Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models

Chao Gong[1,2⋆], Kai Chen[1,2⋆], Zhipeng Wei[1,2], Jingjing Chen[1,2†], and Yu-Gang Jiang[1,2]

[1] Shanghai Key Lab of Intell. Info. Processing, School of Computer Science, Fudan University
[2] Shanghai Collaborative Innovation Center on Intelligent Visual Computing
{cgong20,chenjingjing,ygj}@fudan.edu.cn, {kchen22,zpwei21}@m.fudan.edu.cn

Abstract. Text-to-image models encounter safety issues, including concerns related to copyright and Not-Safe-For-Work (NSFW) content. Despite several methods have been proposed for erasing inappropriate concepts from diffusion models, they often exhibit incomplete erasure, consume a lot of computing resources, and inadvertently damage generation ability. In this work, we introduce Reliable and Efficient Concept Erasure (RECE), a novel approach that modifies the model in 3 seconds without necessitating additional fine-tuning. Specifically, RECE efficiently leverages a closed-form solution to derive new target embeddings, which are capable of regenerating erased concepts within the unlearned model. To mitigate inappropriate content potentially represented by derived embeddings, RECE further aligns them with harmless concepts in cross-attention layers. The derivation and erasure of new representation embeddings are conducted iteratively to achieve a thorough erasure of inappropriate concepts. Besides, to preserve the model's generation ability, RECE introduces an additional regularization term during the derivation process, resulting in minimizing the impact on unrelated concepts during the erasure process. All the processes above are in closed-form, guaranteeing extremely efficient erasure in only 3 seconds. Benchmarking against previous approaches, our method achieves more efficient and thorough erasure with minor damage to original generation ability and demonstrates enhanced robustness against red-teaming tools. Code is available at https://github.com/CharlesGong12/RECE.
WARNING: This paper contains model outputs that may be offensive.

Keywords: Text-to-Image · Concept Erasing · Machine Unlearning

1 Introduction

In recent years, large-scale text-to-image (T2I) diffusion models have exhibited remarkable capability in synthesizing photo-realistic images from text prompts [15,

⋆ Equal contributions.
† Corresponding author.

20, 24, 28]. The exceptional performance of T2I diffusion models is largely due to the vast amount of training data collected from the Internet, which enables the models to imitate a wide variety of concepts. Unfortunately, such powerful models can also be misused to generate copyright infringement and Not-Safe-For-Work (NSFW) image content when conditioned on inappropriate text prompts [11, 30]. Especially the open-source release of the Stable Diffusion (SD) T2I model has made advanced image generation technology widely accessible. To alleviate this safety concern, several recent research efforts have incorporated safety mechanisms into T2I diffusion models, *e.g.* filtering out inappropriate training data and retraining model [22], censoring model outputs with an NSFW safety checker [21], and applying classifier-free guidance to steer the generation away from inappropriate concepts [29]. However, these safety mechanisms either demand expensive computational resources and time [25] or can be easily circumvented by malicious users due to the public availability of code and model parameters in open-source scenario [23].

In response to the drawbacks mentioned above, an alternative is to erase inappropriate concepts from the T2I diffusion model [6–8, 13]. Specifically, given an inappropriate concept described in the text prompt, the pre-trained T2I diffusion model's parameters are fine-tuned to unlearn that concept so that the associated image content cannot be generated. Compared with previous security mechanisms, concept erasure neither requires training the entire model from scratch nor can be easily circumvented even in the case of open-source code. Despite promising progress in concept erasure, there exist several issues. On the one hand, most erasure methods require a high number of iterations to fine-tune considerable amounts of parameters [6, 8, 13], which inevitably degrades the generation capability and consumes a lot of computing resources. Only a recent work called UCE [7] modifies model parameters without fine-tuning using a closed-form solution, ensuring the model maintains original generation capability when erasing concepts. On the other hand, almost all methods fail to sufficiently erase inappropriate concepts, leaving them vulnerable to problematic prompts found by the red-teaming of T2I diffusion models [3, 32, 38]. This results in the unlearned model being compelled to regenerate inappropriate images.

