left[||n - \Phi_\theta(z_t, t, c_i)||^2\right]\}}{\sum_j exp\{-\mathbb{E}_{z,n,t}\left[||n - \Phi_\theta(z_t, t, c_j)||^2\right]\}}. \quad (4)$$

Building on this foundation, the Diffusion Classifier [23] proposes a method for estimating the class of a given image x by finding argmin of the following expression:

$$\underset{c}{\operatorname{argmin}} \sum_{t=1}^{T} \left[||n - \Phi_\theta(z_t, t, c)||^2 \right], \tag{5}$$

where $z_t = \mathcal{E}(x)$ represents the latent encoding of the image x and the goal is to identify the label c from the available set of classes c_i. A notable insight from the Diffusion Classifier is the feasibility of classifying image x even when computations are performed using a single time step t^*.

Single-Timestep Adversarial Attacks on Erased T2I Models. Leveraging insights from the Diffusion Classifier, we re-conceptualize the SD model's loss function in Eq. (1) as a classification mechanism. This perspective allows us to view Textual Inversion (TI) [8] as a form of targeted adversarial attack, where the objective is to optimize the conditioning text embedding c to regenerate the image x [34,58]. Notably, TI is computationally intensive as it necessitates optimization across all time steps.

Prompted by these considerations, our investigation centers around a critical inquiry: *Can adversarial text embeddings be identified with just a single timestep?* We investigate whether adversarial text embeddings can be effectively identified at a singular timestep t^*. Our approach is geared towards nullifying the embedding of a target concept c and its proximate embeddings that might facilitate the regeneration of an erased concept image $\tilde{x}$, such as an explicit image. To this end, we devise a targeted adversarial attack to produce $\tilde{x}$ from a concept-erased diffusion model Φ_θ by introducing an adversarial perturbation δ. The perturbation δ is determined through the optimization:

$$\underset{||\delta||_\infty \leq \epsilon}{\operatorname{argmin}} ||n - \Phi_\theta(\tilde{z}_t, t, c + \delta)||_2^2, \tag{6}$$

where $\tilde{z} = \mathcal{E}(\tilde{x})$ denotes the latent representation of image $\tilde{x}$, ϵ is a small number, and $c = \mathcal{E}_{txt}(y)$ encodes the textual embedding of the targeted concept, for instance, $y =$"nudity". The Projected Gradient Descent (PGD) algorithm [26] is employed to address this optimization problem. Specifically, when selecting timestep $t^* = 500$ as the critical adversarial point, $\tilde{z}_t$ undergoes denoising via $\Phi_\theta(\tilde{z}_t, t, c)$ transitioning from timesteps $t = 1000$ to $t = 500$. Subsequently, the targeted adversarial approach outlined in Eq. (6) is executed to determine δ_{t^*} at $t = 500$. In subsequent denoising steps, $\tilde{z}_t$ is denoised considering the introduced perturbation $c + \delta_{t^*}$.

Our method introduces a distinctive single-timestep adversarial attack, contrasting with prior approaches that required optimization over a wide range of timesteps [3,58]. This approach enables the seamless incorporation of our attack strategy into the adversarial training process specifically for T2I concept erasure.

To assess the efficacy of our method, we execute tests on a Φ_θ model trained for removing the "nudity" concept via the ESD method. Utilizing 142 nudity-centric prompts from the I2P dataset [43], we systematically select timesteps t^*

at 100-step intervals within the $[1, 1000]$ range and compute the Attack Success Rate (ASR), as formulated in Eq.(8), for each t^*. The results, depicted in Fig. 2, unveil a notable capability: the reconstruction of the erased concept is feasible even with the attack confined to a single timestep t^*. Interestingly, after $t^* = 500$, we observe an increasing trend in the ASR. The modified images resulting from the attack visually convey the transition back to the erased concept and reflect the ASR patterns. For additional insights into various concepts and their corresponding ASR trends across different timesteps t^*, the supplementary material offers further information.