Inspired by the idea of adversarial fine-tuning, we propose a **R**eliable and **E**fficient **C**oncept **E**rasure (RECE) method to address the aforementioned challenge, which continually finds new embeddings of the erased concepts during fine-tuning and then enables the unlearned model to erase these new concept embeddings. To speed up the unlearning process, the RECE method builds upon the previous fast and efficient concept erasure method UCE [7], which employs a closed-form editing to only modify the key and value projection matrices in cross-attention layers [33]. Similarly, the RECE method derives new embeddings that most effectively prompt the model to regenerate images of erased concepts, with a closed-form solution based on cross-attention output. Furthermore, a regularization term is introduced to preserve the image generation capability of the model by restricting the deviation of model parameters before and after modification. By editing the model and deriving embedding for multiple epochs,

RECE enables the unlearned model to preserve the image generation ability of unerased concepts and robustly refrains the model from generating images with erased concept content. All the processes above are in closed-form, guaranteeing extremely efficient erasure in 3 seconds. Our major contributions are summarized as follows:

- We present a novel concept erasure method - RECE that uses closed-form parameter editing and adversarial learning schemes for reliable and efficient concept erasing in only 3 seconds.
- RECE sufficiently erases concepts by deriving new embeddings that enable the unlearned model to regenerate erased concepts. In addition, a regularization term is introduced to minimize the impact on the model's capabilities.
- We conduct extensive experiments to validate the effectiveness of RECE for erasing unsafe contents, artistic styles and object classes. Additionally, we assess the robustness of RECE against three red-teaming tools and record fine-tuning durations to highlight the efficiency of RECE.

2 Related Work

T2I Diffusion Models with Safety Mechanisms. In response to the issue of generating inappropriate images in T2I diffusion models, several research has explored solutions to address this concern. Briefly, existing research primarily falls into the following three distinct strategies: The first is filtering the training data and retraining the model [22]. However, retraining on curated datasets not only requires substantial computational resources investment but also results in the generation of inappropriate content [6] and performance degradation [29]. The second is censoring model output through safety checkers [21], or exploiting classifier-free guidance to steer the latent codes away from inappropriate concepts during inference [29]. However, in the case of open-source code, pre-trained T2I diffusion model architectures and parameters are publicly available, so such post-hoc intervention strategies can be easily circumvented by malicious users [23]. The third is fine-tuning the partial parameters of the pre-trained T2I diffusion models to erase the model's representation capability of inappropriate concepts [6–8, 13]. While fine-tuning has been considered an effective strategy to prevent the generation of inappropriate content, current methods consume a lot of computing time and can be easily bypassed by red-teaming tools for T2I diffusion models.

Red-Teaming Tools for T2I Diffusion Models. With the recent popularity of AI, red-teaming has been applied to AI models to enhance model stability by probing functional vulnerabilities [1, 2, 34, 35]. Recent works have also developed red-teaming tools for T2I diffusion models, which is a rarely explored field in AI red-teaming. For instance, Prompting4Debugging (P4D) [3] automatically finds the problematic prompts that would lead to inappropriate content via utilizing prompt engineering techniques and an auxiliary diffusion model

without any safety mechanisms to assess the reliability of deployed safety mechanisms. Conversely, UnlearnDiff [38] does not depend on an auxiliary diffusion model, which leverages the inherent classification capabilities of diffusion models, thereby providing computational efficiency without sacrificing effectiveness. Both works have the main weakness in assuming white-box access to the target model. In response to this issue, Ring-A-Bell [32], a model-agnostic framework capable of constructing attacks without prior knowledge of the target model. Specifically, Ring-A-Bell first performs concept extraction to obtain a holistic representation of inappropriate concepts. Subsequently, Ring-A-Bell automatically produces problematic prompts by leveraging the extracted concepts.

3 Method

3.1 Preliminaries

Text-to-Image (T2I) Diffusion Models In contemporary Text-to-Image (T2I) applications, diffusion models have become the preferred choice [36] since the progressive denoising process [10] empowers them with superior image synthesis ability [4]. To reduce computational complexity, T2I often adopts latent diffusion models [12, 24, 27, 28], which operates on the low-dimensional latent space of a pre-trained variational autoencoder (VAE) [5] and employs a U-Net generative network as the denoising architecture [26]. To incorporate text conditioning into the image generation process, T2I encodes text by language models like CLIP [19, 24] and integrates text embeddings into U-Net through cross-attention layers. Specifically, these layers employ Query-Key-Value (QKV) structure [33] to represent the interactions between text and vision. For a given text embedding c_i, keys and values are generated as $k_i = W_k c_i$ and $v_i = W_v c_i$. These keys compute an attention map by multiplying with the query q_i representing visual features, and then the cross-attention output is computed by attending over values v_i:

$$\mathcal{O} \propto \text{softmax}(q_i k_i^T) v_i. \tag{1}$$

Concept Erasing with Closed-form Solution There are existing erasure methods requiring fine-tuning, such as ESD, CA and SA [6,8,13]. However, such methods are relatively inefficient as they require thousands of fine-tuning steps. In contrast, UCE [7] is an efficient method which modifies the attention weights through a closed-form edit. UCE requires a "source" concept (*e.g.*, "nudity") and a "destination" concept (*e.g.*, empty text " "). Let c_i represent the source embedding, c_i^* denote the corresponding destination embedding, set E denote concepts to erase, and set P denote concepts to preserve. Given a K/V projection matrix W^{old} (a concise notation for W_k^{old} and W_v^{old}), UCE seeks new weights W by editing concepts in E while preserving concepts in P. Specifically, the objective is to find weights such that the output Wc_i for $c_i \in E$ aligns with target values $W^{\text{old}} c_i^*$ instead of the original $W^{\text{old}} c_i$. Meanwhile, to control parameter

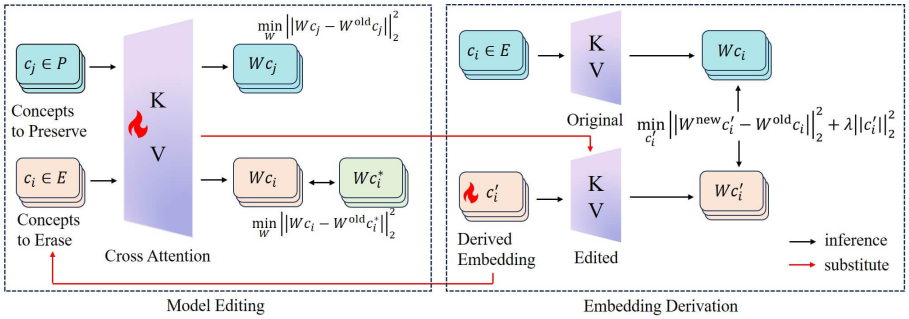

Fig. 1: Overview of the proposed RECE. RECE consists of two main components: model editing and embedding derivation. First, erasing concepts by editing the model with a closed-form solution, and obtaining the edited cross-attention W^{new}. Then, new embedding c_i' can be derived by Eq. (7) given the original cross attention W^{old} and the edited W^{new}. In subsequent epochs, model editing and embedding derivation are looped.

changes, outputs for $c_j \in P$ are preserved as $W^{\text{old}}c_j$ and a L2 regularization term is introduced:

$$\min_W \sum_{c_i \in E} ||Wc_i - W^{\text{old}}c_i^*||_2^2 + \lambda_1 \sum_{c_j \in P} ||Wc_j - W^{\text{old}}c_j||_2^2 + \lambda_2 ||W - W^{\text{old}}||_F^2, \quad (2)$$

where λ_1 and λ_2 are scaling factors preserving the existing concepts. UCE [7] prove that this formula has a closed-form solution:

$$W = W^{\text{old}}(\sum_{c_i \in E} c_i^* c_i^T + \lambda_1 \sum_{c_j \in P} c_j c_j^T + \lambda_2 I)(\sum_{c_i \in E} c_i c_i^T + \lambda_1 \sum_{c_j \in P} c_j c_j^T + \lambda_2 I)^{-1}. \quad (3)$$

UCE directly assigns cross-attention KV matrices using the closed-form solution, eliminating the need for fine-tuning. This makes UCE significantly faster, hence we use UCE in our method.

3.2 Reliable and Efficient Concept Erasure (RECE)

While UCE [7] offers a fast solution for removing undesired concepts from T2I diffusion models, it can still produce undesired content, as illustrated in Tab. 1. This suggests an incomplete erasure of these concepts. To effectively eliminate such undesired concepts, we efficiently erase closed-form embeddings capable of regenerating erased concepts within the unlearned model. The derivation of embeddings and erasure is conducted iteratively to achieve a thorough erasure of inappropriate concepts, as shown in Fig. 1.

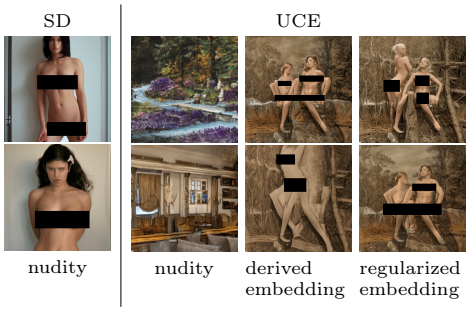

Fig. 2: When given the input prompt "nudity", SD generates images containing nudity content, while UCE generates unrelated images. When given our derived embedding and regularized embedding, UCE generates nude images again. We use ■ for publication purposes.

Finding Target Contents Let us take "nudity" for example. As depicted in the second column of Fig. 2, when directly providing the input prompt "nudity" to UCE models, only landscape or unrelated images are generated. This is because the word "nudity" has been aligned with the empty text " ". However, the erasure of UCE is incomplete. We can generate an adversarial prompt that enables UCE's model to produce images containing nudity content again, similar to those generated by SD when provided with the prompt "nudity".