Adversarial Training for T2I Concept Erasure. Motivated by the findings from our single-timestep adversarial attack experiments, we explore the potential of such attacks to enhance the robustness of concept erasure in T2I models, posing the question: *Can adversarial attacks improve the resilience of concept erasure mechanisms?*

Table 1. Performance comparison of L_{erase} and L_{RACE}

	L_{erase}	L_{RACE}
I2P [43]	0.08	**0.05**
FID [13]	33.12	**25.16**
CLIP-Score [12]	0.726	**0.745**

RACE distinguishes itself from existing approaches [9,21] by aiming to eliminate not only the targeted concept's embedding but also its adjacent embedding within the model's latent space, which could otherwise lead to the inadvertent generation of the erased concept by Φ_θ. We incorporate our adversarial attack into the erasure loss function (L_{erase}), yielding an enhanced adversarial training loss:

$$L_{RACE} := ||\Phi_\theta(z_t, t, c + \delta) - (\Phi_{\theta^*}(z_t, t) - \eta(\Phi_{\theta^*}(z_t, t, c) - \Phi_{\theta^*}(z_t, t)))||_2^2. \quad (7)$$

This method is deliberate, substituting the concept embedding c with $c + \delta$ within the trainable parameters of Φ_θ. This precise adjustment ensures enhanced fidelity of the generated images by mapping the ϵ-neighborhood of the concept embedding to its null representation. Comparative metrics between the direct substitution in L_{erase} and the strategic use of L_{RACE} are provided in Table 1. The latter approach demonstrates promising reductions in ASR for "nudity" prompts within I2P dataset [43] and improvements in image quality metrics, as assessed on the MS-COCO [25]. The RACE methodology is comprehensively detailed in Algorithm 1. To validate RACE's effectiveness, we evaluate the ASR against both white-box and black-box attacks, as elaborated in Sect. 4.

4 Experiments

4.1 Experimental Setting

Datasets. Our assessment of the RACE framework spans various domains, including artistic styles, explicit concepts, and identifiable objects, in line with established benchmarks [3,9,58]. To ensure a uniform image generation process, we standardize key hyperparameters such as the scale of classifier-free guidance and random seeds. Artistic style evaluations leverage shared text prompts from

ESD [9] and UnlearnDiff [58]. For explicit content, we draw from the inappropriate image prompt benchmark [43], selecting a diverse set of prompts encompassing 142 nudity, 98 illegal acts, and 101 violence instances. Objects are curated from a subset of Imagenette [16], known for its distinct and recognizable classes, with text prompts synthesized via ChatGPT [31] following the approach by Kumari et al. [21]. It's crucial to note that these datasets serve solely to gauge the effectiveness of adversarial attacks; the adversarial training component of RACE does not necessitate the use of these specific prompts.

Training Details. In validating RACE, we choose the Erased Stable Diffusion (ESD) model [9] for its ability to erase a wide range of concepts, serving as an ideal testbed to showcase RACE's efficacy. We integrate RACE with ESD, optimizing with attack parameters $\epsilon = 0.1$ and $\alpha = \epsilon/4$. The optimization of model parameters θ utilizes the Adam optimizer [20] at a learning rate of $1e-5$, consistent with ESD. Additional details on attack parameter selection are available in the supplementary material.

Red Teaming Methods. To rigorously test RACE's robustness, we deploy a comprehensive suite of adversarial attacks, spanning both white-box and black-box approaches. Initial assessments utilize the I2P red teaming prompt dataset [43]. In black-box scenarios, we employ PEZ [49], which crafts adversarial prompts via CLIP [35]. In the white-box scenario, methods like P4D [3] and UnlearnDiff [58] are used, which generate adversarial prompts by leveraging gradients from the SD.

Evaluation. To gauge the robustness of RACE, we employ domain-specific classifiers: a ViT-base model [51] pre-trained on ImageNet [4] and fine-tuned on WikiArt [42] for artistic styles, Nudenet [1] for explicit content, and ResNet-50 [10] trained on ImageNet for object removal. We measure robustness using the Attack Success Rate (ASR):

$$\text{ASR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(f\left(SD(\tilde{y}_i)\right) = \tilde{y}_i\right), \tag{8}$$

where $\tilde{y}_i$ is the adversarial prompt, f is the classifier, and N is the number of prompts. Additionally, we assess image quality after applying RACE by generating 5,000 images from the MS-COCO [25] test set, computing the Frechet Inception Distance (FID) score [13] and the CLIP score [12] to evaluate RACE's impact on image fidelity while ensuring concept erasure.