In this section, we introduce our method for deriving the new embedding in UCE's model, which guides UCE to generate nude images. As elaborated in Sec. 3.1, T2I introduces text embeddings into image generation through cross-attention layers, where the projection matrices W_k and W_v are used to transform text embeddings. Let W^{old} denote the projection matrices of the original U-Net before UCE's editing, W^{new} represent the projection matrices after UCE's editing, c denote the embedding of "nudity", and c' signify our derived embedding. If we can find a c' such that $W^{\text{new}}c'$ closely resembles $W^{\text{old}}c$, then c' can guide the edited model to generate nude images like how c guides the original model. More precisely, the objective function is formulated as follows:

$$\min_{c'} \sum_i \|W_i^{\text{new}} c' - W_i^{\text{old}} c\|_2^2, \tag{4}$$

where W_i denotes K/V cross-attention matrices of U-Net. The solution c' derived from Eq. (4) can be viewed as the actual representation of c within the edited model. Evidently, Eq. (4) represents a convex function with respect to c', which possesses a unique global minimum. As derived in Appendix A, Eq. (4) admits a closed-form solution:

$$c' = \left(\sum_i {W_i^{\text{new}}}^T W_i^{\text{new}} \right)^{-1} \left(\sum_i {W_i^{\text{new}}}^T W_i^{\text{old}} \right) c. \tag{5}$$

Given that text conditioning works in the form of embedding in Stable Diffusion (SD), we can use our derived embedding as text conditioning. As illustrated in the third column of Fig. 2, this derived embedding effectively guides UCE's edited model to once again generate nude images, indicating its capacity to represent the concept of "nudity" within edited model. Thus it demonstrates that the erasure process of UCE remains incomplete. To address this issue, we further

remove our derived embeddings c' from UCE's model with Eq. (3) to prevent the generation of nude images.

Regularization Term We can further erase the concept of "nudity" by substituting c in Eq. (2) with our derived embedding c'. However, upon directly erasing c', we observe a significant decline in the model's performance: it struggles to generate high-quality images for unrelated concepts as shown in Fig. 6. Hence, it becomes imperative to devise a method that retains the model's performance while erasing concepts.

Let W^{new1} denote the projection matrices after the last epoch's modification, and W^{new2} denote the projection matrices after the current epoch. Partly, preserving the model's performance entails minimizing the impact on unrelated concepts. Consequently, we define our objective function as follows:

$$\min_{W} \|W^{\text{new2}}d - W^{\text{new1}}d\|_2^2, \tag{6}$$

where d represents an unrelated concept's embedding. Note that W^{new2} is directly influenced by c', as it is derived after erasing c' from W^{new1}. We can obtain a theorem about our objective and its proof is provided in Appendix B:

Theorem 1. *If c' is set to $\mathbf{0}$, Eq. (6) achieves its global minimum of 0.*

Intuitively, if c_i in Eq. (3) is set to zero, the coefficient matrices on the right side will become an identity matrix when multiplied. As a result, W will revert to being equivalent to W^{old}, exerting minimal influence on unrelated concepts due to the absence of further modifications to W. Therefore, we need to introduce a regularization term to the original objective Eq. (4) to ensure that the obtained c' is close to zero, thereby minimizing its influence on the model's performance:

$$\min_{c'} \sum_i \|W_i^{\text{new}}c' - W_i^{\text{old}}c\|_2^2 + \lambda\|c'\|_2^2. \tag{7}$$

As derived in Appendix A, this final objective function also possesses a unique global minimum solution:

$$c' = \left(\lambda I + \sum_i {W_i^{\text{new}}}^T W_i^{\text{new}}\right)^{-1} \left(\sum_i {W_i^{\text{new}}}^T W_i^{\text{old}}\right) c \tag{8}$$

As illustrated in the fourth column of Fig. 2, this regularized embedding can also guide UCE's model to once again generate nude images, indicating its ability to represent the concept "nudity" within UCE's model. With the incorporation of our regularization term, we iteratively apply the erasure process to the refined embedding c' using Eq. (3) over multiple epochs. This ensures thorough concept erasure while safeguarding the overall performance of the model. The algorithm details are elaborated in Algorithm 1.

Algorithm 1: Erase Concepts with *RECE*

Input: Diffusion U-Net θ, concepts set E to erase, P to preserve, epochs T.
Output: Diffusion U-Net θ' with concepts E erased.

```
/* Initialize */
θ' ← θ;
Initialize text embeddings c_i and c_j from E and P;
Extract K&V matrices W^old from the cross attention of θ;
/* Preliminary erasing with UCE */
Obtain W^new by erasing concepts in E with Eq. (3);
for t = 1, ..., T do
    /* Derive new embeddings */
    E' ← {};
    for c_i ∈ E do
        Derive new embedding c'_i using W^new with Eq. (8);
        E' ← E' ∪ {c'_i}
    /* Erasing derived embeddings */
    Obtain W^new' by erasing E' with Eq. (3);
    W^new ← W^new';
    Update θ' by replacing K&V matrices with W^new;
return θ'
```

4 Experiments

In this section, we present the results of our method for erasing inappropriate concepts and artistic styles. We also include the results of object removal in Appendix. We start with SD V1.4 as our base model. Following the implementation in [16], we set λ_1 and λ_2 in Eq. (2) to 0.1. For inappropriate concepts, we perform iterative erasure for 5 epochs and set λ in Eq. (7) to $1e-1$. For artistic style, we conduct erasure for 10 epochs and set λ to $1e-3$. The baselines we will compare with are: SD V1.4 [25], SD V2.1 [31](Stable Diffusion pretrained on an NSFW filtered dataset), SLD [29], ESD [6], CA [13], SA [8], UCE [7]. As for SLD, ESD, SA and UCE, we adhere to the recommended configuration in their papers [6–8,29]. For CA [13], we fine-tune the full weights of U-Net to erase unsafe contents and the cross-attention module to erase artistic styles, according to its documentation[3].

4.1 Unsafe Content Removal

Experimental Setup In this section, we assess the effectiveness of erasing unsafe concepts. We conduct experiments on the Inappropriate Image Prompts (I2P) dataset [29]. The I2P dataset includes various inappropriate prompts, such as violence, self-harm, sexual content, and shocking content. These prompts are collected from real-world, user-generated images based on the official SD. Our

[3] `https://github.com/nupurkmr9/concept-ablation`

Method	Nudity Detection									COCO-30k	
	Breast(F)	Genitalia(F)	Breast(M)	Genitalia(M)	Buttocks	Feet	Belly	Armpits	Total↓	CLIP↑	FID↓
SD v1.4	183	21	46	10	44	42	171	129	646	**31.33**	-
SD v2.1	121	13	40	3	14	39	146	109	485	-	-
ESD-u	14	1	8	5	5	24	31	33	121	30.45	3.73
UCE	31	6	19	8	11	20	55	36	186	31.26	**1.82**
SLD-Med	72	5	34	3	18	19	104	99	354	30.95	2.60
SA	39	9	**4**	**0**	15	32	49	**15**	163	30.57	17.34
CA	**6**	1	9	10	4	14	28	23	95	31.16	7.87
Ours	8	**0**	6	4	**0**	**8**	**23**	17	**66**	30.95	2.82

Table 1: Comparison of performance metrics for content removal methods. (Left) Number of nude body parts detected by Nudenet on I2P dataset with threshold 0.6. (Right) CLIP-score and FID against original SD. F: Female. M: Male. **Bold**: best. Underline: second-best.

evaluation focuses on the erasure of nudity since it is a classical unsafe concept. For each model, we generate one image per prompt in the I2P dataset, resulting in a total of 4703 images. Nude body parts are detected using the Nudenet detector [18], with the threshold set to 0.6. This threshold follows the default settings in I2P[4].

To verify that the unlearned models can still generate normal images, we use COCO-30k [14] with its captions as prompts. COCO-30k is a dataset devoid of unsafe concepts, making it suitable for evaluating edited models' generation capabilities. We evaluate the models' image-text consistency based on CLIP-score [9], and visual similarity against SD-generated images based on FID [17].

Removal Results As depicted in Tab. 1, our method yields the lowest number of nude body parts, while demonstrating impressive specificity in preserving normal content of COCO-30k. CA generates the second-fewest nude body parts but it exhibits poorer performance in terms of FID, indicating a poorer trade-off between generation ability and removal effectiveness. On the other hand, CA fine-tunes full weights and our RECE only fine-tunes cross-attention modules, which will be discussed in detail in Sec. 4.4.

Notably, our method achieves a FID score closely comparable to the top-performing UCE and the second-best SLD, both of which generates a considerably higher number of nude body parts. This suggests that our method minimally impacts the generation of normal content while striving for better removal effectiveness. Additionally, most methods exhibits favorable CLIP-score results thus we consider the CLIP-score performance acceptable as long as it remains within a reasonable range.

In open-sourced conditions, the inference guidance mechanism such as SLD [29] can be easily bypassed by deleting the corresponding code [24]. Large-scale model retraining on NSFW-filtered dataset demands considerable computational re-

[4] `https://github.com/ml-research/i2p`

Fig. 3: Qualitative results of different nudity removing methods. Images for each row are generated using the same prompt. **All prompts originate from I2P benchmark, and we don't intentionally generate inappropriate content of any individual.**

sources [22], and even then, the model SD v2.1 may still generate nude images, as illustrated in Tab. 1.

Qualitative results are illustrated in Fig. 3. In the first row, all erasing methods successfully generate non-nude images, but the results of CA and SA differ significantly from the original SD's. In the second row, all methods except for SD v2.1 avoid generating nude images. Among these, only UCE, SLD, and our method effectively capture the facial features. These findings demonstrate that our method effectively maintains unrelated concepts.