4.2 Robust Concept Erase Against Red Teaming

In our comprehensive analysis, RACE undergoes a series of red teaming evaluations [3,43,49,58], encompassing both white-box and black-box techniques aimed at regenerating concepts targeted by RACE for erasure, as depicted in Fig. 1. The comparative analysis of ASR presented in Table 2 spans diverse conceptual domains, from "Van Gogh"-inspired artistry to explicit content such as "nudity"

Table 2. Attack Success Rate (ASR) against white (○)/black (●) box attacks. We conduct experiments on artistic style (Van Gogh), explicit concepts (nudity, violence, and illegal acts), and objects (church, golf ball, and parachute). "ESD/RACE-Concept" denotes the concept erased from the model. We also measure the quality of T2I models. We can observe that RACE reduces ASR by over 30% (-0.30 or below) in absolute value on Van Gogh, nudity, and church for UnlearnDiff [58], which is the previous SOTA attack method.

	Prompts	PEZ [49]	P4D [3]	UnlearnDiff [58]	CLIP-Score [12]	FID [13]
White/Black Box	●	●	○	○	-	-
ESD[9]-VanGogh	0.04	0.00	0.26	0.36	0.7997	19.16
ESD[9]-Nudity	0.14	0.08	0.75	0.80	0.7931	18.88
ESD[9]-Violence	0.27	0.13	0.84	0.79	0.7834	21.55
ESD[9]-Illegal	0.29	0.20	0.89	0.85	0.7854	21.50
ESD[9]-Church	0.16	0.00	0.58	0.68	0.7896	19.68
ESD[9]-GolfBall	0.04	0.00	0.16	0.16	0.7738	20.64
ESD[9]-Parachute	0.06	0.04	0.48	0.60	0.7865	19.72
RACE-VanGogh	0.00 (-0.04)	0.00 (-0.00)	0.00 (-0.26)	0.04 (-0.32)	0.8024	20.65
RACE-Nudity	0.05 (-0.09)	0.02 (-0.06)	0.49 (-0.26)	0.47 (-0.33)	0.7452	25.16
RACE-Violence	0.11 (-0.16)	0.08 (-0.05)	0.75 (-0.09)	0.68 (-0.11)	0.7374	28.71
RACE-Illegal	0.20 (-0.09)	0.13 (-0.07)	0.85 (-0.04)	0.80 (-0.05)	0.7591	24.87
RACE-Church	0.02 (-0.14)	0.00 (-0.00)	0.26 (-0.32)	0.38 (-0.30)	0.7730	23.92
RACE-GolfBall	0.00 (-0.04)	0.00 (-0.00)	0.10 (-0.06)	0.06 (-0.10)	0.7480	25.38
RACE-Parachute	0.02 (-0.04)	0.00 (-0.04)	0.24 (-0.24)	0.38 (-0.22)	0.7570	26.42

and "violence", extending to tangible objects like "churches", "golf balls", and "parachutes". Remarkably, RACE consistently diminishes ASR, notably surpassing 30% for "Van Gogh" styles, "nudity", and "church" categories, particularly outperforming UnlearnDiff [58], the current state-of-the-art in white-box adversarial methodologies. This marked decline in ASR underscores RACE's heightened robustness and delineates its capability as a potent, computationally efficient defense mechanism for T2I diffusion models against intricate adversarial attacks. Crucially, RACE's methodological advantage stems from its independence from external imagery or prompts, diverging from traditional red teaming techniques reliant on such data for prompt generation.

As demonstrated in Table 2, RACE achieves a significant 33% reduction in ASR for the "nudity" concept, underscoring its effectiveness. To further elucidate how RACE enhances ASR, we analyze the specific categories or elements it targets for removal. Utilizing Nudenet, we enumerate the body parts generated by various models-original SD, ESD, UnlearnDiff, and RACE-when prompted with nudity-related inputs from the I2P dataset. Illustrated in Fig. 3, the analysis reveals that while ESD enhances the original SD's resilience to such prompts, UnlearnDiff manages to bypass ESD's defenses, reconstructing explicit content. In contrast, RACE maintains its robustness even against the sophisticated UnlearnDiff attacks, showcasing the advanced protective capabilities of our approach in safeguarding against the regeneration of sensitive content.