Nudity Bias While our RECE demonstrates remarkable effectiveness in minimizing the generation of nude content, it exhibits limitations in erasing male nudity and similar limitations are also observed in other methods. We count the number of female and male nude body parts in the 4703 I2P images. As illustrated in Tab. 2, the nudity ratios between women and men almost decrease in every method, indicating an erasure bias on female information (excluding SA, which erases more sex-related concepts besides nudity [8]). We attribute this limitation to the inherent bias in the target concept "nudity" within SD, which tends to generate more female-oriented content. To investigate this bias, we randomly selected 20 seeds and employed SD V1.4 to generate 3 images per seed with the prompt "nudity", resulting in a total of 60 images. Surprisingly, almost all of these generated images depict female body part as presented in the last column of Tab. 2. Thus our derived embedding is also biased. Further improvements require an awareness of the biases inherent in the model while performing erasure.

4.2 Artistic Style Removal

Experimental Setup We conduct an evaluation to assess the efficacy of removing artistic styles to address copyright concerns. Following the datasets in ESD [6], we use 20 prompts for each of 5 famous artists—Van Gogh, Pablo Picasso, Rembrandt, Andy Warhol and Caravaggio—and 5 modern artists—Kelly

Metrics	SD	ESD	UCE	SLD-Med	SA	CA	Ours	SD-"nudity"
# Female Nudity	204	15	37	77	48	7	8	89
# Male Nudity	56	13	27	37	4	19	10	1
Female/Male	3.64	1.15	1.37	2.08	12	0.37	0.8	89

Table 2: In the first 7 columns, we counted the numbers of female nudity and male nudity body parts for I2P dataset. The last column is the result of SD v1.4 conditioned by the prompt "nudity". This table highlights the inherent bias within SD. And all each erasure methods can not remove the concept of male nudity very well.

Removal Method	Erase "Van Gogh"			Erase "Kelly McKernan"		
	$\text{LPIPS}_\text{e} \uparrow$	$\text{LPIPS}_\text{u} \downarrow$	$\text{LPIPS}_\text{d} \uparrow$	$\text{LPIPS}_\text{e} \uparrow$	$\text{LPIPS}_\text{u} \downarrow$	$\text{LPIPS}_\text{d} \uparrow$
ESD-x	**0.40**	0.26	0.14	**0.37**	0.21	0.16
UCE	0.25	**0.05**	<u>0.20</u>	0.25	**0.03**	<u>0.22</u>
SLD-Med	0.21	0.10	0.11	0.22	0.18	0.04
CA	0.30	0.13	0.17	0.22	0.17	0.05
Ours	<u>0.31</u>	<u>0.08</u>	**0.23**	<u>0.29</u>	<u>0.04</u>	**0.25**

Table 3: Comparison of LPIPS scores for artistic removal methods. **Bold**: best. <u>Underline</u>: second-best. LPIPS_d indicates overall erasure performance.

McKernan, Thomas Kinkade, Tyler Edlin, Kilian Eng and the series "Ajin: Demi-Human", which have been reported to be imitated by SD [30]. To evaluate our RECE and all the aforementioned baselines, we erase the style of two artists: Van Gogh and Kelly McKernan.

Removal Results We conducted an evaluation based on LPIPS scores [37] compared to the original SD, as detailed in Tab. 3. LPIPS evaluates the perceptual distance between image patches, where higher values indicate greater differences and lower values indicate more similarity. The LPIPS_e is calculated on the erased artist. A higher LPIPS_e value suggests a more effective style removal, and both ESD and our method demonstrate successful erasure of the target style. LPIPS_u is calculated on unerased artists. A lower LPIPS_u indicates a lesser impact on unrelated artists, where our method and UCE effectively maintains unrelated concepts. We also calculate the overall effectiveness by $\text{LPIPS}_\text{d} = \text{LPIPS}_\text{e} - \text{LPIPS}_\text{u}$, which is the difference between erased and unerased artists. Our method performs best in this regard. Qualitative results can be found in Appendix.

4.3 Robustness Against Red-teaming Tools

Experimental Setup To demonstrate the robustness of our RECE in safeguarding against various attack methods, we employ different red-teaming tools, including white-box methods such as P4D [3] and UnlearnDiff [38], and the black-box method Ring-A-Bell [32]. For nudity, as provided by UnlearnDiff [38],

Red-teaming	SD v1.4	SD v2.1	ESD-u	UCE	SLD-Max	SA	CA	Ours
UnlearnDiff	-	-	<u>66.20</u>	79.58	82.39	77.46	**65.49**	**65.49**
P4D	-	-	<u>63.38</u>	80.28	77.46	78.87	**60.56**	64.79
Ring-A-Bell	83.10	72.54	69.72	33.10	66.20	<u>22.54</u>	25.35	**13.38**
Average	-	-	66.43	64.32	75.35	59.62	<u>50.47</u>	**47.89**

Table 4: Robustness of different methods against red-teaming tools, measured by attack success rate(%). **Bold**: best. <u>Underline</u>: second-best.

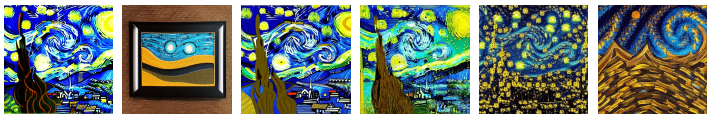

Fig. 4: Qualitative results of different artistic style removal methods for eliminating the style of Van Gogh under Ring-A-Bell's attack. The prompt is "A depiction of a starry night over a quiet town, reminiscent of Van Gogh's famous painting".

we use a set of 143 prompts selected from I2P, each with a nudity score (as determined by NudeNet) above 0.75, and employ the Nudenet detector with a threshold set to 0.45 for detecting inappropriate content.

Results The attack success rates (ASR, %) are summarized in Tab. 4. Our method achieves the best robustness in average. In the case of the black-box attack Ring-A-Bell, our method achieves the lowest ASR at 13.38%, significantly outperforming other methods. In the case of the white-box attack, CA achieves the best performance, while our method performs either the best or very closely to the second-best. Although SA achieves a decent black-box result, it consumes substantial computation resources for generating 5000 prepared images [8]. CA modifies 100% of the U-Net parameters while our method only modifies K&V projection matrices, constituting a mere 2.23% of the U-Net. While UCE also modifies only 2.23% parameters, our method significantly outperforms UCE. This is attributed to our derived embeddings in Eq. (8).

For artistic style removal, we provide qualitative examples in Fig. 4. The first image is generated by the original SD v1.4 without any attack, and the following images are under Ring-A-Bell's attack [32]. Recall that UCE is the second-best artistic style removal method as shown in Tab. 3 but it falls short in robustness against red-teaming tools. Although our method employs a closed-form solution like UCE, it outperforms UCE in robustness. ESD, CA and our method perform similarly well. More results can be found in the appendix.

4.4 Model Editing Duration

To demonstrate the efficiency of different methods, we measured the percentage of parameter modification and editing duration on an RTX 3090 for each

	ESD	UCE	SA	CA	Ours
Modification (%)	94.65	2.23	94.65	100	2.23
Duration (s)	3720	1.2	-	1400	3.4

Table 5: Percentage of parameter modification and model editing duration for different methods. Our method and UCE are significantly ahead of other methods.

method, as shown in Tab. 5. We excluded SLD from the analysis since it operates at inference time rather than modifying the model's weights which can be easily bypassed under open-source conditions. Additionally, we don't include the duration of SA, as it involves generating 5000 images, calculating the Fisher Information Matrix and fine-tuning, which makes it exceptionally slow.

Based on Tab. 5 and Tab. 1, our method achieves the best concept erasure effect in an extremely short time of only three seconds. Our method (5 epochs) and UCE (1 epoch) modify the lowest percentage of parameters with a closed-form solution, resulting in the shortest editing durations. Despite similar durations, our method significantly outperforms UCE in removal effectiveness. Conversely, CA, ESD and SA modify a high percentage of parameters with more time but achieve less impressive removal results.

4.5 Ablation Study

We conduct experiments to elucidate the impact of our derived embedding among different epochs and the effectiveness of the regularization term.

Effect of Derived Embeddings among Epochs

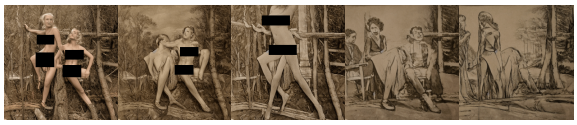

Fig. 5: There is no nudity information from epoch 3 onward; hence, we select the checkpoint after epoch 2 to avoid damaging the model's ability.

We conduct an experiment to expound the impact of our derived embedding. We perform "nudity" erasure for 5 epochs using Algorithm 1. In each epoch, we derive a distinct embedding that represents "nudity". Before the erasure of each epoch, we generate an image using the embedding to test its degree of nudity information, as presented in Fig. 5. Images from epoch 0 to epoch 2 contain nude body parts, indicating that our derived embeddings successfully reveal potential nudity information in the model. Specifically, we opt for the checkpoint after epoch 2, as images from epoch 3 to 4 lack nudity information. Actually, erasing such "not so nude" embeddings in epoch 3-4 would impair the model's normal generation ability, which is an unworthy trade-off.