One caveat to mention is that our experiments reveal a nuanced trade-off between robustness and image quality. While artistic style erasures maintain

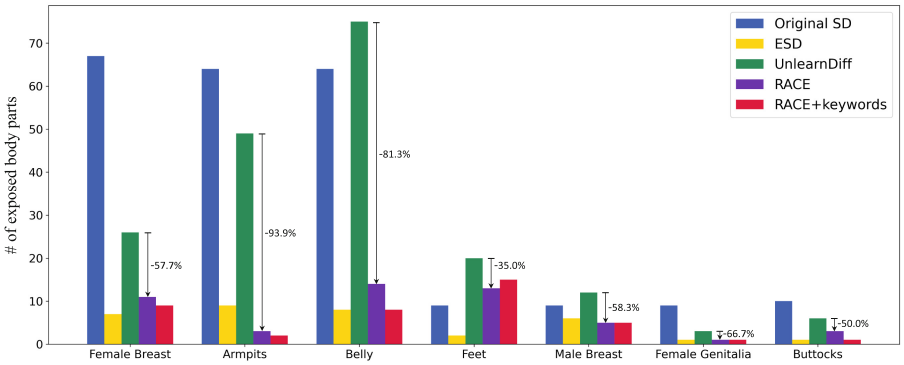

Fig. 3. Although ESD significantly reduces the chance of generating images with exposed body parts, state-of-the-art red teaming methods, such as UnlearnDiff, can be used to bypass ESD's defense and reconstruct explicit content. RACE and its variant can effectively defend the malicious attempts to reconstruct explicit content from the ESD model that erased the concept of nudity.

quality metrics, erasures of other concepts inadvertently degrade image quality. This divergence could be attributed to methodological differences; erasing artistic styles predominantly involves fine-tuning the cross-attention layers of SD as per ESD guidelines, whereas erasing other concepts necessitates adjustments in non-cross-attention layers [9]. Another plausible explanation for the observed trade-off between robust concept erasure and overall image quality could relate to the inherent complexity of differentiating between closely related or overlapping concepts within the model's latent space. As RACE intensifies adversarial robustness, it may inadvertently alter the delicate equilibrium within these conceptual overlaps, leading to unintended modifications in adjacent, non-targeted conceptual representations. This issue highlights the complex tension between precise concept erasure and maintaining the model's overall integrity against attacks. To address the quality concerns arising from this trade-off, we explore a potential strategy for improvement in Sect. 4.4, aiming to reduce the trade-off between targeted erasure with high quality and the model's defensive robustness.

4.3 Disentanglement

Investigating RACE's disentanglement performance, our study focuses on its capability to precisely erase intended concepts without impacting other elements. The evaluation spans both qualitative and quantitative measures. On the qualitative front, we configure separate RACE-enhanced models to specifically erase artistic imprints such as "Van Gogh", "Thomas Kinkade", and "Kilian Eng" from the SD. The illustrative outcomes, showcased in Fig. 4, underscore RACE's erasure precision, ensuring that the excision of one artistic style doesn't lead to the collateral removal of others. This meticulous erasure extends to discrete

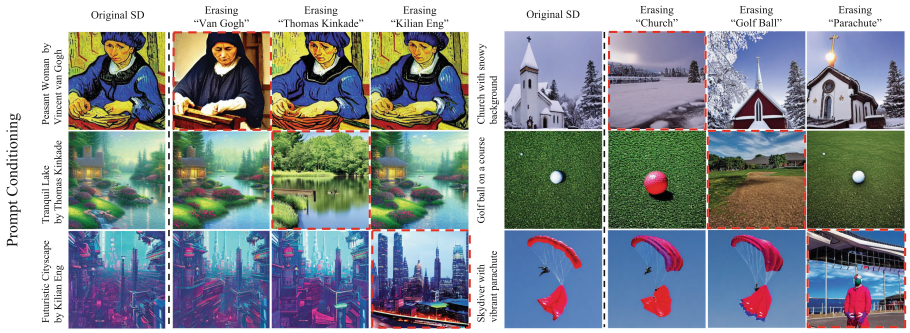

Fig. 4. RACE's Disentanglement in Concept Erasure. This figure highlights RACE's precision in erasing specific concepts, as shown in diagonal images, while preserving unrelated concepts, which is evident in off-diagonal images. For reference, baseline images generated by the original Stable Diffusion (SD) model are also presented.

object concepts like "church", "golf ball", and "parachute", with the generated images further affirming the method's discernment.

Quantitatively, we generate 5,000 prompts with varied random seeds, such as "an image of a [class name]", to produce a diverse set of images, subsequently evaluated using a pre-trained ResNet-50 classifier for top-1 accuracy. As shown in Table 3, RACE showcases an improved capability for object concept erasure compared to ESD. Despite this, there exists a slight decrement in classification accuracy for non-target classes. This effect likely stems from RACE's method of targeting the ϵ-neighborhood surrounding the intended concept, potentially influencing proximate concepts. Nevertheless, we can observe from Fig. 4 that RACE can precisely erase the target concept with minimal visual impacts on other concepts.

Table 3. Accruacy of erased and non-erased classes in Imagenet [4]. We evaluate the classification accuracy of ESD and RACE for the target erased concepts and other non-target concepts.

Erased Concept	Acc. Erased		Acc. Others	
	ESD	RACE	ESD	RACE
Church	0.16	0.02	0.57	0.53
Golf Ball	0.04	0.00	0.45	0.35
Parachute	0.06	0.02	0.57	0.41

4.4 Discussion

Potential Strategy to Improve the Robustness-Quality Trade-off. Our findings underscore the ability of RACE to significantly bolster the SD model's defense against prompt-based adversarial attacks across a variety of concepts. However, as illustrated in Table 2, enhancing robustness appears to inversely impact image quality. To address this dichotomy, we test a refined version of the RACE loss function incorporating a regularization term:

$$L_{RACE+Reg.} := L_{RACE} + \lambda||\theta - \theta^*||_1, \tag{9}$$

Table 4. Ablation Studies for Performance Improvement. Using the "nudity" concept, we evaluate the effectiveness of adding weight regularization and close concept keywords. ASR and quality metrics are measured to assess performance.

	I2P [43]	PEZ [49]	P4D [3]	UnleranDiff [58]	CLIP-Score [12]	FID [13]
ESD	0.14	0.08	0.75	0.80	0.7931	18.88
RACE	0.05	0.02	0.49	0.47	0.7452	25.16
RACE+Reg.	0.07	0.02	0.60	0.62	0.7593	24.42
RACE+keywords	0.02	0.01	0.42	0.46	0.7201	30.97

where θ and θ^* represent the parameters of the RACE and original Stable Diffusion models, respectively, and λ is the regularization strength, set to 0.1 in our experiment. This regularization approach, as evidenced in Table 4, helps to partially reconcile the robustness-quality trade-off, enhancing image quality while maintaining improved robustness over ESD.

Enhancing Concept Erasure. In pursuit of further reducing the ASR, we explore strategies for more comprehensive concept erasure. Recognizing that a target concept may manifest in various synonymous forms, we extend RACE's erasure scope to include semantically related concepts. Leveraging the CLIP text encoder embedded within Stable Diffusion, we identify and subsequently erase concepts closely related to the target, based on their proximity in the CLIP embedding space. For instance, alongside "nudity", we also target synonymous concepts like "nude", "nsfw", and "bare", identified as the top-3 semantically similar terms. In Fig. 3, we can observe that our method equipped with this expanded erasure strategy (denoted as RACE+keywords) is more effective in defending the malicious attempts to bypass the ESD by further reducing the number of exposed body parts. Table 4 also indicates that our strategy indeed fortifies the model's robustness against red teaming tactics. Nonetheless, this broadened concept removal spectrum reaffirms the robustness-quality trade-off, manifesting as a decrement in image quality. This aspect opens an intriguing avenue for future enhancements to the RACE methodology, balancing the twin objectives of robust concept erasure and preserved image fidelity.