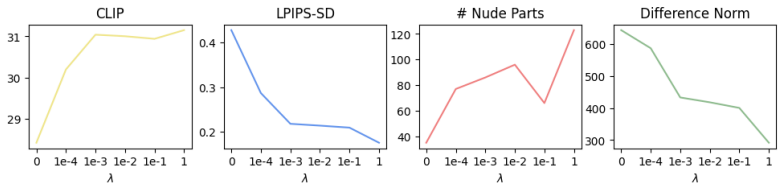

Fig. 6: Ablation study on the impact of the regularization term on model performance.

Effect of Regularization Coefficient We conduct an experiment to assess the influence of our regularization term. Specifically, we select different regularization coefficients λ in Eq. (4), which divide the interval $[0, 1]$ into five parts. The results, presented in Fig. 6, include CLIP-score and LPIPS against the original SD on the COCO-30k validation subset. As λ increases, CLIP score shows an upward trend while LPIPS and the difference between new and old parameters show a downward trend. This indicates the role of the regularization term in preserving the model's ability for unerased content. Furthermore, we recorded the number of nude parts on the I2P benchmark, presented in the third column of Fig. 6. However, the number of nude parts doesn't strictly increase as λ increases, which is counterintuitive. Although the purpose of the regularization term in Eq. (6) is to preserve the model's generation capability, maintaining this capability does not always affect the erasure effect.

5 Conclusion

In this paper, we propose a novel approach for reliably and efficiently erasing specific concepts from Text-to-Image (T2I) diffusion models. Our approach only modifies the cross-attention K&V matrices of U-Net, constituting a mere 2.23% of parameters. While previous methods also edited cross-attention modules, they still exhibited the ability to generate inappropriate images. To tackle this challenge, we derive and erase new embeddings that can represent target concepts within unlearned models. To mitigate the impact on unrelated concepts, a regularization term is introduced during the erasure process. All the above techniques are formulated in closed-form, facilitating rapid editing. This enables the execution of "derive-erase" across multiple epochs, ensuring thorough and robust erasure. Extensive experiments were conducted to validate the effectiveness of our approach in erasing artistic styles, unsafe contents and object classes. Furthermore, we recorded editing durations to underscore the efficiency of our method and evaluated the robustness against red-teaming tools. We believe our RECE has the potential to empower T2I providers in effectively removing undesired concepts, thereby fostering the development of a safer AI community.

Acknowledgements

This project was supported by National Key R&D Program of China (No. 2021ZD0112804).

References

1. Chen, K., Wei, Z., Chen, J., Wu, Z., Jiang, Y.G.: Attacking video recognition models with bullet-screen comments. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 312–320 (2022)
2. Chen, K., Wei, Z., Chen, J., Wu, Z., Jiang, Y.G.: Gcma: Generative cross-modal transferable adversarial attacks from images to videos. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 698–708 (2023)
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. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
5. Doersch, C.: Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908 (2016)
6. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. arXiv preprint arXiv:2303.07345 (2023)
7. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5111–5120 (2024)
8. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. Advances in Neural Information Processing Systems **36** (2024)
9. 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)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Hunter, T.: Ai porn is easy to make now. for women, that's a nightmare. (2 2023)
12. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
13. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22691–22702 (2023)
14. 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)
15. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. pp. 16784–16804. PMLR (2022)

16. Orgad, H., Kawar, B., Belinkov, Y.: Editing implicit assumptions in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7053–7061 (2023)
17. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11410–11420 (2022)
18. Praneeth, B.: Nudenet: Neural nets for nudity classification, detection and selective censoring (2019)
19. 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)
20. 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)
21. Rando, J., Paleka, D., Lindner, D., Heim, L., Tramèr, F.: Red-teaming the stable diffusion safety filter. arXiv preprint arXiv:2210.04610 (2022)
22. Rombach, R.: Stable diffusion 2.0 release (November 2022)
23. Rombach, R.: Tutorial: How to remove the safety filter in 5 seconds (August 2022)
24. 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. pp. 10684–10695 (2022)
25. Rombach, R., Esser, P.: Stable diffusion v1-4 model card. Model Card (2022)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
27. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
28. 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. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
29. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22522–22531 (2023)
30. Setty, R.: Ai art generators hit with copyright suit over artists' images (1 2023)
31. StabilityAI: Stable diffusion 2.1 model card. Model Card (2022)
32. 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)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
34. Wei, Z., Chen, J., Goldblum, M., Wu, Z., Goldstein, T., Jiang, Y.G., Davis, L.S.: Towards transferable adversarial attacks on image and video transformers. IEEE Transactions on Image Processing **32**, 6346–6358 (2023)

35. Wei, Z., Chen, J., Wu, Z., Jiang, Y.G.: Adaptive cross-modal transferable adversarial attacks from images to videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
36. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys **56**(4), 1–39 (2023)
37. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
38. 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)