5 Conclusion

In this work, we present RACE, a novel defense approach designed to protect the Text-to-Image Stable Diffusion models from prompt-based red teaming attacks. RACE effectively strengthens the model's concept erasure capabilities while maintaining computational efficiency, offering a valuable enhancement to the current erasure framework and bolstering defenses against various adversarial techniques. We also observe the robustness-quality trade-off and discuss possible future directions to improve it. This initial contribution lays the groundwork for further exploration, underscoring the critical importance of developing sophisticated defenses in the rapidly evolving domain of generative AI.

Acknowledgements. We would like to express our gratitude to Sangmin Jung for assisting with the experiments. This work is partially supported by the National Science Foundation under Grant No. 2038666 and No. 2101052. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

1. Bedapudi, P.: Nudenet: neural nets for nudity classification, detection and selective censoring, December 2019
2. Carlini, N., et al.: Extracting training data from diffusion models. ArXiv **abs/2301.13188** (2023). https://api.semanticscholar.org/CorpusID:256389993
3. 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)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Devlin, K., Cheetham, J.: Fake trump arrest photos: how to spot an AI-generated image. BBC News (March 2023). https://www.bbc.com/news/technology-68981525
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12873–12883 (2021)
7. Fernandez, P., Couairon, G., Jégou, H., Douze, M., Furon, T.: The stable signature: rooting watermarks in latent diffusion models. arXiv preprint arXiv:2303.15435 (2023)
8. Gal, R., et al.: An image is worth one word: personalizing text-to-image generation using textual inversion. ArXiv **abs/2208.01618** (2022). https://api.semanticscholar.org/CorpusID:251253049
9. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2426–2436 (2023). https://api.semanticscholar.org/CorpusID:257495777
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Heng, A., Soh, H.: Selective amnesia: a continual learning approach to forgetting in deep generative models. ArXiv **abs/2305.10120** (2023). https://api.semanticscholar.org/CorpusID:258740988
12. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: a reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
13. 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. In: Advances in Neural Information Processing Systems, pp. 6626–6637 (2017)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. **33**, 6840–6851 (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)

16. Howard, J., Gugger, S.: Fastai: a layered api for deep learning. Inf. **11**, 108 (2020). https://api.semanticscholar.org/CorpusID:211082837
17. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
18. Kim, C., Min, K., Patel, M., Cheng, S., Yang, Y.: Wouaf: weight modulation for user attribution and fingerprinting in text-to-image diffusion models. arXiv preprint arXiv:2306.04744 (2023)
19. Kim, C., Ren, Y., Yang, Y.: Decentralized attribution of generative models. In: International Conference on Learning Representations (2021)
20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). http://arxiv.org/abs/1412.6980, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015
21. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 22634–22645 (2023). https://api.semanticscholar.org/CorpusID:257687839
22. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1931–1941 (2022). https://api.semanticscholar.org/CorpusID:254408780
23. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2206–2217 (October 2023)
24. Li, S., et al.: Get what you want, not what you don't: image content suppression for text-to-image diffusion models. ArXiv **abs/2402.05375** (2024). https://api.semanticscholar.org/CorpusID:267547985
25. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
26. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
27. Marcelo, P.: Fact focus: fake image of pentagon explosion briefly sends jitters through stock market. Associated Press (May 2023). https://apnews.com/article/pentagon-explosion-fake-image-stock-market-jitters-78d5913603bdf9f17cc7a8c77a72e59b
28. Ni, M., et al.: Ores: open-vocabulary responsible visual synthesis. ArXiv **abs/2308.13785** (2023). https://api.semanticscholar.org/CorpusID:261243073
29. Nichol, A.Q., et al.: Glide: towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning, pp. 16784–16804. PMLR (2022)
30. Nie, G., Kim, C., Yang, Y., Ren, Y.: Attributing image generative models using latent fingerprints. arXiv preprint arXiv:2304.09752 (2023)
31. OpenAI: Chatgpt. Online (2022). https://chat.openai.com/chat. Accessed 24 Feb 2024
32. Patel, M., Jung, S., Baral, C., Yang, Y.: λ-eclipse: multi-concept personalized text-to-image diffusion models by leveraging clip latent space. ArXiv **abs/2402.05195** (2024). https://api.semanticscholar.org/CorpusID:267547418

33. Patel, M., Kim, C.S., Cheng, S., Baral, C., Yang, Y.: Eclipse: a resource-efficient text-to-image prior for image generations. ArXiv **abs/2312.04655** (2023). https://api.semanticscholar.org/CorpusID:266149498
34. Pham, M., Marshall, K.O., Cohen, N., Mittal, G., Hegde, C.: Circumventing concept erasure methods for text-to-image generative models. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=ag3o2T51Ht
35. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
37. Rando, J., Paleka, D., Lindner, D., Heim, L., Tramèr, F.: Red-teaming the stable diffusion safety filter. arXiv preprint arXiv:2210.04610 (2022)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, June 2022
39. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
40. 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: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22500–22510 (2022). https://api.semanticscholar.org/CorpusID:251800180
41. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural Inf. Process. Syst. **35**, 36479–36494 (2022)
42. Saleh, B., Elgammal, A.: Large-scale classification of fine-art paintings: learning the right metric on the right feature. ArXiv **abs/1505.00855** (2015). https://api.semanticscholar.org/CorpusID:14168099
43. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: mitigating inappropriate degeneration in diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22522–22531 (2022). https://api.semanticscholar.org/CorpusID:253420366
44. Schuhmann, C., et al.: Laion-5b: an open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402 (2022)
45. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? Investigating data replication in diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6048–6058 (2022). https://api.semanticscholar.org/CorpusID:254366634
46. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Understanding and mitigating copying in diffusion models. ArXiv **abs/2305.20086** (2023). https://api.semanticscholar.org/CorpusID:258987384
47. Tsai, Y.L., et al.: Ring-a-bell! How reliable are concept removal methods for diffusion models? ArXiv **abs/2310.10012** (2023). https://api.semanticscholar.org/CorpusID:264146485

48. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: encoding visual concepts into textual embeddings for customized text-to-image generation. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15897–15907 (2023). https://api.semanticscholar.org/CorpusID:257219968
49. Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., Goldstein, T.: Hard prompts made easy: gradient-based discrete optimization for prompt tuning and discovery. Adv. Neural Inf. Process. Syst. **36** (2024)
50. Wen, Y., Kirchenbauer, J., Geiping, J., Goldstein, T.: Tree-ring watermarks: fingerprints for diffusion images that are invisible and robust. arXiv preprint arXiv:2305.20030 (2023)
51. Wu, B., et al.: Visual transformers: token-based image representation and processing for computer vision. ArXiv **abs/2006.03677** (2020). https://api.semanticscholar.org/CorpusID:219531480
52. Wu, Y., Zhang, J., Kerschbaum, F., Zhang, T.: Backdooring textual inversion for concept censorship. ArXiv **abs/2308.10718** (2023). https://api.semanticscholar.org/CorpusID:261049298
53. Yang, Y., Gao, R., Wang, X., Xu, N., Xu, Q.: Mma-diffusion: multimodal attack on diffusion models. ArXiv **abs/2311.17516** (2023). https://api.semanticscholar.org/CorpusID:265498727
54. Yu, N., Skripniuk, V., Abdelnabi, S., Fritz, M.: Artificial fingerprinting for generative models: rooting deepfake attribution in training data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14448–14457 (2021)
55. Yu, N., Skripniuk, V., Chen, D., Davis, L., Fritz, M.: Responsible disclosure of generative models using scalable fingerprinting. arXiv preprint arXiv:2012.08726 (2020)
56. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: learning to forget in text-to-image diffusion models. ArXiv **abs/2303.17591** (2023). https://api.semanticscholar.org/CorpusID:257833863
57. Zhang, K.A., Xu, L., Cuesta-Infante, A., Veeramachaneni, K.: Robust invisible video watermarking with attention (2019)
58. Zhang, Y., et al.: To generate or not? Safety-driven unlearned diffusion models are still easy to generate unsafe images ... for now. ArXiv **abs/2310.11868** (2023). https://api.semanticscholar.org/CorpusID:264289091
59. Zheng, Y., Yeh, R.A.: Imma: immunizing text-to-image models against malicious adaptation. ArXiv **abs/2311.18815** (2023). https://api.semanticscholar.org/CorpusID:265506125